EXPLANATION IN CAUSAL INFERENCE

Methods for Mediation and Interaction

TYLER J. VANDERWEELE

OXFORD

Explanation in Causal Inference

Explanation in Causal Inference

Methods for Mediation and Interaction

TYLER J. VANDERWEELE

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016
www.oup.com

Oxford is a registered trade mark of Oxford University Press in the UK and in certain
other countries.

Library of Congress Cataloging-in-Publication Data
VanderWeele, Tyler.
Explanation in causal inference : methods for mediation
and interaction / Tyler VanderWeele.
pages cm
Summary: "A comprehensive book on methods for mediation and interaction.
The only book to approach this topic from the perspective of causal
inference. Numerous software tools provided. Easy-to-read and
accessible. Examples drawn from diverse fields. An essential reference
for anyone conducting empirical research in the biomedical or
social sciences"– Provided by publisher.
ISBN 978-0-19-932587-0 (hardback)
1. Social sciences – Research. 2. Social sciences – Methodology. 3. Causation. I. Title.
H62.V3238 2015
001.4′22 – dc23 2014029661

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

To my teachers, from whom I have learned to reason well; and to my family and loved ones, from whom I have had the support to do so.

CONTENTS

PREFACE

This book developed out of a course and set of lectures on the topic of methods for mediation and interaction that I have offered at the Harvard School of Public Health. The course was structured so as to be accessible to second-year graduate students in applied disciplines—such as epidemiology and the social and behavioral sciences—who had had a one-year introductory sequence in statistics. The lectures and the present book approach these topics from a counterfactual-based perspective on causal inference. Many of the students attending the lectures at Harvard had previously acquired an introductory knowledge of causal inference and counterfactuals in other courses at the university. In trying to bring the material in this book to a broader audience, a decision needed to be made as to whether to also presuppose, from the reader, a similar background in such introductory principles of causal inference.

After some thought, along with conversations with colleagues, I decided that, so as to attempt to extend the reach of this book as broadly as possible, I would not presuppose such a background, but that I would rather describe as many of the methods and assumptions as possible without requiring specific appeal to the notation of counterfactual-based logic. The methods presented in the book are shaped by such ideas, but the reader is not required to have had any previous exposure to them. The reader may end up acquiring a background in counterfactuals and causal inference simply by reading the book, but this would not be required to benefit from the book. There are many fine book-length introductions to counterfactuals and causal inference elsewhere (Morgan and Winship, 2007; Pearl, 2009; Hernán and Robins, 2015; Imbens and Rubin, 2015). Occasionally, when appeal to counterfactual notation is necessary (as occurs in some of the later chapters in the book), the reader is explicitly warned and can pass over these sections if desired. I believe that in almost all such cases, the reader could skip over these sections without jeopardizing comprehension of the material that appears subsequently in the book.

To the best of my knowledge, this book is unique in addressing the topics of mediation and interaction from a counterfactual-based perspective on causal inference. The book, I believe, would be unique in addressing either topic from a counterfactual-based perspective. While there have been a few book-length

treatments of interaction from a more statistical perspective (Aiken and West, 1991; Jaccard, 2001; Jaccard and Turrisi, 2003), I was, at the time of the writing of this book, aware of only one book-length treatment of the topic of mediation: David MacKinnon's *Statistical Mediation Analysis* (MacKinnon, 2008). Although there is some overlap between MacKinnon's book and the present text, this overlap is not (with the first edition of his book at least) very substantial, and a reader could benefit from the reading of both books. David MacKinnon's book discusses methods for mediation that have developed within the social science literature, primarily psychology. It is a reasonably comprehensive text in that regard. The causal inference literature on mediation, which is the focus of this book, developed, in part motivated by, but also largely independently of, that social science literature. The causal inference literature has focused more on the formal definitions, the substantive assumptions, and the actual interpretation of effect estimates that are part of mediation analysis. While assumptions are, to a certain extent, discussed in some of the work on mediation in the social sciences, many social science methods papers ignore the substantive confounding assumptions that are needed for a causal interpretation of effect estimates. As will be seen in the chapters of this book, this can sometimes bring about very misleading results, leading the analyst toward incorrect conclusions. While the social sciences have gradually been seeing increasing attention paid to the confounding assumptions once again, these assumptions are still very often ignored when the techniques are used in practice, especially in psychology. The book will devote considerable discussion to these issues, and it is concerning these issues that the counterfactual-based perspective on causal inference is especially valuable when thinking about mediation.

In many ways, however, the social science literature on mediation is far ahead of the causal inference literature in the scope and settings for which mediation analysis techniques have been developed. However, many of these techniques from the social sciences have not yet been framed within the counterfactual framework; and, as such, the assumptions required for the use of these techniques, and the precise interpretation of the effect estimates from these techniques, are not yet clear. It is my belief that, over the next decade, there will be a powerful synthesis of the methods that have developed in the social science literatures with the rigorous foundation provided by the counterfactual framework. It is my hope that this book is a step toward that synthesis.

I have endeavored to write this book in as accessible a manner as possible. The only knowledge this book presupposes is an approximately year-long sequence in applied statistics through linear and logistic regression. As such, the book would be appropriate for use as a second-year graduate textbook in applied disciplines, or perhaps even as an advanced undergraduate text. Some readers, even with this background, may still find some of the material challenging. I have put considerable effort into making the book as accessible as possible. It is conceivable that it could be simplified yet further still, but I believe that I have, at least in the core central Chapters 2, 3 and 9, come close to reaching the limits of my ability in this regard. I have taught a number of short courses on this material, at conferences in epidemiology and in the social sciences. These courses have, I think, been pitched at a slightly higher technical level than the present book. However, I do believe that many of the participants in these courses came away with substantially improved understanding of the methodology presented. And so it is my hope for this book also that, being

pitched at a slightly more accessible level, it would likewise be capable of assisting a broad range of readers. Perhaps I overestimate what I have actually been able to convey in the courses, or underestimate the importance of face-to-face teaching, but I truly have endeavored to make much of the book, especially Chapters 2, 3, and 9, yet more accessible than the courses that I teach.

While the primary audience of the book is applied empirical researchers in the biomedical and social sciences, the book has also been written with the hope of appealing to another audience as well: statisticians and methodologists. As discussed below, considerable further methodological development in the some of the topics covered by this book is needed, and it will fall to statisticians and other methodologists to carry out this work if further progress is to be made. Although the book is targeted primarily for a nonstatistical reader, the book may appeal to statisticians and methodologists in two respects. First, it provides a relatively thorough and comprehensive treatment of the causal inference literature on mediation and interaction and discussion of what questions and problems are still open—the book may thus serve to help map out the field, as well as future areas of inquiry. Second, while the text of the book has been written so as to be relatively accessible, the book also includes a rather lengthy technical appendix that contains all of the formal and theoretical development of the methods in the book. The Appendix provides the corresponding formal notation, definitions, propositions, theorems, proofs, and so on, for the more descriptive and accessible coverage in the body of the text. For a more technical reader, the Appendix could almost be read as a theoretical and rigorous, albeit terse, text in its own right, though the main text would still effectively be necessary for discussion of motivation and significance. In any case, the Appendix does compile, in one place, formal statements and proofs of the major results in the field, material that can, at present, only be found by searching appendices of papers, online supplements at journals, and authors' technical reports, all of which often have conflicting notations across sources. The technical appendix here brings all of this material together in one place and in a common format. It is hoped that this second feature of the book, for the statistical reader, will be at least as useful as the first.

With regard to the selection and the ordering of the material, I have been asked both "Why not just write a book on mediation alone? Why include interaction?" and also "If you are going to include interaction, ought not the material on interaction precede that on mediation in order?" Concerning the first question, the broad topic of the book is explanation in causal inference; it is a book on understanding mechanisms. In my view, interaction is an important part of understanding mechanisms and can often be an important aspect of explanation. Such issues are dealt with at greater length in this book's first, introductory, chapter, and extensively in Part II of this book, especially in Chapter 10. To omit this material on interaction would, I believe, have considerably weakened the range of methods and concepts that a reader would have access to in understanding mechanisms and explanation in causal inference. It is, moreover, my experience that many empirical researchers in the biomedical and social sciences feel that they "understand" interaction and believe that what they need to learn about is mediation, but that, in reality, they would in fact benefit considerably from further training on concepts and methods for interaction as well. As will be clear in Part II of this book, questions of interaction extend far beyond whether a product term in a statistical model is significant.

The material on interaction has been included in this book to deepen the readers' understanding of this important set of topics.

Concerning the second question, "Ought not the material on interaction precede that on mediation?," I am left with considerably more uncertainty. It may have indeed been preferable to reverse these two parts of the book. The decision on the present ordering was ultimately made on pragmatic grounds. As suggested above, I believe it is likely that the primary market for this book will be those wanting to learn more about mediation. If this is so, it seemed preferable to give such readers what they wanted up front without it seeming necessary to work through numerous chapters before getting to the material in which they are principally interested. This decision on the ordering of the material is in some ways in tension of course with one of the reasons I gave above for including interaction in the book in the first place. This decision concerning the ordering has been somewhat of a balancing act between meeting what I perceive as reader demands and encouraging new learning in areas with which readers may believe themselves to already be familiar.

As with nearly any text, this book too has its own shortcomings. The book is, of course, strongly shaped by my own perspective, knowledge, and biases. In a book that attempts to survey a field, biases are in part exerted by the selection of material. While I have attempted to be quite broad and inclusive in the work that is mentioned and cited, priority has been given to devoting larger sections of the text to methods that, given the current state of methods development and software tools, can be relatively easily implemented now, at the time of writing. This prioritization is in line with the primary intended audience of this book, empirical researchers in the biomedical and social sciences. In my own methodological research, I have oriented a good deal of my work toward methods that can easily be employed. Much of the book thus does, admittedly, describe methods that I have worked on and developed. I have tried also to devote considerable space to methods developed by other researchers, especially those that can be easily implemented in practice. The work of many other methodologists is discussed throughout the book. In my view, the major comprehensive alternative set of methods and software for mediation, from the causal inference perspective, to my own is that which has been developed by Kosuke Imai and colleagues. This work is discussed in Chapters 2, 3, and 5. I believe that their techniques and software are very useful and cover a number of cases that my own methods do not. I include discussion of their methods in all courses that I teach, and they are included in this book as well. As will be seen in Chapter 2, the methods that they have developed and that I and my colleagues have been working on are very closely related. In many cases, we have been working on methods to address the same set of issues and problems but using different approaches; and it has often been a pleasure watching how they have addressed a similar question or issue but using a different technique. I apologize to them if I have not done full justice to their methodological work in this book; there have inevitably been omissions on my part. The book's content has been shaped not only by prioritizing methods that can be easily implemented, but also by those that I best know and understand and could easily write about, which once again has likely resulted in an overemphasis on my own work. This book might be viewed as much a research monograph as it is a textbook. Notably absent also from the book is discussion of many of the methods and software options that have come out of the social science literature outside of the counterfactual tradition. As noted above, much of this is still yet to be

framed within a counterfactual-based perspective on causal inference; this material has thus not been included in the present edition of this book, and so a reader will have to turn elsewhere for descriptions of these other software options.

This brings us to another limitation of the book, which is the choice of software. This was not an easy decision. Different disciplines give priority to different software packages. Some of the methods for mediation and interaction have been developed in some software packages but not others. A book that attempted to describe all methods in multiple software packages would become too cumbersome. I have tried to strike a balance. In Chapter 2, which describes the core set of methods for mediation, and in Chapter 9, which describes the core set of methods for interaction, multiple software packages are covered. In Chapter 2, macros and commands for mediation are described in SAS, SPSS, Stata, and R. Fairly general methods are now available in all of these packages. In the remainder of this book (outside of Chapters 2 and 9), the software implementation of methods is given in SAS but references are provided for the use of similar techniques in other software packages where available. I apologize to readers unfamiliar with SAS. Some decision had to be made, and SAS was the software package with which I was most familiar and had easiest access to existing code for methods for mediation and interaction. SAS is also, in my experience, one of the packages most commonly employed by biomedical researchers, one of the primary intended disciplinary audiences of the book. For users of other software packages, hopefully the inclusion of the SAS code will provide a template as to what code might look like in other packages as well and how it could be adapted. And once again, references to other papers with similar code for other software packages are given wherever possible. The software resources for this book can be downloaded at: www.oup.com/us/causalinference/.

Many of the ideas, techniques, and methods described in this book are quite new, some of the material being published in journal articles almost concurrently with the book itself. As such, a number of the methods that the book presents have not yet had an opportunity to become widely disseminated and thus have only just begun to be applied to data. Consequently, another shortcoming of the book is that, in some cases, there were a limited number of options with regard to the choice of empirical studies with which to illustrate the methods. Whenever possible, I have tried to select examples that are of interest from a substantive perspective, and not simply of methodological curiosity. However, this has not always been possible, and thus some of the examples really serve more as toy illustrations than as substantive applications. This is especially so with some of the sections in Chapter 5. In these cases, I thought that it was nevertheless preferable to have a toy example rather than no illustration at all. It is my belief that, even in these instances, the methods presented are, and will be, of importance in application; the difficulty in finding an ideal example in these settings is not because the methods are not useful, but because they are still very new.

Yet another shortcoming of this book is one that really cannot be avoided: The methods in this field are developing very rapidly and within a few years this book may well be out of date. An updated edition of the book may be desirable very quickly; but by the time such an update is reasonable, the field may already be too vast to summarize in a single book. Another related shortcoming of the book is that it inevitably does not address all settings that may be of interest in assessing mediation and interaction. The book is, of necessity, limited to the current state

of methodological development. In some settings it would be very helpful if methods were already available, but they are not; at least, not yet. Some discussion of what methodological developments remain to be done is given at the end of each chapter. Further synthesis of what will be needed in future methods development, along with a summary of the end-of-chapter discussions, is then also provided in the book's final chapter. Tremendous progress has been made in methods for mediation and interaction over the past decade. However, as will be seen throughout the book, this is still a very active area of methodological research, and much work remains. It is my hope that, in spite of the book's limitations, it will nevertheless be a useful guide to methodology for assessing mediation and interaction, that it will increase the understanding of mechanisms and of causal explanation, and that it will do so, hopefully, ultimately, in ways that advance not only knowledge, but human health, society, and flourishing.

I am indebted to many people, as well as several institutions, for their assistance in the writing and publishing of this book. Thanks first to Abby Gross, Emily Perry, and Molly Balikov, along with the other Oxford University Press staff, for helping to see this book through the production process. Thanks to John Willett and Dick Murnane for their introduction to Abby and for conversations on the topic of book writing. Thanks to Judy Singer for what proved to be a formative conversation on the level and intended audience of the present book. With regard to the content of the book, special thanks to Jamie Robins, from whom I learned causal inference and without whose teaching and mentorship I would not have been able to take on the research presented in the book. Thanks also to Thomas Richardson, Eric Tchetgen Tchetgen, Dustin Tingley, Teppei Yamamoto, Kosuke Imai, Sander Greenland, Bhramar Mukherjee, Elizabeth Ogburn, Theis Lange, John Jackson, Peng Ding, Zhichao Jiang, Linda Valeri, Etsuji Suzuki, Yasutaka Chiba, StijnVansteelandt, and Jan Vandenbroucke for numerous helpful suggestions and comments, as well as for catching a number of my errors that had previously been in the text; thanks to them also for, in many cases, their own very valuable contributions to the field, many of which are covered in the book. I do apologize to them and to my readers if any of the material is unfairly represented or misrepresented and for any errors that may still be present; for these I take full responsibility.

Thanks also to many students, colleagues, and collaborators for countless questions, many of which motivated the methodological developments that this book presents. Thanks to the National Institutes of Health for financial support of my methodological research in this area. Thanks to the Harvard School of Public Health and Harvard's Department of Epidemiology for granting me leave for the fall of 2012 to begin the task of book writing. Thanks to Paul and Margaret Isenman for the use, that fall, of their apartment in Paris, where the majority of this book was in fact written. And finally, special thanks to Lisa, to whom, by the time this book is published, I will be married and who, for a number of months, graciously put up with my absence from Boston so that I was able to escape the demands of ordinary work life, in order to be able to write. Thanks to her also for her love and support throughout this process.

Tyler J. VanderWeele
Cambridge, Massachusetts

PART ONE

Mediation Analysis

1

Explanation and Mechanism

The topic of this book is explanation in causal reasoning. More specifically, we will be concerned with empirical methods that can provide insight into the explanation of causal phenomena. We will consider what we can learn about causal explanation from data, and also with the limitations of what we can learn. We will ask questions not simply about whether an exposure affects an outcome but why, and for whom, and how. In slightly more technical language, we will be concerned principally with "mediation" and "interaction."

The Latin root of "mediation" is "mediarai," to divide into two equal parts or settle a dispute by intervening, related to the past participle "mediatus," acting through an intermediate agent. The word "mediation" has been used in the social science literature for some time for the state in which one cause affects some intermediate that, in turn, goes on to affect an outcome. Mediation then, as considered in this book, is concerned with both (a) the processes by which one exposure or variable or state affects another and (b) why something occurs. The English word "interaction" is derived from the Latin roots "inter-" (among, between, or during) and "agere" (to act). In the context of thinking about effects of exposures or causes, interaction is understood as action between or among two or more causes.

Most of the book will be concerned with concepts and methods to more formally define and assess these phenomena of mediation and interaction. We will describe various statistical techniques that may help with assessing these phenomena, and we will discuss (a) the assumptions required to interpret statistical analyses causally and (b) the robustness, or lack thereof, of causal conclusions to violations in the assumptions. In this chapter, however, we will consider the nature of explanation and its relation to causation itself in broader, more conceptual, terms. We will consider different forms of explanation when we are reasoning about causation, and we will discuss how the phenomena of mediation and interaction provide different types of explanations for cause–effect relationships. We will also describe what might motivate a researcher to investigate these phenomena of mediation and interaction empirically, and we will conclude this chapter with a brief description of the remainder of the contents of this book.

1.1. CAUSAL INFERENCE AND EXPLANATION

The formal and technical approach we will take with regard to addressing questions of mediation and interaction is that of the "potential outcomes" or "counterfactual" framework for causal inference, which is now widely being employed in formal methodological work in statistics, epidemiology, economics, sociology, psychology, education, computer science, and other disciplines. The framework provides a formal and technical notation to conceptualize causation. This is done principally by conceiving of what might have occurred had some action or state been otherwise than it was. If some outcome would have differed had some exposure or action been other than it was, then we would say that the exposure or action causes or affects the outcome. The idea of conceptualizing causation in terms of counterfactual states goes at least as far back as Hume (1748). Within the statistics literature, formal notation for this counterfactual approach was described by Neyman (1923) in the context of randomized agricultural experiments. The framework was later developed by Rubin (1974, 1978) and extended to observational studies. The framework was further extended to multiple exposures and exposures that vary over time by Robins (1986) and was related to graphical representations of causality by Spirtes et al. (1993) and Pearl (1995, 2001).

Although we could conceptually think about what would have occurred under each of two or more actions, in practice we typically only know what actually did occur. The outcome that would have occurred in the counterfactual state in which we took an action other than the one that was in fact taken is essentially missing or unknown. This is what makes causal inference challenging with empirical data. We can define the causal effect in an individual instance as the difference in the outcomes that would have occurred under each of two potential actions, but we will not in general know what this difference or effect is for an individual. The outcomes that would have occurred under each of two potential actions are often referred to as "potential outcomes" or "counterfactual outcomes."[1] Although it is thus difficult to draw conclusions about causal effects for specific individuals, it is sometimes possible to make inferences about such effects on average for a population. Randomization of the action or exposure or intervention can help ensure that the groups receiving the different actions or interventions are comparable on average and thus that any difference in the outcomes between the groups receiving different actions or interventions is attributable to the action itself, rather than to some other factors. With observational data, in which the actions are not randomized, we might

1. Here we will use "potential outcomes" and "counterfactual outcomes" interchangeably. Some authors prefer one term over the other, and sometimes the two terms are distinguished linguistically. Further details on the linguistic distinctions that are sometimes drawn are given in the Appendix. Here we will follow what has for the most part become common convention; with slight abuse of linguistic nuance, we will use the two terms interchangeably. None of the analytic development is dependent on this, and readers can themselves distinguish the two terms or preferentially use one over the other.

still try to control for various other factors that might explain differences in the outcomes across intervention groups other than the intervention itself. Such control can help attribute differences in outcomes to the cause or action under study, but with observational data in the absence of randomization, one cannot in general be sure that such control has been adequate.

Arguably not every instance of causation falls within this counterfactual framework. An event may be the cause of some outcome without it being the case that if the former event had not happened, then the outcome would not have happened either. This can occur if a specific event was in fact *the* cause of the outcome, but if this cause hadn't been operative, some other cause would have been, and thus the outcome would have occurred anyway. This phenomenon is sometimes referred to as "overdetermination" in the philosophical literature. Although basic counterfactual conditions are thus perhaps not necessary for causation, there is general consensus that if a phenomenon does fall within the purview of the counterfactual framework, then it constitutes an instance of causation. The framework thus does not provide a formal characterization of all aspects of causality; and numerous questions about causation, such as the criteria by which we identify the actual cause of an event, are also essentially left unaddressed by the potential outcomes framework.[2] We will return to certain philosophical questions concerning causation and counterfactuals in the final chapter of the book. What the counterfactual framework allows for principally is a set of definitions that provide either criteria or sufficient conditions indicating that some event or exposure was a cause of another—not necessarily *the* cause, but a cause. The framework moreover provides formal criteria concerning when we can draw such conclusions about causation from empirical data; in other words for the assumptions that need to be made, or are sufficient, to move from conclusions about association to conclusions about causation.

The methods and approaches described in this book utilize this counterfactual or potential outcomes perspective. The book, however, does *not* presuppose familiarity with this counterfactual framework. But it is out of this counterfactual tradition that the methods described in the book have been developed. For the interested reader a few texts are now available or in development (Morgan and Winship, 2007; Hernán and Robins, 2015; Imbens and Rubin, 2015; and Pearl, 2009, for a graphical perspective) that give the formal details of this counterfactual framework. However, again, the book itself does not require prior knowledge of the framework. The book only presupposes some familiarity with basic statistics, specifically linear and logistic regression methods. The reader will gradually become acquainted with this counterfactual approach to thinking about causation as the book progresses, and Section 2.16 in the next chapter provides some discussion about how the framework relates to concepts of mediation.

Although the book does not presuppose familiarity with causal inference, it does draw upon insights from and explains in intuitive terms the ideas about

2. Although there have been attempts to do so, these are generally considered to have been unsuccessful (Hall and Paul, 2003; Collins et al., 2004; Menzies, 2004; Halpern and Pearl; 2005; VanderWeele, 2009e; Glymour et al., 2010).

and approaches to mediation and interaction from the causal inference literature employing counterfactuals. This counterfactual approach can be extended to address questions of mediation and interaction by extending the counterfactual notation to include either (a) contrary-to-fact settings of not only the primary exposure, but also the mediator, in the case of mediation or (b) include contrary-to-fact settings of two different exposures in the case of interaction. Methods for mediation and interaction have been developed in the social science literature as well, often relying only on a statistical framework without a formal framework for causality (see MacKinnon, 2008, for an excellent overview of these methods). The literature on mediation in causal inference has helped clarify the assumptions required to interpret the results of the methods in the social sciences causally. We will see this in Chapters 2 and 3. It has also helped to extend these methods to new settings. We will see this in Chapters 2–6. Finally, it has also provided a set of techniques—sometimes referred to as sensitivity analysis—that allow investigators to assess how robust conclusions are violations of assumptions required for a causal interpretation. This will be the focus of Chapter 3 but will arise again in Chapters 4 and 5.

In some ways, the causal inference literature on mediation is behind that which has developed in the social sciences insofar as methods designed for more complex settings such as multilevel or longitudinal data are available in the social sciences but not yet well developed within the causal inference literature. However, what the causal inference literature often does is to take methods from the social sciences and clarify what assumptions must be made for effect estimates to have a causal interpretation and to clarify what the interpretation of the effect estimates actually is. We will see this very clearly in Chapter 6, in which the complexities of longitudinal data make the methods, models, and effect estimates in the social science literature very difficult to interpret without a formal counterfactual framework. As we will discuss further in the final chapter of the book, there will likely continue to be a fruitful synthesis as ideas from the causal inference literature and methods from the social science literature come together and inform one another.

The book has been written to be a relatively accessible introduction to this counterfactual approach. When possible, concepts, ideas, definitions, and assumptions are described in intuitive terms, and methods are described in a way so as to be accessible for a reader familiar with only linear and logistic regression. When further statistical knowledge is presupposed by any particular section, the reader will be alerted to this, but in such instances it will be possible to skip over any of these more demanding sections and continue with the remainder of the book without impediment. A description of which chapters require which others, along with other comments about the organization of the book, is given in Section 1.4. For more technical readers, who may be interested in the formal technical details of the counterfactual framework, the theoretical justification of the methods, or the mathematical proofs of the results, a lengthy technical appendix has been provided at the end of the book with all definitions given more formally, and with all results stated formally and given with proof. The Appendix could almost be read by the technical or statistical reader as a stand-alone text. However, only the text of the book itself really provides the motivation and relevance of the results stated formally in the

Appendix. The Appendix is provided as a repository of results so that the methodologist interested in developing methods further can gain insight into the current state of the field, but its purpose is also to make clear how the counterfactual approach to causal inference can and has shed light on methods for mediation and interaction.

1.2. FORMS OF EXPLANATION AND TYPES OF MECHANISMS

When we explain a phenomenon, we put it with a particular context so that the phenomenon itself is better understood. Questions about explanation often begin with "Why...?" or "How...?" Causation itself is sometimes seen as a type of explanation. An outcome may be explained by reference to its cause. The cause may explain why the outcome came about. Many instances of explanation in science do in fact make reference to causation and can be thought of as a form of causal explanation. However, not all explanations are causal. Explanation may also make reference to mathematical or logical deduction or to human intention (cf. Lipton, 2009). We will briefly discuss forms of noncausal explanation in the final chapter of this book. However, until then, the focus of this book will be on *causal* forms of explanation.

With causation itself, a distinction is sometimes drawn between "type causation" and "token causation." Token causation involves statements of the form "X caused Y (in this particular instance)." Type causation involves statements of the form "X causes Y (in general)."Token causation explains the particular instance of some event or outcome by the particular instance of a cause. Type causation explains some of the instances of the outcome by the general presence of the cause. When used in actual data analysis, the methods from the potential outcomes causal inference literature are principally concerned with type causation—that is, of statements of the form "X causes Y." However, token causation and type causation are not unrelated. Once we have established type causation ("X causes Y in general"), we know there must be instances of token causation ("X in fact caused Y in this particular instance") even if we cannot identify the particular instances. In such cases, if X causes Y in general, then at least some instances of the outcome Y are explained by X. Likewise, statements of token causation ("X caused Y" in these particular cases) imply statements of type causation ("X causes Y" in general). Both token causation and type causation offer an explanation of this particular instance of the outcome, or of instances in general, by making reference to the cause of the outcome.

Specific explanations often push our understanding one level deeper. However, rarely do explanations provide a complete account of the phenomenon in view. An explanation often involves various other facts or phenomena that themselves may be in need of further explanation. We might explain the presence of an outcome by reference to a cause of the outcome. The explanation for the outcome is its cause. However, we might further ask the question "Why does the cause itself affect the outcome?" or "Why does the cause affect the outcome in certain instances and not in others?" These are questions of explanation not concerning simply the presence of the outcome, but rather concerning the cause–effect relationship itself. We seek an explanation not of the outcome, but of the phenomenon of causation. We want

to know why or how or when the cause affects the outcome. And it is these questions of explanation concerning the phenomenon of causation itself that are the focus of this book.

1.2.1. How an Effect Occurs—the Phenomenon of Mediation

One way to explain a cause–effect relationship is to explain how it is that the cause affects the outcome. We might describe the mechanism by which this occurs. Such a mechanism may be conceived of as an account of how the cause and certain initial states lead to particular final state (the outcome) through a process or a series of processes involving different intermediate stages. Appeal to such a mechanistic account may be given at more or less detailed levels.[3] The first half of this book is devoted to explanation of this type, explanation that makes reference to the processes by which a cause affects an outcome. More specifically, we will describe what we can and cannot learn about such mechanisms and processes from the statistical analysis of empirical data. The methods that are described will focus on the setting in which an investigator believes that a particular intermediate state or variable or exposure may be responsible for some of, or most of, the effect of the cause on the outcome. The methods attempt to assess what portion of the effect of the cause on the outcome is in fact operating through that particular intermediate and what portion might be through other mechanisms or pathways. The effect of the cause on the outcome that operates through the intermediate of interest is sometimes referred to as an indirect effect or mediated effect. The effect of the cause on the outcome that is not through the intermediate of interest is sometimes referred to as the direct effect or unmediated effect; however, it is important to keep in mind that such effects are direct only relative to the intermediate of interest; there will likely be other intermediates or mechanisms that account for other aspects of the effect of the cause on the outcome. The phenomenon whereby a cause affects an intermediate and the change in the intermediate goes on to affect the outcome is what is generally referred to as the phenomenon of "mediation," and the set of techniques by which a researcher assesses the relative magnitude of these direct and indirect effects is sometimes referred to as "mediation analysis." The intermediate itself is sometimes referred to as a "mediator."

Formal approaches to defining such direct and indirect effects and methods and sufficient assumptions for evaluating these effects and assessing mediation are described in Chapter 2 and developed further throughout Part I of this book. We will consider not only the setting of assessing the phenomenon of mediation for a single intermediate, but also more complex settings in which the outcome may be the time to an event (Chapter 4), or when multiple intermediates are of interest (Chapter 5), or when the cause and/or the intermediate may vary

3. See the work of Machamer et al. (2000) for a fuller description the notion of mechanism as employed in various scientific disciplines.

over time (Chapter 6). Throughout we will also focus on the strong assumptions needed to assess this phenomenon of mediation with empirical data. To help accommodate these strong assumptions, we will describe sensitivity analysis techniques (Chapter 3) that allow a researcher to assess how strong of a violation of the assumptions would be required to substantially alter conclusions concerning mediation.

1.2.2. For Whom an Effect Occurs—the Phenomenon of Interaction

Explaining how it is that a particular cause affects an outcome is one form of explanation concerning a cause–effect relationship. It is the form of explanation we will refer to as mediation. It is the form of explanation that will be the focus of Part I of this book. Another form of explanation concerning a cause–effect relationship might relate to explaining when, or for whom, a cause affects a particular outcome. We may observe that a cause affects an outcome in some instances but not in others, or that the cause may affect the outcome to differing extents in different contexts. Knowing for whom, or in which contexts, a cause affects an outcome provides deeper understanding of the cause–effect relationship and provides an explanation for why a cause may give rise to an outcome in some instances but not in others. The phenomenon whereby the effect of a cause on an outcome varies across individuals is sometimes referred to as "effect heterogeneity."

It will in general not be possible to know what would have happened to a particular individual both with and without the cause of interest, and so it will be difficult to assess effect heterogeneity at the individual level. Often, instead, to assess effect heterogeneity, sets of individuals are grouped together on the basis of some shared characteristic or exposure, and the effect of the cause of interest is assessed across the subgroups so defined to see if they differ. If we do find such heterogeneity, we might then think that the characteristic or exposure defining the subgroups somehow explains the effect heterogeneity. Perhaps the characteristic or exposure defining the subgroups somehow interacts with the primary cause of interest in its effects on the outcome. Or perhaps the characteristic or exposure defining the subgroups is somehow related to a different state or characteristic that interacts with the primary cause of interest in its effects on the outcome. The phenomenon whereby one exposure, characteristic, or state somehow alters the effect of a different exposure, characteristic, state, or cause is often referred to as one of "interaction" or "moderation" or sometimes "effect modification." The phenomenon helps explain why a particular cause sometimes affects the outcome and sometimes does not. It is this phenomenon of interaction that is the focus of Part II of this book. We will consider statistical methods that can be used with empirical data to assess this phenomenon (Chapter 9), different ways of conceptualizing the phenomenon of interaction (Chapters 9 and 10), how sensitive conclusions about interaction are to violations in assumptions (Chapter 11), and also how various study design considerations can help strengthen inferences concerning the phenomenon of interaction (Chapters 12 and 13).

1.2.3. Mechanisms in the Phenomena of Mediation, Interaction, and Interference

The methods in this book for *mediation* can help explain how it is that a cause affects an outcome. The methods in this book for *interaction* can help explain for whom a cause affects an outcome. However, these two types of explanation of a cause–effect relationship are in fact not entirely unrelated. Both phenomena can simultaneously be present, and attributing how much of an effect is due to mediation or interaction or both or neither when the two may simultaneously be present is the focus of Chapter 14, which presents a framework to unite the assessment of these two phenomena. The phenomena of mediation and interaction are also both related in different ways to the concept of a mechanism. The phenomenon of, and methods for, mediation are those that are most naturally thought of as assessing mechanisms—a delineation how the cause and certain initial states lead to a particular final state (the outcome) through a process or a series of processes involving different intermediate stages. However, as we will see in Chapter 10 of this book, assessing interaction can also shed light on mechanisms. More specifically, we might conceive of causation as consisting of different mechanisms that are each sufficient to bring about a particular outcome, with each mechanism itself requiring various conditions or causes or states to be set in motion. For a specific mechanism to operate, some set of such conditions is necessary, and each mechanism is itself sufficient for the outcome. This conceptualization of causation is sometimes referred to as a "sufficient cause" model of causation and appears in the philosophical (Mackie, 1965), epidemiologic (MacMahon and Pugh, 1967; Rothman, 1976), legal (Wright, 1988), and psychological (Cheng, 1997; Novick and Cheng, 2004) literatures. It will be discussed further in Chapter 10. In Chapter 10, we will also see that certain empirical tests for interaction can give insight into the mechanisms themselves. Specifically, we will see that in some cases we are able to test for the presence of a mechanism involving two or more specific causes. Thus it is not only the phenomenon of mediation that sheds light on mechanisms; the phenomenon of interaction can give insight into mechanisms as well. Discussion of the relationship between the sufficient cause conceptualization of causation and the counterfactual conceptualization can be found in Chapter 10 and elsewhere (Greenland and Poole, 1988; Greenland and Brumback, 2002; Flanders, 2006; VanderWeele and Hernán, 2006; VanderWeele and Robins, 2008; VanderWeele and Richardson, 2012; VanderWeele, 2012d). More detailed and formal accounts of the relationships between mediation, interaction, and mechanism within the counterfactual and sufficient cause frameworks are given elsewhere (Hafeman, 2008; VanderWeele, 2009f, 2011c, 2014; Suzuki et al., 2011) and are discussed in some further detail in Chapters 10 and 14 of this book.

We have, thus far, kept explanations of how an effect occurs, and for whom an effect occurs, distinct. However, this distinction blurs somewhat in the context of yet another phenomenon, that of "interference" or "spillover effects," which is the subject of Chapter 15 of this book. Interference is the phenomenon whereby the exposure or state of one individual can affect the outcome of another. In many settings, this sort of phenomenon does not arise: Whether one cancer patient

receives surgery or chemotherapy is not likely to affect the survival outcome of a different cancer patient. However, in other contexts interference is quite clear. Whether a person is vaccinated might well affect whether a family member or a friend is infected. This phenomenon of interference is common whenever an outcome depends upon social interactions between individuals. In such contexts, the question of "how?" and "for whom?" become blurred. The mechanism by which an effect occurs for one individual often involves the exposure or outcome of a different individual. Assessing and explaining causal effects in this context is challenging and often requires reference to both the phenomena of mediation and interaction to fully understand the effects that may be present. Indeed we will discuss in Chapter 15 that various relations hold between the phenomena of mediation and interaction on the one hand and interference on the other. We will consider in that chapter what we can learn about the mechanisms whereby causal effects arise when interference and spillover effects are present. Explanation in these contexts for one individual will in general require reference to the exposures and outcomes of other individuals.

1.3. MOTIVATIONS FOR ASSESSING MEDIATION, INTERACTION, AND INTERFERENCE

We have discussed in the previous section how the phenomena of mediation, interaction, and interference relate to the explanation of cause–effect relationships. The methods described in this book are focused on the empirical study of these phenomena. In addition to gaining insight into the explanation of causal effects, several theoretical and practical considerations also motivate the empirical study of mediation and interaction.

1.3.1. Motivations for Assessing Mediation

Methods for mediation help understand the mechanisms, pathways, and intermediates whereby a cause affects an outcome. There are a number of motivations for wanting to understand the phenomenon of mediation. In some instances, the motivation may principally be simply explanation and understanding. For example, in Chapter 2 we will consider an example from genetic epidemiology in which genetic variants were found to be associated with both smoking behavior and lung cancer. A question that arose in this context was whether the variants affected lung cancer only because they affected smoking and we know that smoking causes lung cancer or whether the variants affected lung cancer through pathways other than through smoking (Amos et al., 2008; Hung et al., 2008; Thorgeirsson et al. 2008). This question was of some scientific interest because 50 years earlier, Fisher (1958) had proposed that there might be a genetic variant that affected both smoking and lung cancer. We will return to this question and example in Chapter 2 and again later in Chapter 14 to see how methods for mediation can give insight into the pathways and mechanisms by which causal effects arise.

Empirically studying mediation can also help confirm and refute theory. For example, it has been repeatedly found that low socioeconomic status (SES) during

childhood is associated with adverse health outcomes later in life. However, there remains debate as to whether this is because low SES during childhood affects adult SES, which in turn affects adult health (a "social trajectory" model), or whether childhood SES affects adult health through pathways other than through adult SES (a "latent effects/sensitive period" model), or both. Understanding the role of adult SES in the association between childhood SES and health outcome can provide evidence in settling which of these theoretical models is better supported. We will return to this question in Chapter 5 and use empirical methods for mediation to evaluate some of the evidence for these different theories.

Another motivation sometimes given for the empirical study of mediation is to refine interventions. We may be in a setting in which we have established through a randomized trial that an intervention has a beneficial affect on average for the study population. We might be interested in further refining the intervention so as to increase the magnitude of the effect. This might be done by altering or improving components of the intervention that target a particular mechanism for the outcome. However, before proceeding with such refinements, it might be thought desirable to know whether, and the extent to which, the mechanism targeted is an important pathway from the intervention to the outcome. If the mechanism targeted explains a large portion of the effect, then refining the intervention further to target this mechanism may be desirable. However, if the mechanism is found not to be important, then it may be best to redirect efforts at refining the intervention elsewhere. The methods for mediation in Part I of this book can help assess the relative importance of various mechanisms.

Likewise, if an intervention has an effect on an outcome, then knowing something about the mechanisms and pathways by which these effects arise may allow for the discarding of components of an intervention which perhaps are not ultimately important for the outcome. As an example of this, in Chapter 3 we will consider a randomized trial of a cognitive therapy intervention (Strong et al., 2008) that was found to have a beneficial effect on depression symptoms. However, it was also noted that the intervention had an effect on the use of antidepressants: Those in the cognitive behavioral therapy group were more likely to use antidepressants during follow-up. This led to questions concerning whether the cognitive behavioral therapy intervention had a beneficial effect on depressive symptoms simply because it led to higher antidepressant use, or whether the intervention affected depressive symptoms through other pathways—for example, by changing the thought and behavioral patterns of the participants. If the intervention were only beneficial because of higher use of antidepressant, then the cognitive behavioral aspects of the intervention could perhaps be abandoned without much loss and a more cost-effective intervention just focusing on antidepressant adherence could be developed. Alternatively, it may have been the case that the intervention was effective both because of increased antidepressant use and because of cognitive behavioral changes. The methods for assessing mediation in Chapters 2 and 3 can again be useful in assessing the relative contribution of these various pathways, and we will return to this example again in Chapter 3.

Methods for mediation might also be of interest in settings in which an intervention is found not to have an effect on an outcome. By considering various

possible mechanisms and intermediates, it may be possible to assess whether the intervention did not affect the outcome because it failed to affect the intermediate, or whether the intervention did in fact affect the intermediate but rather the intermediate under consideration failed to change the outcome. Such knowledge may be useful in determining whether an intervention needs to be refined by better targeting a particular mechanism or intermediate, or whether the mechanism or intermediate that was targeted was in fact the wrong one because it had little effect on the outcome. It is also possible that an intervention might affect the outcome positively through one mechanism and negatively through a different mechanism. Assessing mediation can also help identify such settings.

We might also be interested in empirically assessing mediation because in some settings we may not be able to intervene on the primary exposure or cause of the outcome directly and so we might be interested in whether we can eliminate a detrimental effect of an exposure by intervening instead on some particular mechanism or intermediate. For example, in the context of the genetic epidemiology example above, we cannot intervene directly on the genetic variants, but we might be interested in how much of the effect of the genetic variants on lung cancer we could block if we could intervene to eliminate smoking. We will revisit this question in Chapter 2. As we will see there, the portion of the effect that operates through a particular mechanism (e.g., by changing smoking) and the portion of the effect that could be eliminated if we intervened on a mechanism (e.g., completely eliminated smoking) may be very different from one another, and different methods and different measures of effect are useful in these different contexts. These two measures to assess the portion of the effect that operates through a particular mechanism (e.g., by changing smoking) and the portion of the effect that could be eliminated if we intervened on a mechanism will diverge when both mediation and interaction are simultaneously present so that the exposure both affects and interacts with the mediator. The methods described in Chapter 2 can accommodate such settings with mediation and interaction simultaneously present, and additional refinements that help understand the respective roles of these two phenomena are further described in Chapters 7 and 14. Similar motivations concerning the effects of interventions on a mechanism when we cannot intervene directly on the exposure itself likewise arise in health disparities research. We may, for example, find differences in a health outcome across racial groups. We obviously cannot intervene on race, but we might be interested in the extent to which the health disparities across racial groups might be reduced or eliminated if we could intervene to equalize education levels across racial groups. We will return to such questions and methods to address them and their relation to mediation in Chapter 7.

It is also sometimes proposed that the study of mechanisms can make more plausible the claim that the exposure of interest does in fact cause the outcome in the first place (Hafeman and Schwartz, 2009). While this is likely the case with approaches from the basic sciences[4], it is less clear that this motivation and argument is reasonable when a researcher is relying only on the statistical analysis

4. See also Glennan (2009) for an overview of the broader philosophical position that evaluating mechanisms helps establish causality itself.

of data. As will be seen in the next chapter, the assumptions that need to be made to empirically assess mediation are in general much stronger than those that need to be made to assess overall causation of an exposure–outcome relationship. Thus in most cases, if we are unsure about overall causation, we will be unsure about mediation as well, and so methods for mediation will not in general lend additional credibility to the weaker claim of an exposure–outcome cause–effect relationship. An exception can occur if it is thought possible to evaluate causation for exposure–intermediate relationship and also for the intermediate–outcome relationship but not directly for the exposure–outcome relationship.

1.3.2. Motivations for Assessing Interaction

There are also a number of practical and theoretical considerations that motivate the study of interaction. One of the most prominent of these is that, in a number of settings, resources to implement interventions may be limited. It may not be possible to intervene on or treat an entire population. Resources may only be sufficient to treat a small fraction. If this is the case, then it may be important to identify the subgroups of individuals in which the intervention or treatment is likely to have the largest effect. As will be discussed in Chapter 9, methods for assessing additive interaction can help determine which subgroups would benefit most from treatment. For example, we will see there that the effects of asbestos exposure is far more harmful for smokers than for non-smokers and it might then be thought desirable, in removing asbestos from homes, to target the homes of smokers first. Even in settings in which resources are not limited and it is possible to intervene on everyone, it may be the case that a particular intervention is beneficial for some individuals and harmful for others. In such cases, it is very important to identify those groups for which treatment may be harmful and refrain from treating such persons. Techniques for assessing these so-called "qualitative" or "crossover" interactions will also be discussed in Chapter 9 and are useful in this regard. We will describe there, for example, a breast cancer treatment setting in which, for young patients under age 50 with low progesterone receptor levels, treatment without tamoxifen led to higher proportions who were disease-free at 3 years, but for all other groups (who were either older or had higher progesterone receptor levels, or both), treatment with tamoxifen led to higher proportions who were disease-free at 3 years. Here we would likely want to give young patients with low progesterone receptor levels the treatment without tamoxifen while giving others the treatment with tamoxifen. Other, more sophisticated methods also described in Chapter 9 can help identify groups of individuals, based on a large number of covariates, who would or would not benefit, or who would benefit to the greatest extent, from treatment.

Another reason sometimes given for empirically assessing interaction is that it may shed insight on the mechanisms of the outcome. We briefly mentioned above in Section 1.2.3 how this might be so. Chapter 10 is devoted to how the empirical study of interactions can shed insight into the mechanisms for the outcome. Yet another reason sometimes given for studying interaction is that leveraging interactions that may be present may in fact help increase power in testing for the overall

effect of an exposure on an outcome. In some settings, by jointly testing for a main effect and for an interaction simultaneously, it is possible to detect an overall effect when a test ignoring the interaction would otherwise not be able to detect the effect. It has been proposed that this may be especially important in the context of studying genetic variants when many variants are being tested and correction for such multiple testing reduces power, whereas allowing for the joint test may increase power to detect the effects. We will consider such methods for joint testing in Chapter 12.

As noted above, one of the motivations for studying interaction is to identify which subgroups would benefit most from intervention when resources are limited. However, in some settings it may not be possible to intervene directly on the primary exposure of interest, and one might instead be interested in which other covariates could be intervened upon to eliminate much or most of the effect of the primary exposure of interest. In these cases, methods for attributing effects to interactions, discussed in Chapter 9, can be useful in assessing this and identifying the most relevant covariates for intervention. For example, we will see there that in the genetic epidemiology example already mentioned above in Section 1.3.1 concerning genetic variants associated with lung cancer, although we cannot intervene directly on the genetic variants themselves, if we were able to eliminate smoking, this would in fact also eliminate almost all of the effect of the variants. Methods for attributing effects to interaction can help establish such results. These various motivations for empirically assessing interaction will all be discussed further, and illustrated in practical examples, in Part II of this book.

1.3.3. Motivations for Assessing Interference

There are likewise a number of motivations for studying interference or spillover effects. First, when spillover effects are present so that one person's exposure affects another's outcome, this can be very important in evaluating the cost-effectiveness of an intervention. If an intervention on one person benefits not only that person but others also through spillover, then ignoring the spillover effects will lead to an underestimate of the true cost-effectiveness of the intervention or program. Understanding interference and spillover can also help determine the design of an intervention program for a population and what proportion it is necessary to give treatment to in order to achieve a desired outcome for the population. For example, knowing how large the effect of one person's exposure on another's outcome is can help determine what proportion of a population would need to be vaccinated to substantially reduce disease in a population. Because the vaccination of one person may also protect unvaccinated persons (since the vaccinated person may not be infected and thus may not transmit the disease to others), it may be possible to substantially reduce infection in an entire population by only treating some smaller fraction of that population. The methods and approaches described in Chapter 15 can help assess this phenomenon, sometimes referred to as "herd immunity." The methods in Chapter 15 likewise can be useful in understanding the extent to which the effect of one person's exposure on another's outcome is because the exposure of the first person prevented or caused the outcome of the first person, which then

affected the second person and the extent to which the exposure of the first person affected the outcome of the second through other mechanisms. For example, a weight loss program that someone participates in might also affect the weight of his wife. This might be because the husband's weight loss program leads to weight loss for the husband, which also motivates the wife to lose weight; alternatively, even if the husband does not lose weight, information from the weight loss program may be passed from husband to wife, leading to weight loss for the wife. Being able to assess these different mechanisms for the spillover effect may be important in intervention design and refinement.

In Chapter 15 we will also discuss issues of spillover effects and contagion in social networks. When spillover effects are present so that the exposure of one individual affects the outcomes of others, using social network data can be helpful in identifying for which individuals it is best to target treatment so that the treatment effects have the maximum possible diffusion over the population. It may be that targeting persons who are more central in a social network will have larger effects on the population as a whole. Using social network data, it may also sometimes be possible to discern whether the effects on a population are maximized by targeting central individuals, or cliques of individuals, or by targeting unrelated individuals throughout the social network. Such considerations can likewise be important in intervention design. These various motivations for, and methods related to, the empirical study of interference and spillover effects will be described Part III of this book.

1.4. ORGANIZATION OF THIS BOOK

This book is divided into three parts. Part I concerns mediation, Part II concerns interaction, and Part III concerns interference and spillover effects and also the relation between mediation and interaction. Parts I and II can be read independently. The reader could start in either place. Part III assumes familiarity with at least the basic notions of mediation contained in Chapter 2 and the basic notions of interaction contained in Chapter 9.

The dependencies among the chapters is given Figure 1.1. In Part I on mediation, once Chapters 2 and 3 are read, then any of the other chapters in Part I (i.e., Chapters 4–8) can all be read independently of each other, in any order. In Part II on interaction, once Chapter 9 is read, then any of the other chapters in Part II (i.e., Chapters 10–13) can all be read independently of each other, in any order. Part III of the book concerns interference and the relation between mediation and interaction; it consists of Chapters 14–16. Once Chapter 2 on mediation and Chapter 9 on interaction have been read, the chapters of Part III (i.e., Chapters 14–16) can be read independently of each other, in any order.

The present chapter has provided a brief introduction to the relations between causation and explanation, different forms of causal explanation, and the motivations for empirically assessing mediation, interaction, and interference. Chapter 2 provides the basic introduction to the concepts of mediation and regression-based approaches for mediation. The chapter describes the assumptions that are needed

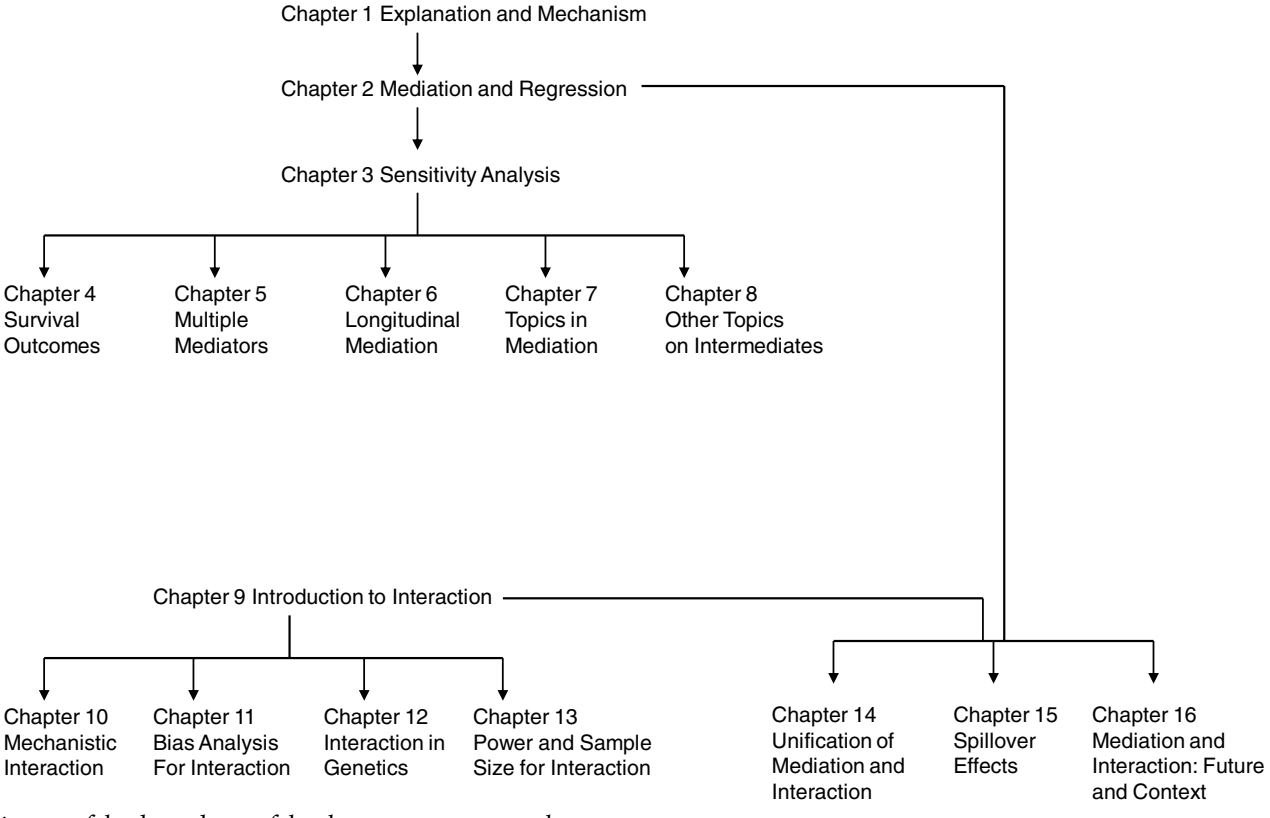

Figure 1.1 Diagram of the dependence of the chapters upon one another.

for a causal interpretation of the direct and indirect effect estimates, describes software to implement methods, and also gives some discussion of study design considerations in mediation analysis. Chapter 3 describes the use of sensitivity analysis to assess how robust one's conclusions are for total effects, direct effects, and mediated effects with regard to potential biases due to various forms of unmeasured confounding and measurement error in the variables. Chapter 4 considers similar methods and sensitivity analysis techniques for survival or time-to-event outcomes. Chapter 5 considers methods for handling multiple mediators simultaneously. Chapter 6 considers methods that can be employed to assess mediation when the exposure and/or mediator varies over time. Chapter 7 consists of a series of selected topics mostly concerned with the interpretation of effect estimates in mediation and also with alternative estimation and identification approaches to direct and indirect effects. Chapter 8 consists of a series of topics that are related to intermediates but do not constitute mediation per se (i.e., do not constitute the phenomenon whereby an exposure changes an intermediate and that change in the intermediate changes the outcome). Chapter 8 thus considers the topics of principal stratification, surrogate outcomes, instrumental variables, and Mendelian randomization. These topics are sometimes confused with mediation but in fact address other types question that are of interest in different contexts. These various other approaches are described in Chapter 8; and their relations with, and distinctions from, mediation are discussed.

In Part II of this book, Chapter 9 provides the basic introduction to interaction analysis and forms the foundation for the remaining chapters in Part II. Chapter 9 describes basic notions of and measures of interaction, the role of confounding in interpreting interaction analyses, how to present interaction analyses, other specific types of interaction such as mechanistic interaction and qualitative interaction, methods for identifying subgroups for which to target an intervention, and also methods for identifying secondary exposures that can amplify or eliminate the effects of the primary exposure if it is not possible to intervene directly on the primary exposure itself. Chapter 10 discusses different types of mechanistic interaction and how these relate to the sufficient cause framework briefly mentioned in Section 1.2.3. Discussion is given to how such mechanistic interaction (in which an outcome occurs if both of two exposures are present but not if just one or the other is present) is distinct from statistical measures of interaction described in Chapter 9, but how it is still sometimes possible to test empirically for such mechanistic interaction and what the limits of inferences concerning biology are with regard to such mechanistic interaction. Chapter 11 describes sensitivity analysis techniques to assess how robust one's conclusions about interaction are with regard to biases due to unmeasured confounding and measurement error. Chapter 12 considers some special issues that are relevant primarily to genetics and gene–gene and gene–environment interaction, but potentially also of interest in other contexts as well; such topics include the use of the case-only estimator to boost power to detect interaction, the use of joint main-effect and interaction-effect tests to boost power, and issues of multiple testing and how to correct for this. Chapter 13 describes power and sample size calculations for different types of additive, multiplicative, and mechanistic interaction in a variety of different study designs.

Part III of this book describes the relation between mediation and interaction and also provides an introduction to spillover effects. In Chapter 14, a decomposition is given that partitions a total effect of an exposure on outcome in the presence of an intermediate into four parts: that due to mediation alone, that due to interaction alone, that due to both mediation and interaction, and that due to neither mediation nor interaction. The decomposition makes clearer the respective roles of mediation and interaction and helps unify these phenomena. All of the other decompositions given in the book in both Parts I and II are in fact simply special cases of this four-way decomposition. Software code is also provided to implement this four-way decomposition. Chapter 15 provides an introduction to the phenomenon of interference or spillover effects, sometimes also referred to as "social interaction." Extensions to the counterfactual approach to causal inference allowing for such interference are described, and basic definitions of spillover effects are given. Methods to assess how much of a spillover effect is due to contagion versus other mechanisms are described; tests for specific forms of interference or spillover are related to the tests for mechanistic interaction; challenges in assessing spillover with multiple individuals per cluster and with observational data are described, and the discussion is extended further to the setting of social networks. Finally, Chapter 16 concludes the book with some discussion of the current state of methods for mediation and interaction and where further methodological development is perhaps most needed, and then it closes with discussion of some of the broader philosophical issues concerning explanation in causal inference.

2

Mediation: Introduction and Regression-Based Approaches

In the previous chapter we described the basic concepts, ideas, and motivations for mediation analysis. In this chapter we will describe a regression-based approach to mediation. The approach is based on the counterfactual or potential outcomes framework in causal inference (Neyman, 1923; Rubin, 1974, 1978, 1990; Robins, 1986; Pearl, 2009). Familiarity with this framework will not be presupposed, but it is this framework that will allow us to formalize and also generalize various approaches to mediation that have been used within epidemiology and the social sciences and allow us to incorporate interactions between the exposure and mediator. The chapter is organized as follows. The first section discusses the approach to mediation analysis sometimes referred to as the "product method" or "product-of-coefficients method" and made popular by Baron and Kenny (1986). The second section provides an introduction to the counterfactual approach that gives more general definitions of direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001; VanderWeele and Vansteelandt, 2009, 2010; Imai et al., 2010a,b). We describe how these can be estimated within a regression framework, allowing for exposure–mediator interaction, provided that certain no-confounding assumptions hold. In the following section, these no-confounding assumptions required for a causal interpretation of direct and indirect effect estimates are described in more detail. We then discuss how the approach also can be applied to binary outcomes and binary mediators and how the approach can be adapted to case–control studies that are popular in epidemiology but less common in the social sciences. The next section discusses the relationship between (a) the approach using the counterfactual framework and (b) other popular approaches to mediation analysis. The chapter continues with instructions for using macros in SAS, SPSS, and Stata to implement automatically this regression-based approach to mediation. The input and output of these procedures are described in detail. We then consider an example of such an analysis that comes from the genetics literature. In the context of this example, we also discuss two important measures that may arise in considering mediators: (i) the proportion mediated and (ii) the proportion of the effect eliminated by an intervention on the mediator. We then turn

our attention to study design considerations that are of importance when assessing mediation. For the interested reader, we then also describe the more formal and technical definitions of the effects from the counterfactual framework. We conclude with discussion of an alternative simulation-based approach to estimate direct and indirect effects and describe R code that can be used to implement this.

2.1. CLASSIC REGRESSION APPROACH TO MEDIATION ANALYSIS

The practice of mediation analysis has been highly influenced by the work of Baron and Kenny (1986). The causal diagram in Figure 2.1 was the one they used to conceptualize the role of a mediator variable. In this graph, which represents a simple mediation model, A denotes an exposure (or treatment) variable, M denotes the mediator, and Y denotes the outcome variable. In Section 2.6 we will discuss the criteria that Baron and Kenny (1986) proposed for mediation. Baron and Kenny also suggested a parametric approach to estimate direct and indirect effects, which will be the focus of this section. The approach is often simply referred to as the "Baron and Kenny approach." However, it had antecedents in the literature (Hyman, 1955; Alwin and Hauser, 1975; Judd and Kenny, 1981; Sobel, 1982) and is also more generally referred to as the "product method" or "product of coefficients method." Let A be the treatment, Y the outcome, M the mediator, and C the additional covariates. For the case of a continuous mediator and a continuous outcome, consider the following regression models:

$$\mathbb{E}(M|A=a, C=c) = \beta_0 + \beta_1 a + \beta_2' c \tag{2.1}$$

$$\mathbb{E}(Y|A=a, M=m, C=c) = \theta_0 + \theta_1 a + \theta_2 m + \theta_4' c \tag{2.2}$$

The original Baron and Kenny approach did not have covariates, but the same general approach applies with covariates (i.e., $\beta_2' c$ and $\theta_4' c$ were not included in the original models by the authors; here c is considered a vector and may contain multiple confounders). Baron and Kenny proposed that the direct effect be assessed by estimating θ_1; the indirect effect could be assessed by estimating $\beta_1\theta_2$. The direct effect is thus the coefficient of the exposure in the model for the outcome that includes the mediator as a covariate. The indirect effect is the coefficient of the exposure in the mediator model times the coefficient of the mediator in the outcome model. The direct effect can be conceived of as the treatment effect on the outcome at a fixed level of the mediator variable; this is different from the total effect, which represents simply the overall effect of exposure or treatment

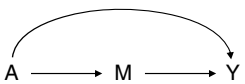

Figure 2.1 A simple mediational model with exposure A, mediator M, and outcome Y.

on the outcome. The indirect effect can be conceived of as the effect on the outcome of changes of the exposure which operate through mediator levels. For these expressions using regression coefficients to have a causal interpretation as direct and indirect effects, fairly strong assumptions about confounding need to be made. We will discuss these assumptions in detail in Section 2.3.

2.2. COUNTERFACTUAL APPROACH TO MEDIATION ANALYSIS: CONTINUOUS OUTCOMES

While the concept of mediation is theoretically appealing, the methods traditionally used to study mediation empirically have important limitations concerning their applicability in models with interactions or nonlinearities (Robins and Greenland, 1992; Pearl, 2001). Recent work on mediation analysis in the causal inference literature has emphasized the importance of articulating confounding control assumptions needed for a causal interpretation and has also extended definitions and results for direct and indirect effects to settings in which nonlinearities and interactions are present (Robins and Greenland, 1992; Pearl, 2001). In the next section we will describe the no-confounding assumptions needed for a causal interpretation in detail. In this section we will consider how the causal inference approach to mediation can be used to extend the Baron and Kenny approach to allow for exposure–mediator interaction.

The causal inference literature introduced counterfactual-based definitions of direct and indirect effects to formalize and generalize the approach in the social sciences. These counterfactual-based effects were formulated by Robins and Greenland (1992) and Pearl (2001). The effects defined in the causal inference literature are described intuitively below and more formally in Section 2.16. These counterfactual-based direct and indirect effects can be estimated from the regression models, provided that certain no-confounding (or confounding control) assumptions, also described below, hold and statistical models are correctly specified (VanderWeele and Vansteelandt, 2009, 2010). Suppose we have a continuous outcome and continuous mediator and that the mediator regression remains as in model (2.1) while the outcome regression now allows for an exposure–mediator interaction and takes the form

$$\mathbb{E}(Y|A=a,M=m,C=c)=\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c \qquad (\mathbf{2.3})$$

From models (2.1) and (2.3), what can be defined as the controlled direct effect (*CDE*), natural direct effect (*NDE*), and natural indirect effect (*NIE*) (described below), for a change in exposure from level a^* to level a can be estimated as follows:

$$\begin{aligned} CDE(m) &= (\theta_1+\theta_3 m)(a-a^*) \\ NDE &= (\theta_1+\theta_3\beta_0+\theta_3\beta_1 a^*+\theta_3\beta_2' c)(a-a^*) \\ NIE &= (\theta_2\beta_1+\theta_3\beta_1 a)(a-a^*) \end{aligned}$$

Note that these expressions are simply combinations of the coefficients from the regression models (2.1) and (2.3). If the exposure is binary, then $a = 1$ and $a^* = 0$. However, the formulae above can be used for a continuous exposure as well for any two levels of the continuous exposure a and a^*. The expressions above generalize those of Baron and Kenny to allow for interactions between the exposure and the mediator. Note that if interaction is absent, so that $\theta_3 = 0$, the controlled direct effect and the natural direct effect are equal to the direct effect obtained using Baron and Kenny approach (θ_1) times ($a - a^*$), and the natural indirect effect is equal to the indirect effect of the Baron and Kenny approach ($\theta_2\beta_1$) times ($a - a^*$). All of the other terms in the expressions above involve θ_3 and essentially are there to account for potential exposure–mediator interaction.

For simplicity, assume that the exposure A is binary and that we are comparing two exposure levels, $a = 1$ and $a^* = 0$. The controlled direct effect ($CDE(m)$) expresses how much the outcome would change on average if the mediator were fixed at level m uniformly in the population but the treatment were changed from level $a^* = 0$ to level $a = 1$. The natural direct effect (NDE) expresses how much the outcome would change if the exposure were set at level $a = 1$ versus level $a^* = 0$ but for each individual the mediator were kept at the level it would have taken, for that individual, in the absence of the exposure. This NDE captures what the effect of the exposure on the outcome would remain if we were to disable the pathway from the exposure to the mediator. The natural indirect effect (NIE), in contrast, expresses how much the outcome would change on average if the exposure were fixed at level $a = 1$ but the mediator were changed from the level it would take if $a^* = 0$ to the level it would take if $a = 1$. This NIE captures the effect of the exposure on the outcome that operates by changing the mediator. More formal definitions of these effects explicitly in terms of counterfactuals are given in the Appendix and in Section 2.16. While controlled direct effects are often of greater interest in policy evaluation because they consider what the effect of the exposure would be if we were to intervene on the mediator across the population (Pearl, 2001; Robins, 2003; VanderWeele, 2013a), natural direct and indirect effects may be of greater interest in evaluating the action of various mechanisms and the importance of different pathways and for effect decomposition (Robins, 2003; Joffe et al., 2007). We will return to these points later. These effects given above are conditional on the level of the covariates $C = c$. If C were set at its average level, $\mathbb{E}[C]$, in the expressions above, then we would obtain marginal effects on average for the entire population.

An important property of the natural indirect effect and the natural direct effect is that the total effect decomposes into the sum of these two effects. The total effect (TE) can be defined as how much the outcome would change overall for a change in the exposure from level $a^* = 0$ to level $a = 1$. In this counterfactual-based approach, the total effect decomposes into the natural direct and indirect effects even in models with interactions or nonlinearities (Pearl, 2001).

The expressions given above involving the coefficients of models (2.1) and (2.3) will be equal to the effects we have just discussed under certain no-confounding assumptions described in the next section. These no-confounding assumptions allow for a causal interpretation of the direct and indirect effects. These

assumptions are also needed for the expressions from the Baron and Kenny approach to have a causal interpretation.

2.3. ASSUMPTIONS ABOUT CONFOUNDING

Consider the relation between the variables in Figure 2.2 which might encompass a wide range of scenarios in mediation analysis. A careful study of this diagram will be useful in clearly formulating the confounding control assumptions for the direct and indirect causal effects of interest. The variables in the graph are: exposure (A), mediator (M), outcome (Y), and covariates $(C = (C_1, C_2))$, the latter of which could be exposure–outcome confounders (C_1) and mediator–outcome confounders (C_2). All the comments below will still hold if C_1 affects C_2 or if C_2 affects C_1.

For example, suppose we wanted to assess, for drug-addicted persons, whether a rehabilitation program, with methadone as treatment (A), leads to increased work activity (Y) and whether the level of illicit drug use (M) may mediate some of this effect. In this example, drug use may be a potential mediator (M) of the relationship between the methadone treatment (A) and the work activity outcome (Y). The level of methadone may affect drug use which may in turn affect work activity. In addressing this question and interpreting associations as causal effects, the investigator must think carefully about and try to control for variables that may be confounders of the exposure–outcome relationship (C_1) or of the mediator–outcome relationship (C_2). For example, there might be (a) social and biological factors, such as income and hypertension status (C_1), that affect decisions about the level of treatment (A) and the work activity outcome (Y) or (b) other factors, such as neighborhood of residence or alcohol consumption (C_2), that affect both the level of drug use (M) and the working activity outcome (Y).

For the effect estimates to have a causal interpretation, control must be made for the confounding variables. In order to be able to estimate the controlled direct effect using the formula above, two assumptions are needed. We must assume [assumption (A2.1)] no unmeasured confounding of the treatment–outcome relationship and [assumption (A2.2)] no unmeasured confounding of the mediator–outcome relationship. The measured covariates C included in the models need to suffice to control for confounding for these two relationships. The first of these

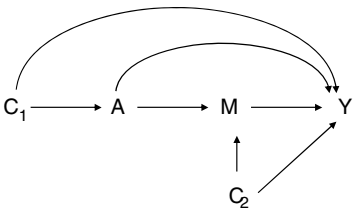

Figure 2.2 Mediation with exposure A, mediator M, and outcome Y with exposure-outcome confounders C_1 and mediator-outcome confounders C_2.

assumptions would be automatically satisfied if treatment were randomized, but even with randomized treatment the second assumption might not be satisfied. If we refer to the example above, to control for [assumption (A2.1)] confounding of the treatment–outcome relationship, the investigator must adjust for common causes of the exposure and the outcome—for example, information on income and hypertension status and any other treatment–outcome confounding variable (C_1) in the analysis. To control for [assumption (A2.2)] mediator–outcome confounding, the investigator must adjust for common causes of the mediator and the outcome—for example, alcohol consumption and neighborhood of residence or any other mediator–outcome confounding variable (C_2). In practice, both sets of covariates would simply be included in the overall set C for which adjustment is made; the investigator does not need to distinguish in this regression approach the treatment–outcome and the mediator–outcome confounding variables, but the collection of covariates must include both sets for estimates to have a causal interpretation.

The assumptions we have described are for controlled direct effects; the identification of natural direct and indirect effects uses these two assumptions above, along with two additional assumptions. In particular, for natural direct and indirect effects to be identified from the data, there must also be [assumption (A2.3)] no unmeasured confounding of the treatment–mediator relationship. Control must be made for variables that cause both the level of treatment and the level of the mediator. In the context of our example, hypertension is a factor that might influence the use of treatment as well as the level of drug use, and it would need to be controlled for in the analysis. This third assumption, like the first, would also be satisfied automatically if the treatment were randomized. Finally, for the natural direct effect and indirect effects to be identified, it also needs to be the case that [assumption (A2.4)] there is no mediator–outcome confounder that is affected by the exposure (i.e., no arrow from A to C_2 in Figure 2.2). This will often be a strong assumption. It essentially requires that there is nothing on the pathway from the exposure to the mediator that also affects the outcome. It may be more plausible if the mediator occurs shortly after the exposure (VanderWeele and Vansteelandt, 2009). It would, however, be violated in Figure 2.3 because the variable L affects both the mediator and the outcome and is itself affected by the exposure. In Chapter 5 we consider approaches for assessing pathways when this fourth assumption is violated. In Chapter 3 we will consider sensitivity analysis techniques for violations of the other assumptions. These sensitivity analysis techniques can help an investigator to assess how strongly an unmeasured confounding variable must be to substantially change results and how robust estimates are to potential violations of the assumptions.

It should be noted that assumptions (A2.1), (A2.2), and (A2.3) also require an assumption of temporal ordering. This assumption of temporal ordering is implicitly or explicitly present in various approaches to mediation analysis (Cole and Maxwell, 2003). In particular, the assumption of no unmeasured confounding of the treatment–outcome relationship implicitly assumes that the treatment

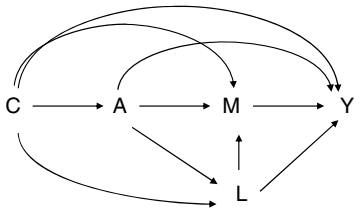

Figure 2.3 An example of a mediator-outcome confounder L that is affected by the exposure A.

temporally precedes the outcome. The assumption of no unmeasured confounding of the mediator–outcome relationship implicitly assumes that mediator precedes temporally the outcome. Finally, the assumption of no treatment–mediator confounding implicitly assumes that the treatment precedes the mediator. Formally, the no-unmeasured-confounding assumptions require that associations reflect causal effects; if the temporal ordering assumptions were not satisfied, then neither would the no-unmeasured-confounding-assumptions since associations would not represent causal effects. We will return to these issues of temporal ordering and questions of study design in mediation analysis in Section 2.15.

In summary, controlled direct effects require [assumption (A2.1)] no unmeasured treatment–outcome confounding and [assumption (A2.2)] no unmeasured mediator–outcome confounding. Natural direct and indirect effects require these assumptions and also no unmeasured treatment–mediator confounding [assumption (A2.3)] and no mediator–outcome confounder affected by treatment [assumption (A2.4)]. It is important to note that randomizing the treatment is not enough to rule out confounding issues in mediation analysis. This is because randomization of the treatment rules out the problem of treatment–outcome and treatment–mediator confounding but does not guarantee that the assumption of no confounding of mediator–outcome relationship holds. This is because even if the treatment is randomized, the mediator generally will not be. This was pointed out by Judd and Kenny (1981), MacKinnon (2008), and James and Brett (1984) but unfortunately not mentioned in the popular paper by Baron and Kenny (1986). If there are confounders of the mediator–outcome relationship for which control has not been made, then direct and indirect effect estimates will not have a causal interpretation; they will be biased. Unmeasured mediator–outcome confounding will not bias estimates of the total effect but will bias estimates of direct and indirect effects. This is true for the controlled direct effect and natural direct and indirect effects described above and also for the effects described by Baron and Kenny. Investigators should think more carefully about and collect data on and control for such mediator–outcome confounding variables when mediation analysis is of interest. If the investigator is aware that unmeasured confounding may be an issue in his or her study, sensitivity analyses (VanderWeele, 2010a; Imai et al., 2010a) should be implemented. Sensitivity analysis will be the topic of the next chapter. When the assumptions do not hold, biases can sometimes be quite extreme as will be demonstrated in some of the examples in the next chapter.

2.4. BINARY AND COUNT OUTCOMES

2.4.1. Binary Outcomes and Continuous Mediators

We have thus far considered only the case in which both outcome and mediator are continuous. The results can be extended to cases in which one or both of the mediator and outcome variables are binary.

For example, when the outcome is binary and mediator is continuous, the model for the mediator could still be taken as the linear regression model in (2.1) and the outcome could be modeled via a logistic regression

$$\text{logit}[P(Y=1|A=a,M=m,C=c)] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \quad \textbf{(2.4)}$$

For this case, provided that the outcome is relatively rare (less than 10% is often used as a cut-off) and that confounding assumptions (A2.1)–(A2.4) hold, and the error term in the linear regression model in (2.1) is normally distributed with variance σ^2, the natural direct and indirect effects on the odds ratio scale are given by (VanderWeele and Vansteelandt, 2010)

$$\begin{aligned} OR^{CDE}(m) &= \exp\{(\theta_1 + \theta_3 m)(a - a^*)\} \\ OR^{NDE} &= \exp\{(\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2' C + \theta_3\theta_2\sigma^2)(a - a^*) \\ &\quad + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\} \\ OR^{NIE} &= \exp\{(\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)\} \end{aligned}$$

With these odds ratios, the total effect is equal to the product of the natural direct and indirect effects (rather than the sum).

The direct and indirect effects given above will apply if the outcome is rare but will be biased if the outcome is common and logistic regression is used to model the outcome. We discuss this point further in Section 2.6. If the outcome is common, then the investigator can estimate the causal effect by running a log-linear model instead of the logistic regression, and the formulas for direct and indirect effect given above will then apply and have a risk ratio interpretation.

Standard errors for the expressions above can be obtained using the delta method [given in Valeri and VanderWeele (2013) or in the Appendix] or by using bootstrapping techniques. The software described below in Sections 2.7–2.9 will calculate these automatically. The assumption that the error term in the mediator regression (2.1) is normally distributed in fact can be dropped for the natural indirect effect and the controlled direct effect, but it can only be dropped for the natural direct effect if there is no exposure–mediator interaction—that is, if $\theta_3 = 0$ (Tchetgen Tchetgen, 2013).

2.4.2. Count Outcomes and Continuous Mediators

Similar formulae for direct and indirect effects along with their standard errors also extend to count outcomes when using Poisson or negative binomial models (Valeri and VanderWeele, 2013) and the SAS, SPSS, and Stata macros, described below,

will implement these as well. For example, if the mediator follows the linear regression model in (2.1) and the outcome follows either a Poisson model or a negative binomial model with the mean of Y conditional on the exposure A, mediator M, and covariates C given by $\exp(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c)$, then the controlled direct effect and the natural direct and indirect effects on the rate ratio scale will again be given by the same expressions as above, that is,

$$\begin{aligned} RR^{CDE}(m) &= \exp\{(\theta_1+\theta_3 m)(a-a^*)\} \\ RR^{NDE} &= \exp\{(\theta_1+\theta_3\beta_0+\theta_3\beta_1 a^*+\theta_3\beta_2' C+\theta_3\theta_2\sigma^2)(a-a^*) \\ &\quad +0.5\theta_3^2\sigma^2(a^2-a^{*2})\} \\ RR^{NIE} &= \exp\{(\theta_2\beta_1+\theta_3\beta_1 a)(a-a^*)\} \end{aligned}$$

and the standard errors for these effects are likewise given by the formulas provided in the Appendix. Again the macros will implement this automatically. Wang and Albert (2012) also describe methods for estimating natural direct and indirect effects from zero-inflated Poisson and negative binomial models for the outcome.

2.4.3. Case–Control Designs

In many epidemiologic studies with a binary outcome, a case–control design is used. In this design, cases (those with the outcome) are oversampled relative to controls (those without the outcome, or a sample from the study base). However, provided that inclusion in the sample depends only on the outcome (and not on the exposure conditional on the outcome), odds ratios relating the outcome to covariates can still be calculated that correspond to what one would have obtained in a cohort study. However, for the purposes of mediation, as noted above, not just one, but two, regressions are involved: one for the mediator and one for the outcome. Although the outcome regression parameters for Y that are required for the direct and indirect effects would be consistently estimated in a logistic regression, the case–control study design needs to be explicitly taken into account for the mediator regression. This is because, in the case–control design, the oversampling takes place with regard to the outcome Y, but a regression is now being fit for a different outcome, namely M. If the outcome Y is rare (which is often the motivation for using a case–control study design to begin with), then the approach described above to estimate direct and indirect effects can still be employed by fitting the outcome model for Y for the entire sample, but fitting the mediator model for M just among the control subjects. If the outcome is rare, then this regression will approximate what one would have obtained for the mediator regression in a cohort study. The coefficients from the two models can then be combined as above to obtain direct and indirect effects. The macros described below will also implement this approach involving case–control data automatically when specified to do so. Alternatively, a weighting approach can be used with case–control data; see VanderWeele and Vansteelandt (2010) for further details.

2.5. BINARY MEDIATORS

A similar approach, allowing for exposure–mediator interaction, also works with binary mediators. Suppose that the mediator is binary and the outcome is continuous and that the following models fit the observed data:

$$\mathbb{E}[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M=1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

If the covariates C satisfy no-confounding assumptions (A2.1)–(A2.4) above, then average controlled direct effect and the average natural direct and indirect effects on the outcome difference scale are given by

$$CDE(m) = (\theta_1 + \theta_3 m)(a - a^*)$$
$$NDE = \{\theta_1(a-a^*)\} + \{\theta_3(a-a^*)\}\frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1+\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}$$
$$NIE = (\theta_2 + \theta_3 a)\left\{\frac{\exp[\beta_0 + \beta_1 a + \beta_2' c]}{1+\exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1+\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}\right\}$$

The expressions for the direct and indirect effects are once again simply combinations of the coefficients from the two regressions above. These effects were derived in Valeri and VanderWeele (2013), and standard errors are also given there and in the Appendix. The macros described below will implement this approach and estimate standard errors and confidence intervals automatically.

Similarly, suppose that both the mediator and the outcome are binary. Suppose that the outcome is rare and that the following models are fit to the observed data:

$$\text{logit}\{P(Y=1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M=1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

If the covariates C satisfied assumptions (A2.1)–(A2.4) above, then the conditional controlled direct effect and natural direct and indirect effects on the odds ratio scale would be given by

$$OR^{CDE}(m) = (\theta_1 + \theta_3 m)(a - a^*)$$
$$OR^{NDE} = \frac{\exp(\theta_1 a)\{1+\exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}{\exp(\theta_1 a^*)\{1+\exp(\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c)\}}$$
$$OR^{NIE} = \frac{\{1+\exp(\beta_0 + \beta_1 a^* + \beta_2' c)\}\{1+\exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2' c)\}}{\{1+\exp(\beta_0 + \beta_1 a + \beta_2' c)\}\{1+\exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}$$

These expressions apply also if the outcome is not rare and log-linear rather than logistic models are fit to the data; the expressions are then for direct and indirect effect risk ratios rather than odds ratios. The mediator does not need to be rare for these expression to apply. Once again, derivations and standard errors for these are given in Valeri and VanderWeele (2013), but the macros described below will implement this approach and give estimates and confidence intervals automatically.

2.6. COMPARISON OF APPROACHES: PRODUCT-OF-COEFFICIENT AND DIFFERENCE METHODS

The counterfactual approach to mediation analysis displays all its power and flexibility when the causal relationships under study are complex and the investigator needs to depart from simple linear models and allow for nonlinearities and interactions. In this section we describe some of the advantages of employing the counterfactual framework to causal mediation that we presented in the previous sections by comparing it to other popular methods to address questions of mediation. In this comparison we will focus on the so-called product-of-coefficients method, the difference method, and the MacArthur approach (Kraemer et al., 2008). We first describe the traditional statistical approaches used in the social sciences and epidemiology, and we then discuss what the counterfactual approach contributes over and above them and comment on the relation between the various approaches.

2.6.1. Traditional Approaches to Mediation Analysis

Modern approaches to mediation have been inspired by the work of the Wright (1921), who developed the path analysis method. Path analysis is now viewed as a special case of structural equation modeling (SEM). Structural equations methods allow for the estimation of direct and indirect effects by modeling covariance and correlation matrices. Most mediation analyses in psychological studies have been conducted using the SEM approach (Baron and Kenny, 1986; Judd and Kenny, 1981; MacKinnon, 2008). Methods to improve estimation and inferential procedures for SEM-based mediation analyses have continued to develop (e.g., MacKinnon, 2008; Sobel, 1982). Structural equation models are often criticized for not adequately addressing issues of confounding/endogeneity in inferring causal relationships. However, if such issues of confounding are adequately addressed by including all relevant confounders (as described in detail above) in the structural equation model, then the SEM approach can be a useful tool. The counterfactual approach has placed strong emphasis on the no-confounding assumptions and conceptual definitions of causal effects (cf. Robins and Greenland, 1992; Pearl, 2001; Cole and Hernán, 2002; Glynn, 2012); recently, a number of authors have used the counterfactual framework to translate the SEM approach within the counterfactual framework (e.g., VanderWeele and Vansteelandt, 2009; Imai et al., 2010a; Pearl, 2011).

A special case of the SEM approach is that with only one exposure variable, one mediator, and one outcome of interest, as considered in the much-cited paper of Baron and Kenny (1986). According to Baron and Kenny, the following criteria need to be satisfied for a variable to be considered a mediator: (i) The exposure variable should be associated with the mediator, (ii) in the model for the outcome that includes the exposure and mediator, the mediator should be associated with the outcome, (iii) in the model for the outcome that includes only the exposure, the exposure should be associated with the outcome, and (iv) when controlling for the

mediator, the association between the exposure and outcome should be reduced, with the strongest demonstration of mediation occurring when the path from the exposure to the outcome variable, when controlling for the mediator, is zero.

While requirements (i) and (ii) have generally been accepted as important for establishing mediation, requirement (iii) has been critiqued by many scholars (MacKinnon, 2008): The relationship between A and Y need not be statistically significant for M to be a mediator. The reason for this is that the effect of A on Y may be zero or close to zero when direct and mediated effects have opposite signs. This phenomenon is sometimes called inconsistent mediation. Requirement (iv) is also not necessary because mediation can be partial or complete. When mediation is complete, after controlling for M (and for confounding), the direct path from A to Y would be zero. When mediation is partial, the path from A to Y can still be of substantial magnitude, but the effect should be reduced if mediation is indeed present. This four-step approach is also often problematic when used in practice because conditions (i)–(iv) are often assessed using significance testing and associations can of course be present even if not "statistically significant." The likelihood of getting the correct answer from significance testing for all four conditions will sometimes be quite small.

Among traditional SEM methods that give *estimates* of direct and indirect effects, the product method and the difference method are used most commonly. Assume a simple mediation model with no exposure–mediator interaction. The rationale behind the product method is that mediation depends on the extent to which the exposure A changes the mediator M, β_1 from equation (2.1), and the extent to which the mediator affects the outcome Y, θ_2 from equation (2.2). The product method estimator of the indirect effect is then simply $\beta_1\theta_2$. Sobel (1982) proposed a test for a mediated effect from the product method estimator.

The difference method approach is implemented by fitting an outcome model with the mediator as in equation (2.2) and also an outcome model with no mediator:

$$\mathbb{E}[Y|A=a, C=c] = \theta_0^\dagger + \theta_1^\dagger a + \theta_4^\dagger c \qquad \textbf{(2.5)}$$

The value of the mediated or indirect effect is then estimated by taking the difference in the coefficients from equations (2.5) and (2.2), $\theta_1^\dagger - \theta_1$; this corresponds to the reduction in the independent variable effect on the dependent variable when adjustment is made for the mediator. The difference method is used more commonly in epidemiology, whereas the product method is used more commonly in the social sciences. The difference method in epidemiology was described in some detail in textbook form as early as Susser (1973) and is still commonly employed today, often ignoring the concerns about mediator–outcome confounding discussed above.

The product method and the difference method are, in some models, closely related. For a continuous outcome on the difference scale, the product method and the difference method will in fact coincide. The algebraic equivalence of the indirect effect using the product method, $\beta_1\theta_2$, and the difference method, $\theta_1^\dagger - \theta_1$, was shown by MacKinnon et al. (1995) for ordinary least squares in linear models with

continuous outcomes and discussed also in Alwin and Hauser (1975). The product method and difference method diverge, however, when using a binary outcome and logistic regression (MacKinnon and Dwyer, 1993), a point to which we return below.

When mediation models include an exposure–mediator interaction term in the outcome regression, this is a particular case or a variant of what is sometimes referred to as "moderated mediation" (James and Brett, 1984; Preacher et al., 2007). Moderated mediation considers the case in which a covariate moderates the mediated effect (cf. MacKinnon, 2007). When the treatment itself is the moderator [as considered in Preacher et al. (2007)], the effect of the mediator is allowed to vary by treatment status; or, conceived of in another way, the effect of treatment is allowed to vary with (i.e., it interacts with) the mediator. In this setting, Preacher et al. (2007) derived an indirect effect estimator in the context of moderated mediation using the product method. However, in Preacher et al. (2007) this indirect effect does not sum with what they define as the direct effect to give a total effect, whereas the natural direct effect and the natural indirect effect, given above, do have this decomposition property of summing to a total effect. This is important in assessing what portion of the total effect is due to mediation.

The MacArthur approach (Kraemer et al., 2008) gives criteria somewhat different than that of Baron and Kenny in assessing mediation and allows also for assessing exposure–mediator interactions. This approach to mediation analysis is based on the assumption that temporal antecedence and association are necessary (but not sufficient) for a causal relationship. The approach allows for nonlinear relations among variables to qualify as mediation as long as there is a relationship between the exposure A and the mediator M. In particular, it is proposed, first, that if there is no association between A and M, if M precedes A, and if the $A \times M$ interaction is significant, then the variable M is to be considered as a moderator rather than a mediator. Second, for M to be a mediator for the effect of A on outcome Y, A should precede M and M should precede Y, the variables A and M should be correlated, and either the main effect of M on the outcome or the $A \times M$ interaction should be significant. These traditional approaches have generally not explicated the confounding assumptions, described above in Section 2.3, needed for a causal interpretation.

2.6.2. Comparison of Traditional Approaches with the Counterfactual Approach when There Are Interactions and Nonlinearities

One of the chief advantages of the counterfactual approach to mediation analysis is that it allows for the decomposition of a total effect into a direct effect and an indirect effect even when there are interactions and nonlinearities. As noted above, some of the statistical approaches, such as that of Preacher et al. (2007) or Kraemer et al. (2008), allow one to assess mediation even when there is exposure–mediator interaction. In fact, the indirect effect of Preacher et al. (2007) for continuous outcomes when there is an exposure–mediator interaction is equivalent to the one given in Section 2.2. However, neither Preacher et al. (2007) nor Kraemer et al. (2008) give a definition of a direct effect such that

the sum of the direct and indirect effects equals a total effect. Such a decomposition is important in assessing the relative contributions of the pathways through and not through the mediator. The counterfactual approach provides a general approach to do effect decomposition irrespective of the statistical model and irrespective of possible interactions. The counterfactual approach coincides with the criteria for mediation of the MacArthur approach (Kraemer et al., 2008) but provides actual direct and indirect effect estimates that combine to a total effect and makes clear the no-unmeasured-confounding assumptions needed for a causal interpretation.

The counterfactual approach also helps in understanding mediation with binary outcomes and binary mediators. As noted above, with a binary outcome and logistic regression, the product method and difference method give different results (MacKinnon and Dwyer, 1993). In fact, neither one in general will be equal to an estimate of an indirect effect with a causal interpretation (VanderWeele and Vansteelandt, 2010). VanderWeele and Vansteelandt (2010) did, however, show that when there is no exposure–mediator interaction (and when all the no confounding assumptions hold), the product method and difference method will be approximately equivalent when the outcome is rare, and both of them will then be equal to the natural indirect effect, once the product or difference method estimator is exponentiated. The problem with dichotomous outcomes arises when the outcome is common and has to do with the fact that logistic regression uses the odds ratio which is a measure that is "noncollapsible" (Greenland et al., 1999). Viewed intuitively, the problem occurs because when the outcome is common, the odds ratio does not approximate the risk ratio, and the extent of this lack of approximation can vary with the other covariates in the models. With a common outcome, the odds ratios with the mediator in the model versus without the mediator in the model are thus not directly comparable, and so the difference method essentially breaks down.

The problem arises because as we add covariates to the logistic regression model (even if these are not confounders), the coefficients tend to increase in magnitude (the coefficients of the exposure with different sets of covariates in the model are thus not comparable). What can happen then is if we add the mediator to the outcome model, then the coefficient of the exposure in the logistic regression may go up somewhat because of the additional variable (cf. Robinson and Jewell, 1991) but down somewhat because of mediation. It might then look like the coefficient of the exposure does not change at all even though there is in fact mediation. We would then draw the wrong conclusion from the difference method. In fact, because of this noncollapsibility of odds ratios, it can be shown that, with logistic regression, the difference method is conservative for mediation. That is to say, if one uses the difference method and the confounding assumptions hold, the difference method will in general underestimate the indirect effect when used with logistic regression (Jiang and VanderWeele, 2014). Thus if the difference method with logistic regression indicates the presence of a mediated effect, then there is in fact evidence for a mediated effect; however, if the difference method does not indicate a nonzero estimate of the indirect effect, this does not indicate that there is no mediation;

there may still be mediation; the difference method does not allow one to draw conclusions in this case because the difference method is conservative.

The risk ratio does not suffer this problem, and it is for this reason that we propose using a log-linear model when the outcome is common. The problem does not arise with log-linear models and risk ratios even if the outcome is common. The problem also effectively goes away for logistic regression when the outcome is rare since then the odds ratios approximate risk ratios. The counterfactual approach moreover also allows us to define and estimate direct and indirect effects when the outcome is binary and an exposure–mediator interaction is present. The counterfactual approach provides a versatile framework to derive direct and indirect effects and to do effect decomposition even with binary variables and nonlinear models.

As is perhaps now clear from this discussion, the traditional statistical approach and the counterfactual approach to mediation will in some settings coincide. For linear models and log-linear models, they will coincide when there is no exposure–mediator interaction; for logistic models, they will coincide when there is no exposure–mediator interaction and when the outcome is rare (VanderWeele and Vansteelandt, 2009, 2010). Thus, before an investigator proceeds with one of the traditional approaches (the product method or difference method), he or she should (i) consider whether control has been made for exposure–outcome confounders, mediator–outcome confounders, and exposure–mediator confounders and (ii) consider whether exposure–mediator interaction might be present; and if the outcome is binary and logistic regression is used, check whether the outcome is rare. If the no-unmeasured-confounding conditions are satisfied, there is no interaction, and the outcome is rare if logistic regression is used, then proceeding with the traditional statistical approaches is fine. If there are exposure–mediator interactions, then it is still possible to estimate direct and indirect effects using the approach described in Section 2.2. We consider the issue of including exposure–mediator interaction further in Section 2.12. If the outcome is common, a log-linear model can be used. If there are confounders of the exposure–outcome, mediator–outcome, or exposure–mediator relationship, then, to the extent possible, these should be controlled for in the models; otherwise, sensitivity analysis techniques (VanderWeele, 2010; Imai et al., 2010a) can be used, as described in the next chapter. Importantly, these no-confounding assumptions are required not only for the counterfactual approach described above but also for the product method and difference method estimators for these estimators to be interpreted causally. The assumptions are often not acknowledged with the product and difference method estimators but the assumptions are still being made, even if not acknowledged.

As a final point of discussion, we note that even in the presence of interaction and nonlinearities, the product method may be used to test for mediation in some cases in which the estimates are not themselves interpretable as estimates of an indirect effect. In other words, to test for mediation, we can test for whether the product of the coefficients is nonzero even if this product is not equal to a causal indirect effect measure. For example, with logistic model with common outcome, the product method estimates will not in general have a causal interpretation as a

natural indirect effect. It is nonetheless the case that although the product-method estimator is not itself a measure of an indirect effect, the product method still gives a valid test for the presence of a mediated effect, provided that the confounding control assumptions (A2.1)–(A2.4) in Section 2.4 hold and that the models are correctly specified (VanderWeele, 2011b; see Appendix). The intuition is that even if the product of the coefficients is not equal to a causal indirect effect, if the product is nonzero, then there must be an effect of the exposure on the mediator and an effect of the mediator on the outcome, and under the confounding control assumptions (A2.1)–(A2.4), this would also imply the presence of a natural indirect effect. Thus, the product-method approach can still be useful in *testing* for mediation even when there are interactions and nonlinearities. As noted above, under the confounding control assumptions (A2.1)–(A2.4) the difference method provides a conservative test (conclusions can only be drawn in one direction); the product method provides a valid test more generally. For *estimation*, however, and for decomposing a total effect into a direct and indirect effect, rather than just testing, methods from the counterfactual approach such as those described above can be employed.

2.7. DESCRIPTION OF THE SAS MACRO

Macros have been designed to enable the investigator to easily implement mediation analysis in the presence of exposure–mediator interaction (Valeri and VanderWeele, 2013) accounting for different types of outcomes (normal, dichotomous-logistic or dichotomous log-linear, Poisson, negative binomial) and mediators of interest (continuous or dichotomous). The logit link for dichotomous outcomes should only be used if the outcome is rare. If the outcome is not rare, then the log link can be used (though the outcome model may not always converge). When using a linear model, the direct and indirect effects are given on the outcome difference scale. When using a log link, the effects are on the risk ratio scale; when using a logit link, they are on the odds ratio scale; and when using a Poisson or negative binomial model, they are on the rate ratio scale. The macros for SAS, SPSS, and Stata all provide estimates, and confidence intervals for the direct and indirect effects were previously defined. The estimates assume that the model assumptions are correct and the confounding control assumptions discussed in the previous sections hold. In this section we will provide a description of the SAS macro; we will then provide similar abbreviated descriptions of the SPSS macro and the Stata command. After that we will give a short hypothetical example illustrating the macro output and then give an empirical example from genetic epidemiology where these methods were used. A reader who intends to only use SAS could skip the sections on SPSS and Stata and go directly to the section containing the hypothetical example. Users of SPSS and Stata would be best at least skimming the SAS section as some of the macro options are described in greater detail in this section than in the following ones. In Section 2.18 we also discuss R code that can be used to estimate direct and indirect effects.

2.7.1. Basic SAS Macro

The SAS macro has been developed using version 9.2 of SAS. In order to implement mediation analysis via the *mediation macro* in SAS, the investigator first opens a new SAS session and inputs the data which has to include the outcome, treatment, and mediator variables. The investigator inputs also the covariates to be adjusted for in the model. Macro activation requires that the investigator then save the macro script and input information in the statement

```
%mediation(data= , yvar= , avar= , mvar= , cvar= , a0= ,
   a1= , m= , nc= , yreg= , mreg= , interaction= )
run;
```

First one inputs the name of the dataset saved in the working directory (data=, then the name of the outcome variable (*yvar*=), the treatment variable (*avar*=), the mediator variable (*mvar*=), and the other covariates (*cvar*=). Categorical variables need to be coded as a series of dummy variables before being entered as covariates. The macro *dumvar* from MCHP SAS Macros, for example, can be used for this purpose. Then the investigator needs to specify the baseline level of the exposure a^* ($a0$=), the new exposure level a ($a1$=). If the mediator is binary, then $a1 = 1$ and $a0 = 0$. The investigator must also specify (a) the level of mediator m at which the controlled direct effect is to be estimated and (b) the number of covariates to be used (*nc*=). When no covariates are entered, then the user still needs to write the command *cvar*= and needs to specify *nc*=, but inputs nothing after the equal sign. The user must also specify which types of regression have to be implemented. In particular, for the outcome model, either linear, logistic, log-linear, poisson, or negbin can be specified (*yreg*=). When using a log-linear model for a binary outcome, it is good to check the output to make sure that the model has in fact converged. For the mediator model, either linear or logistic regressions are allowed (*mreg*=). Finally, the analyst needs to specify whether an exposure–mediator interaction is present (*interaction*= true or false).

The software provides the following output: First the regression output for outcome and mediator models is provided. The output in the SAS macro is derived from the procedures of proc reg when the variable is continuous, proc logistic when the variable is binary. When the outcome is specified as Poisson, negative binomial, or log-linear, the procedure proc genmod is employed. If the dataset contains missing data, the software implements a complete case only analysis. A table with direct and indirect effects together with total effects follows. The effects are reported for the mean level of the covariates C. The table contains standard errors, along with confidence intervals for each effect.

2.7.2. Other Options in the SAS Macro

The reduced output is the default option. The table will just display controlled direct effect, natural direct effect, natural indirect effect, and total effect described above. When the option *output=full* is used, both conditional effects and effects

evaluated at the mean covariate levels are shown. For a continuous mediator, these effects will also be equal to the marginal effects on average for the population. When output equal to full is chosen as an option, the investigator must enter fixed values for the covariates C at which to compute conditional effects. The macro statement is as follows:

```
\%mediation(data= , yvar= , avar= , mvar= , cvar= , a0= ,
   a1= , m= , nc= , yreg= , mreg= , interaction= , output=,
   c=)
run;
```

When *output=full* is added, then, in addition to the controlled direct effect, along with the natural direct and indirect effect described above, two other effects are displayed. The natural direct and indirect effects we have been considering are sometimes called the "pure" natural direct effect and the "total" natural direct effect (Robins and Greenland, 1992). We can also consider the "total" natural indirect effect and the "pure" natural indirect effect. For a binary exposure the total natural direct effect expresses how much the outcome would change on average if the exposure changed from level $a^* = 0$ to level $a = 1$, but the mediator for each individual was fixed at the natural level that it would have taken at exposure level $a = 1$. The pure natural indirect effect expresses how much the outcome would change on average if the exposure were controlled at level $a^* = 0$ but the mediator were changed from the natural level it would take if $a^* = 0$ to the level that would have taken at exposure level $a = 1$. We discuss these effects further in Chapters 7 and 14 (cf. Robins and Greenland, 1992; Robins, 2003; VanderWeele, 2013b). These effects are additionally reported if the user selects *output = full*. If there is no exposure–mediator interaction, the "pure" and "total" natural direct effects will coincide and the "pure" and "total" natural indirect effects will coincide.

The investigator also has the option of implementing mediation analysis when data arise from a case–control design, provided that the outcome in the population is rare. To do so, the option *casecontrol=true* can be used. In this case the macro statement changes to

```
\%mediation(data= , yvar= , avar= , mvar= , cvar= , a0= ,
  a1= , m= , nc= , yreg= , mreg= , interaction= ,
  casecontrol=)
run;
```

Finally, the investigator can choose whether to obtain standard errors and confidence intervals via the delta method or a bootstrapping technique. The default is the delta method. To use bootstrapping, the option *boot=true* can be given. In this case the macro will compute 1000 bootstrap samples from which causal effects are obtained along with their standard errors (s.e.) and percentile confidence intervals ($p\_95\_CI_{\text{lower}}$, $p\_95\_CI_{\text{upper}}$). If the investigator wishes to use a higher number of bootstrap samples, instead of "true" he or she inputs the number of bootstrap samples desired (e.g., *boot=5000* would estimate standard errors and confidence intervals using 5000 bootstrap samples). The use of bootstrap for standard errors is

generally to be preferred if the sample size of the original sample is small, as it will lead to more accurate inferences than the delta method (MacKinnon, 2008). However, these issues are less important if the original sample is large, if this is the case, the use of delta method standard errors may be preferred because of computational efficiency. [For example, Ananth and VanderWeele (2011) conducted a mediation analysis using a sample of 26,000,000 individuals, and bootstrapping would have been computationally infeasible.] In general, for a common binary outcome in which "yreg=log-linear" is specified, bootstrapping standard errors often performs better than the delta method due to convergence issues. When using the bootstrap the macro statement changes to

```
\%mediation(data= , yvar= , avar= , mvar= , cvar= , a0= ,
  a1= , m= , nc= , yreg= , mreg= , interaction= , boot=)
run;
```

As noted above, if the investigator wants to add a categorical variable as covariate, this must be recoded as a series of indicator variables. For example, if a covariate, named catvar, takes four levels (1,2,3,4), we could construct three "dummy" or "indicator" variables, named, for example, ivar2, ivar3, and ivar4, leaving the first value as the reference. The variable ivar2 would take the value 1 for all observations which had catvar=2 and would take 0 for all other observations. The variable ivar3 would take the value 1 for all observations that had catvar=3 and would take 0 for all other observations, etc. The macro *dumvar* mentioned previously requires the user to list the categorical variables (e.g., catvar) that need to be transformed in the input *dvar*. The user needs also to input the prefix of the name of the dummy variables (e.g., ivar) that will be generated. Categorical variables can be both character and numerical using *dumvar*. For example, we can run the following:

```
dumvar data=data dvar=‘‘catvar’’ prefix=‘‘ivar’’
  drop=‘‘ivar1’’
```

Running this command will generate three indicator variables: "ivar2," "ivar3," "ivar4." For more examples, see http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1048.

2.8. DESCRIPTION OF THE SPSS MACRO

The SPSS macro that is provided was developed under the version 19.0, and it performs exactly the same tasks described in the previous section for the SAS macro. However, we point out some small differences that the investigator has to take into account when running mediation analysis using SPSS software.

Before invoking the mediation macro, the user has to open a new SPSS session and needs to specify the path in which he or she wants to save relevant estimates from the mediator and outcome regressions. This is simply done by running this command:

```
DEFINE !path()‘‘C:$\backslash $’’!ENDDEFINE
```

In between the quotation marks the path is defined, here for example the path "C:\" has been entered. For the newer versions of SPSS for Windows users, when specifying the path name, one should not use "C:\" since the newer version in Windows does not allow saving files to "C:\".

Macro activation requires that the macro script is then saved as a syntax file (the syntax file should be called from the session that has just been opened) and information is input in the following statement:

```
mediation data= / yvar= /avar= /mvar= /cvar= /NC= /a0=
 /a1= /m= /yreg= /mreg= /interaction= [/casecontrol= /boot
 =/nobs =/Output= /c=]
```

In the command above, the code outside the square brackets is mandatory while the code within square brackets is optional; these options are described further below. With regard to the part of the code that is required, first one inputs the name of the dataset (including the path, e.g. data="C:\mydata.sav"), then the name of the outcome variable (*yvar*=), the treatment variable (*avar*=), the mediator variable (*mvar*=), the other covariates (*cvar*=). Categorical variables need to be coded as a series of dummy variables before being entered as covariates. The macro *dummit* can be used for this purpose. Then the investigator needs to specify the baseline level of the exposure a^* (*a*0=), the new exposure level *a* (*a*1=), the level of mediator *m* at which the controlled direct effect is to be estimated, and the number of covariates to be used (*NC*=). When no covariates are entered, then the user still needs to write the command *cvar*= and needs to specify that *nc=0*. The user must also specify which types of regression have to be implemented. In particular, either LINEAR, LOGISTIC, LOG-LINEAR, POISSON, or NEGBIN can be specified in the option *yreg*. Logistic links for *yreg* can be used for rare dichotomous outcomes; otherwise, for dichotomous outcomes that are not rare, log links should be used for the outcome regression and the effects are given on the risk ratio scale. When using a log-linear model for a binary outcome, it is good to check the output to make sure the model has in fact converged. For the option *mreg*, either LINEAR or LOGISTIC regressions are allowed. If the dataset contains missing data, the software implements a complete case-only analysis. Finally, the analyst needs to specify whether an exposure–mediator interaction is present (TRUE or FALSE). As optional inputs, the investigator can use the option *casecontrol=TRUE*, when the data arise from a case–control study and the outcome is rare. A more complete output (described in the previous section) can be obtained using the option *Output=FULL* and entering the values for the covariates at which to compute causal effects conditional on those covariate values (*c*=). In order to enter the covariate values, the investigator needs to create a separate dataset that contains those values. For example, if two covariates *C* are present in the model and the value at which the investigator wants to fix the first is 4 and the value at which the investigator wants to fix the second is 10, at the beginning of the script the following commands need to be run:

```
Matrix.
compute c=make(1,2,0).
```

```
compute c(1,1)=4.
compute c(1,2)=10.
SAVE c(1,:) /OUTFILE=''C:$\backslash $c.sav''.
end matrix.
```

After having created a dataset for the covariate values, the user can specify the option *Output = FULL/ c="C:\c.sav"* to obtain the more complete output. If the *output=FULL* is requested and covariates are present, then the user is also required to specify *c="C:\c.sav"* to specify the level of the covariates at which conditional effects are to be estimated. If the investigator wishes to obtain bootstrap standard errors, he or she can use the option *boot=TRUE* followed by the number of observations in the dataset (*nobs=*) to compute causal effects and standard errors with 1000 bootstrap replications (or "boot=n," where n is the desired number of bootstrap samples). Otherwise, the default option is delta method standard errors.

As we mentioned in the previous section, if the investigator needs to add a categorical variable as covariate, a series of indicator variables needs to be generated. The SPSS macro *dummit* works very similarly to the SAS macro. In particular, the investigator needs to call the macro followed by three sets of parentheses. In the first set of parentheses the number of levels is entered, and in the second set of parentheses the name of the variable needs to be specified. Finally, in the third set of parentheses, the prefix for the new variables is entered. For example, if the variable we need to recode is "smoking," which takes levels "never," "past," and "current," then we can run the following macro:

```
dummit (3) (smoking) (smoke)
```

This macro would generate the following variables: "smokedum2" and "smokedum3." The category "never" is automatically taken as a reference. More examples can be found by using the following link:

http://www.glennlthompson.com/?p=92

Other related code in M-Plus, based on what was presented here in SPSS, can be found in Muthén (2012).

2.9. DESCRIPTION OF THE STATA MACRO

The Stata command "paramed" (Emsley et al., 2014) likewise implements the same range of analyses as the SAS and SPSS macros considered above; it was developed under Stata version 12.1. It uses slightly different syntax. The "paramed" package can be placed into the user's "ado/plus/p" folder or downloaded using "ssc install paramed."

The command can be used with the following statement:

```
paramed varname, avar() mvar() cvars() a0() a1()
 m() yreg() mreg()
```

First one inputs the name of the outcome variable ("varname" in the statement above), then the name of the exposure and mediator variables are put in the

parentheses within "avar()" and "mvar()" respectively, and the names of the other covariates are placed within the parentheses with "cvars()." When no covariates are being used, "cvars()" can be omitted. Categorical covariates need to be coded as a series of indicator variables before being entered as covariates. Unlike the SAS and SPSS macros, the number of covariates do not need to be given as a separate input. Then the investigator needs to specify the baseline level of the exposure a^* within parentheses in "a0()" and the new exposure level a in "a1()", along with the level of mediator m at which the controlled direct effect is to be estimated in "m()". The user must also specify which types of regression are to be implemented for the outcome regression in parentheses in "yreg()" and for the mediator regression in parentheses in "mreg()." For the outcome regression, this can be either linear, logistic, log-linear, Poisson, or negative binomial. Logistic links for yreg can be used for rare dichotomous outcomes; otherwise for dichotomous outcomes that are not rare, log links should be used for the outcome regression, and the effects are given on the risk ratio scale. For mediator regression, either linear or logistic regressions are allowed. If the dataset contains missing data, the software implements a complete case-only analysis.

The option "nointer" excludes the exposure–mediator interaction in the analysis and computes direct and indirect effects accordingly. The option "case" is to be specified if the data arise from a case–control study and the outcome is rare, and then the mediator regression is run just among the controls as described above. A more complete output (described in Section 2.7) can be obtained using the option "full" and by entering the values for the covariates at which to compute causal effects conditional on those covariate values; this is done by putting the covariate values in parentheses in "c()," which must also be given as an option if the "full" option is employed. If the investigator wishes to obtain bootstrap standard errors, he or she can use the option "boot" to compute causal effects and standard errors with 1000 bootstrap replications; to change the number of replications, the user can use the option "reps()" and input the number of replications in the parentheses; the option "seed()" can be used with the seed specified in parentheses if the user wishes to use a particular seed for the bootstrapping. Finally, the option "level()" can be used to specify the level at which the confidence interval is calculated.

2.10. HYPOTHETICAL EXAMPLE WITH OUTPUT

We present in this section a hypothetical example of using the mediation macro in SAS to illustrate the output. Output from the other software packages would be similar. We implement the analyses on a modified version of the fictitious dataset used by Preacher and Hayes (2004) to illustrate the output. In the next section we will describe a real example using the methods described in this chapter. The hypothetical example is used only to illustrate output. In the hypothetical example, residents of a retirement home diagnosed as clinically depressed are randomly assigned to receive 10 sessions of a new cognitive therapy ($A = 1$) or 10 sessions of an alternative therapeutic method ($A = 0$). After Session 8, the positivity of the evaluation the residents make for a recent failure experience is assessed (M). Finally, at the end of Session 10, the residents are given a questionnaire to measure

life satisfaction (Y). We might assess whether the cognitive therapy's effect on life satisfaction is mediated by the positivity of attributions of failure experiences.

The new dataset that we employ differs with respect to Preacher and Hayes' only in the way in which the outcome is simulated. In particular, the exposure and mediator variables are the same, but now the outcome is simulated as a normally distributed variable with mean equal to the linear regression estimated with the original data [the coefficients given in the outcome regression in Preacher and Hayes (2004)] plus a new term, the exposure–mediator interaction term, with coefficient equal to $\theta_3 = 0.5$, indicating a positive interaction, and standard deviation equal to the standard error of the residuals obtained from the outcome regression using Preacher and Hayes data (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html).

We first consider the case in which the interaction between the therapy and the attributions of negative experiences is omitted by the investigator. After having saved the dataset and inserted the SAS macro script, we run the following command:

```
\%mediation(data=dat, yvar=satis, avar=therapy, mvar=attrib,
  cvar= , a0=0, a1=1, m=0, nc= ,  yreg=linear, mreg=linear,
  interaction=false)
run;
```

The first output provided is the results of the outcome regression:

Dependent Variable: satis

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.71479	0.20449	-3.50	0.0017
therapy	1	0.66788	0.30147	2.22	0.0354
attrib	1	0.67186	0.16923	3.97	0.0005

Then the output of the mediator regression is given:

Dependent Variable: attrib

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.35357	0.21837	-1.62	0.1166
therapy	1	0.81857	0.29902	2.74	0.0106

Then the direct effects and indirect effects are given, including estimates for the controlled direct effect, the natural direct and indirect effect, and the total effect.

Obs	Effect	Estimate	s_e_	p Value	Lower	Upper
1	cde=nde	0.66788	0.30147	0.026733	0.07700	1.25877
2	nie	0.54997	0.24403	0.024215	0.07167	1.02827
3	total effect	1.21785	0.33475	0.000275	0.56174	1.87396

Note that when there is no exposure–mediator interaction, the controlled direct effect and natural direct effect coincide. We then run the mediation macro with the correctly specified outcome regression model that includes the exposure–mediator interaction term. We use the following command:

```
\%mediation(data=dat, yvar=satis, avar=therapy, mvar=attrib,
 cvar= , a0=0, a1=1, m=0, nc= , yreg=linear, mreg=linear,
 interaction=true)
run;
```

The output from the outcome regression is the following (the mediator regression will be the same):

Dependent Variable: satis

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.84424	0.19646	-4.30	0.0002
therapy	1	0.62132	0.27901	2.23	0.0348
attrib	1	0.30575	0.21913	1.40	0.1747
int	1	0.74464	0.31251	2.38	0.0248

We obtain the following estimates for the effects:

Obs	Effect	Estimate	s_e_	p Value	_95__CI_ Lower	_95__CI_ Upper
1	cde	0.62132	0.27901	0.02596	0.07446	1.16818
2	nde	0.35804	0.34759	0.30298	-0.32323	1.03931
3	nie	0.85981	0.28782	0.00281	0.29568	1.42395
4	marginal total effect	1.21785	0.33407	0.00027	0.56307	1.87263

In this example, the estimate of the indirect effect is biased downward if the interaction term is omitted, and is smaller in magnitude and has a larger p value. Moreover, when the interaction term is correctly added to the model, controlled direct effects and natural direct effects differ. The estimates above could be interpreted as the natural direct and indirect effects if the no-confounding assumptions (A2.1)–(A2.4) hold. Recall that even if the treatment is randomized, we still need to control for mediator–outcome confounding. In this hypothetical example, no covariates were controlled for and so the assumptions would probably be violated. The hypothetical example is included only for the purposes of illustrating the output of the macros. In the next section we will consider an example with real data and evaluate the confounding control assumptions more carefully.

2.11. EMPIRICAL EXAMPLE IN GENETIC EPIDEMIOLOGY

In this section we present an empirical example addressing a question about mediation in genetic epidemiology that was of recent interest. In 2008, three studies

(Amos et al., 2008; Hung et al., 2008; Thorgeirsson et al., 2008) found associations between genetic variants on chromosome 15q25.1 (rs8034191) and lung cancer. These variants were known also to be associated with smoking behavior, raising the question of whether the association of the variants with lung cancer is primarily through smoking or through other pathways. In addition to possible effects of genetics variants on 15q25.1 on lung cancer either through or independent of smoking, a third possible explanation of the associations was proposed (Thorgeirsson and Stefanson, 2010): that the variant may increase individuals' vulnerability to the harmful effect of tobacco smoke, a form of gene–environment interaction. The approach to mediation analysis above to estimate direct and indirect effects allowing for exposure–mediator interaction was employed and used with data from a lung cancer case-control study of 1836 cases and 1452 controls conducted at the Massachusetts General Hospital (VanderWeele et al., 2012a).

Cigarettes per day were used as a measure of smoking intensity. Linear regression was used for models of smoking intensity, measured as square root of cigarettes per day so as to better approximate a linear fit, though analyses using total cigarettes per day gave similar results. Logistic regression was used to model lung cancer status both with and without a smoking $\times$ variant interaction term. Covariates included in the models were genotype, age, sex, college education, and smoking duration; models for lung cancer also included smoking intensity (square root of cigarettes per day). Analyses that omitted smoking duration as a covariate gave qualitatively similar conclusions. The regression for smoking intensity and the regression for lung cancer risk were combined as in Section 2.4 to obtain direct and indirect effects using odds ratios for mediation analysis for a dichotomous outcome. The direct effect can be interpreted as the odds ratio comparing the risk of lung cancer with the genetic variant present versus absent if smoking behavior were what it would have been without the genetic variant. The indirect effect can be interpreted as the odds ratio for lung cancer for those with the genetic variant present comparing the risk if smoking behavior were what it would have been with versus without the genetic variant. The analyses assumed that conditional on the covariates there is no confounding of [assumption (A2.1)] the exposure–outcome relationship [assumption (A2.2)] the mediator–outcome relationship, and [assumption (A2.3)] the exposure–mediator relationship and that [assumption (A2.4)] there is no effect of the exposure that itself confounds the mediator–outcome relationship. The assumptions of no confounding of the effect of the exposure on the mediator and on the outcome [assumptions (A2.1) and (A2.3)] are likely to hold approximately when the exposure is a genetic variant with analysis restricted to a single ethnic group (in this analysis Caucasians); this is generally assumed in genetic studies. Assumption (A2.2) is that there is there is no unmeasured confounding of the mediator–outcome relationship. This may be less plausible in this example. A variable like neighborhood of residence might affect both smoking behavior (e.g., through cigarette adverting) and lung cancer (e.g., through air pollution). This assumption will be made in the analysis, but in the next chapter we will describe sensitivity analysis techniques that can be used to evaluate the sensitivity of conclusions to this assumption. Assumption (A2.4) is that none of the mediator–outcome confounders are themselves affected by the exposure. We would in general not expect

these specific genetic variants to affect a mediator–outcome confounder like neighborhood of residence. However, a variable like smoking duration might plausibly be associated with both cigarettes per day and lung cancer and is itself affected by the genetic variants. In this study, there was little evidence of association of the variants with smoking duration. Moreover, the analyses that did and did not control for smoking duration gave similar results.

Analyses from the study indicated strong evidence for a direct effect and suggested that the indirect effect is small. Ignoring possible gene–environment interaction gave a direct effect odds ratio of 1.35 (95% CI:1.21–1.52; $P = 3 \times 10^{-7}$) and an indirect effect odds ratio of 1.01 (95% CI:0.99–1.02; $P = 0.15$) per variant allele, with 3.6% of the increased risk mediated by smoking (see discussion in Section 2.13 regarding the proportion mediated). Allowing for smoking-by-gene interaction, for changes from 0–1 and from 0–2 variants alleles respectively, direct effect odds ratios are 1.31 (95% CI:1.15–1.49; $P = 3 \times 10^{-5}$) and 1.72 (95% CI:1.34–2.21; $P = 2 \times 10^{-5}$) and indirect effect odds ratios are 1.01 (95% CI:0.99–1.02; $P = 0.15$) and 1.03 (95% CI:0.99–1.07; $P = 0.16$), with 4.2% and 6.3% of increased risk mediated by smoking intensity respectively. Although accounting for interaction increases the indirect effect and the proportion mediated somewhat, most of the effect still appears to be direct. The confidence interval for the indirect effect is relatively narrow; the mediated effect is a relatively small portion. Most of the effect seems to be not by changing the number of cigarettes per day. Note, however, that the mediator here is cigarettes per day, and it is possible that other aspects of smoking (e.g., depth of inhalation) could potentially be responsible for the effect of the genetic variants on lung cancer.

One final point here is worth noting. In this example, and in any in which the exposure variable is not binary, there will be multiple exposure comparisons that could be done; for example, we could compare 0 and 1 variant alleles, or we could compare 0 and 2 variant alleles (or we could compare 1 to 2 variant alleles). In the presence of exposure–mediator interaction, the proportion mediated need not be the same across various comparisons. In the example just considered, when we allowed for interaction, we had a proportion mediated comparing 0 and 1 variant allele of 4.2% and a proportion mediated comparing 0 and 2 variant alleles of 6.3%. We could also compare, say, 1 and 2 variant alleles. In this case the proportion mediated is 5.4%. The proportion mediated need not be the same in the presence of exposure–mediator interaction (even comparing the same-sized exposure change—for example, 0 to 1 variant allele versus 1 to 2) because the importance of the interaction will vary depending on what reference exposure level is being considered.

2.12. WHEN TO INCLUDE AN EXPOSURE-MEDIATOR INTERACTION

In the previous sections and in the examples, we have considered potential exposure–mediator interaction. We have discussed how one of the advantages of the counterfactual-based approach to mediation is to be able to decompose a

total effect into direct and indirect effects even in the presence of such exposure–mediator interaction. We have considered in Sections 2.3–2.5 regression-based estimators for direct and indirect effects that allow for such exposure–mediator interaction in statistical models. A natural question that then arises is when to include such interaction in the models and when to omit them. An investigator might be tempted to only include such exposure-mediator interactions in the models if the interaction terms are statistically significant. However, for reasons that we will now describe, this approach is problematic. It is problematic because power to detect interaction tends to be very low unless the sample size is very large. In Chapter 13 of this book we will consider power and sample size formulae for interaction; and, there again, we will see that very large sample sizes are required to have reasonable power to detect interaction as statistically significant. The approach of only including exposure–mediator interaction in the statistical models if they are significant can thus be problematic because such exposure–mediator interaction may be important in capturing the dynamics of mediation even when the interaction terms are not "statistically significant" because the sample size is small. A better approach to deciding whether or not to include such exposure–mediator interaction in the outcome model is perhaps to include them by default and only exclude them if they do not seem to change the estimates of the direct and indirect effects very much. If the *magnitude* of the interaction term in the statistical outcome model is quite small and including versus excluding the exposure–mediator interaction does not change the direct and indirect effect estimates much, then it is probably safe to exclude the interaction. Otherwise, it is perhaps best to leave the exposure–mediator interaction in the outcome model. Again accounting for such interaction can be important even if the interaction coefficient is not statistically significant.

It is possible to have an exposure–mediator interaction that is not statistically significant but that increases the magnitude of the natural indirect effect by two- or threefold. In other cases, when exposure–mediator interaction is in fact present and it is ignored, there are settings in which, by ignoring it, one is led to precisely the wrong conclusion (e.g., that most of the effect is mediated when in fact most of it is direct). In the next chapter, we will see one such example in which ignoring exposure–mediator interaction leads to precisely the wrong conclusion. Even if the exposure-mediator interaction term is not statistically significant, it can be important in capturing the dynamics of mediation. It is also possible to have an exposure–mediator interaction term that is not statistically significant but that, by including it, substantially increases the power to detect the indirect effect. The exposure–mediator interaction tends to have two effects on the power to detect a mediated effect. On the one hand, including the exposure-mediator interaction requires an extra statistical parameter, reducing the degrees of freedom and thereby reducing power; however, often including the exposure–mediator interaction increases the magnitude of the natural indirect effect (by more fully capturing the dynamics of mediation), thereby increasing the power to detect the mediated effect; we saw this occur in the hypothetical example in Section 2.10.

The exposure–mediator interaction terms in the models in Sections 2.3–2.5 are not, at present, being given any sort of mechanistic interpretation, although we will consider such mechanistic or causal interpretations in Part II of the book. For the

time being, the exposure–mediator interaction is simply being included to allow for additional model flexibility, and this additional model flexibility can often be important for understanding the extent of mediation. The question may arise why we have focused so much on exposure–mediator interaction rather than exposure–covariate interaction or mediator–covariate interaction. The reason we have focused on exposure–mediator interaction is that the exposure and mediator are the two variables in the model for which we are giving the effects of these variables a causal interpretation. It is thus important to understand the relationships of these variables with the outcome as comprehensively as possible. The covariates, on the other hand, are included in the model for purposes of confounding control. While in some instances incorporating exposure–covariate or mediator–covariate interactions in the model might be important for ensuring that confounding control is adequate, control for only the main effects of the covariates can often do a reasonably good job at confounding control even in some settings in which the relationship between the covariates and the outcome is not strictly linear (cf. Maldonado and Greenland, 1993; Greenland and Maldonado, 1994). If exposure–covariate or mediator–covariate are thought to be important, then the methods described in Sections 2.17 and 2.18 below can be used to carry out mediation analysis in these settings.

In summary, exposure–mediator interactions can be important in capturing the dynamics of mediation, in fully accounting for the mediated effect, and sometimes even for increasing power to detect mediation. These reasons apply even when the interaction term in the model is not statistically significant. Exposure–mediator interactions certainly will not always be important, and in many cases they can be omitted without problems. If they are relatively small in magnitude and their inclusion or omission does not substantially change the estimates of the direct and indirect effects, then they can safely be omitted. Otherwise, however, it is probably best to include them in the models to allow for adequate model flexibility concerning the variables for which a causal interpretation is desired.

2.13. PROPORTION MEDIATED

To assess the extent to which the total effect of the exposure on the outcome operates through the mediator, a measure often called the "proportion mediated" is sometimes used. When the effects are used on the difference scale, the proportion mediated is then just defined as the ratio of the natural indirect effect to the total effect, that is,

$$PM = \frac{NIE}{TE}$$

This measure in some sense captures how important the pathway through the intermediate is in explaining the actual operation of the effect of the exposure on the outcome. It measures what would happen to the effect of the exposure on the outcome—by how much it would be reduced—if we were to somehow disable the pathway from the exposure to the intermediate. In the hypothetical example in Section 2.10 we had a natural direct effect of 0.358, a natural indirect effect of 0.859, and a total effect of 1.217. The proportion mediated would then just

be the ratio of the natural indirect effect to the total effect (i.e., $0.859/1.217 = 70.5\%$) and we might say that 70.5% of the effect of therapy on life satisfaction was mediated by positive attribution. Confidence intervals can be obtained by bootstrapping or some analytic formulae are given in the online supplement of Valeri and VanderWeele (2013).

This proportion-mediated measure can be a helpful summary. However, it is subject to a number of limitations. First, it has been shown to be a highly variable measure (MacKinnon, 2008). The confidence intervals for it are often quite wide. If the confidence interval is being used simply to assess whether some of the effect is mediated, then it is much better simply to use the confidence interval for the natural indirect effect itself. Second, the measure is also problematic when the natural direct effect and natural indirect effect operate in different directions. One can then obtain a proportion mediated much larger than 100%, and the measure is no longer really meaningful. Even more dramatically, if the natural direct and indirect effects are of roughly the same magnitude but of opposite signs, then the total effect will be close to zero and the proportion-mediated measure then takes the natural indirect effect and divides it by a number close to zero which will result in an enormous (and again meaningless) proportion. This problem also creates even further instability in the estimate of the proportion mediated, again often resulting in large standard errors. The measure can be a useful summary but should only be used when direct and indirect effects are in the same direction.

Our discussion of this proportion mediated measure has thus far been focused on natural direct and indirect effects calculated on the difference scale. However, as we have discussed, when an outcome is binary, often a ratio scale is used. When we use an odds ratio or risk ratio scale to estimate natural direct and indirect effects, we can still obtain a proportion-mediated measure on the risk difference scale, but doing so requires a particular transformation. Specifically, suppose we use odds ratios and have a rare outcome (or alternatively replace the odds ratios with risk ratios). If we have a natural direct effects odds ratio of OR^{NDE} and a natural indirect effect odds ratio of OR^{NIE}, then provided that the outcome is rare, we can calculate the proportion mediated on a risk difference scale by (VanderWeele and Vansteelandt, 2010)

$$\frac{OR^{NDE}(OR^{NIE} - 1)}{(OR^{NDE} \times OR^{NIE} - 1)}$$

To gain some intuition why this holds, consider the following example. Suppose we have an outcome that is rare so that odds ratios approximate risk ratios. If the unexposed risk for a particular individual is 0.1% with a natural direct effect odds ratio of $OR^{NDE} = 9$, then the risk of the outcome if the exposure were present but the mediator were fixed to what it would have been if the exposure had been absent would be $0.1\% \times 9 = 0.9\%$. A change from the exposure being absent with the mediator set to what it would be if the exposure were absent to a setting in which the exposure is present but the mediator is set to what it would be if the exposure were absent thus changes the risk from 0.1% to 0.9% for an overall change of $0.9\% - 0.1\% = 0.8\%$ on the difference scale; this is essentially the natural direct effect now on the difference scale rather than the odds ratio scale. Now if the natural indirect effects odds is $OR^{NIE} = 2$, then the risk of the outcome if the exposure were present would be

$0.1\% \times 9 \times 2 = 1.8\%$. A change from the exposure being present with the mediator set to what it would be if the exposure were absent (which was a risk of 0.9%) to a setting in which the exposure is present but the mediator is set to what it would be if the exposure were present (where the risk is 1.8%) changes the risk from 0.9% to 1.8% for an overall change of $1.8\% - 0.9\% = 0.9\%$ on the difference scale; this is essentially the natural indirect effect now on the difference scale rather than on the odds ratio scale. The total effect on the difference scale comparing the exposure absent (with risk 0.1%) to the exposure present (with risk 1.8%) is a total effect of $1.8\% - 0.1\% = 1.7\%$. Thus on the difference scale we would have a proportion mediated of $(0.9\%)/(1.7\%) = 52.9\%$; that is, on the risk difference scale, 52.9% of the increased risk would be due to the mediator. The formula above would likewise give $OR^{NDE}(OR^{NIE}-1)/(OR^{NDE} \times OR^{NIE}-1) = 9(2-1)/(9 \times 2-1) = 52.9\%$. Essentially the formula goes through the reasoning above automatically to capture this proportion mediated on the risk difference scale.

Note in this hypothetical example, the proportion mediated on the risk difference scale is above 50%, even though the natural direct effect odds ratio, $OR^{NDE} = 9$, is larger than the natural indirect effect odds ratio, $OR^{NIE} = 2$. This is essentially because the natural direct effect and natural indirect effect are using different reference levels of risk. The natural direct effect is using the unexposed risk, 0.1%, as a reference level of risk in calculating the natural direct effect ratio when the exposure is present but the mediator is set to what it would be if the exposure were absent, $OR^{NDE} = 0.9\%/0.1\% = 9$. The natural indirect effect is using the risk when the exposure is present but the mediator is set to what it would be if the exposure were absent, 0.9%, as the reference level of risk in calculating the natural indirect effect when setting the exposure to present and the mediator to what it would be if the exposure were present, $OR^{NIE} = 1.8\%/0.9\% = 2$. Thus, although the natural indirect effect is larger on the risk difference scale, it looks smaller on the ratio scale because it is working with a different reference level of risk when calculating the ratio. For this reason, if the proportion mediated is to be used when working with binary outcomes and ratio measures, it is important to use the formula above for the proportion mediated for risk differences when calculating this measure. As before, the measure should only be used if the natural direct and indirect effects operate in the same direction.

If we return to the example from genetic epidemiology above, when we ignored the potential gene-smoking interaction, we obtained a direct effect odds ratio of 1.35 (95% CI:1.21–1.52; $P = 3 \times 10^{-7}$) and an indirect effect odds ratio of 1.01 (95% CI:0.99–1.02; $P = 0.15$) per variant allele. We could then calculate a proportion mediated measure of $OR^{NDE}(OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1) = 1.35(1.01 - 1)/(1.35 \times 1.01 - 1) = 3.7\%$. When we allowed for smoking-by-gene interaction, for changes from 0–1 and from 0–2 variants alleles respectively, we obtained direct effect odds ratios of 1.31 (95% CI:1.15–1.49; $P = 3 \times 10^{-5}$) and 1.72 (95% CI:1.34–2.21; $P = 2 \times 10^{-5}$) and indirect effect odds ratios of 1.01 (95% CI:0.99–1.02; $P = 0.15$) and 1.03 (95% CI:0.99–1.07; $P = 0.16$). For a change from 0–1 alleles we could calculate the proportion mediated as $OR^{NDE}(OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1) = 1.31(1.01 - 1)/(1.31 \times 1.01 -$

$1) = 4.1\%$. For a change from 0–2 alleles we could calculate the proportion mediated as $OR^{NDE}(OR^{NIE} - 1)/(OR^{NDE} \times OR^{NIE} - 1) = 1.72(1.03 - 1)/(1.72 \times 1.03 - 1) = 6.7\%$. In all cases, however, most of the effect seems to operate through pathways other than by increasing the number of cigarettes per day.

2.14. PROPORTION ELIMINATED

Although natural direct and indirect effects are useful in assessing the importance of particular pathways, they do not correspond to any particular intervention that we could actually carry out in practice. This is because they require, for example, for each individual, fixing the mediator to the level it would have been in the absence of exposure, whereas we do not in general know what those values are for those persons actually exposed. Natural direct and indirect effects, although helpful for assessing the impact of different pathways, are thus of more limited interest from a policy perspective.

An alternative proportion measure may be of more interest in policy settings. The controlled direct effect fixing the intermediate to level $M = m$, $CDE(m)$, captures the effect of the exposure on the outcome if the intermediate were set, possibly contrary to fact, to level m. This is an intervention we might hope to be able to carry out in practice. We might hope that by intervening on the intermediate, we could block a substantial part of the effect of the exposure on the outcome (Robins and Greenland, 1992; VanderWeele, 2013a). A proportion measure that could then be used and that would be of policy relevance would be the proportion of the effect of the exposure on the outcome that could be eliminated by intervening to set the intermediate to some fixed level m. We might call this measure the proportion eliminated and denote it by $PE(m)$. On a difference scale, this would be

$$PE(m) = \frac{TE - CDE(m)}{TE}$$

that is, the difference between the total effect and the controlled direct effect fixing the mediator level to m (which measures the extent of the effect that is eliminated by fixing the mediator to level m) divided by the total effect itself, to obtain a proportion. If this "proportion eliminated" (PE) were large and we wanted to prevent the effect of the exposure on the outcome, we might try to implement policies to intervene on the intermediate. Whereas the proportion mediated essentially captures what would happen to the effect of the exposure if we were to somehow disable the pathway from the exposure to the intermediate (setting it to its natural value in the absence of the exposure), the proportion eliminated measure captures what would happen to the effect of the exposure on the outcome if we were to fix the intermediate to the same fixed value $M = m$ for all persons.

Importantly, the "proportion eliminated" measure will not always equal the proportion mediated. Suppose that the exposure and the intermediate interacted but that the exposure had no effect on changing the intermediate itself. In this case the indirect effect of the exposure on the outcome through the intermediate would be

0 (because the exposure does not change the intermediate) and we would have a proportion mediated of $PM = NIE/TE = 0/TE = 0\%$. However, if, with interaction, the effect of the exposure were large with the intermediate but small without the intermediate, then the "proportion eliminated" by fixing the intermediate to 0 might be substantial. If we were to fix the intermediate to 0 for everyone, the exposure may not have much of an effect on the outcome; that is, $CDE(m = 0)$ might be quite small, and the proportion eliminated $PE(m = 0) = [TE - CDE(m = 0)]/TE$ might be close to 100%. In the extreme case in which there is a "pure interaction" (so that the exposure has no effect on the outcome unless the intermediate is present) but if it is also the case that the exposure has no effect on the intermediate itself, then the proportion mediated is 0% but the proportion eliminated by fixing $m = 0$ is 100%.

More generally the proportion mediated measure $PM = NIE/TE$ and the proportion eliminated measure $PE(m) = [TE - CDE(m)]/TE$ may differ because, in the presence of an interaction between the exposure and the intermediate, we may have a different proportion eliminated measure for every value of m. Because the total effect decomposes into a natural direct effect and natural indirect effect ($TE = NIE + NDE$), we can re-express the proportion mediated as $PM = [TE - NDE]/TE$. In the technical language of the causal inference literature, the proportion eliminated and the proportion mediated may differ because the controlled direct effect may not equal the natural direct effect; the natural direct effect essentially averages over the various controlled direct effects. The two measures will coincide when there is no interaction between the exposure and the intermediate (either at the individual level (Robins, 2003) or at the expected population level under no confounding assumptions (A2.1)–(A2.4) (VanderWeele and Vansteelandt, 2009).

If we have estimated a controlled direct effect using odds ratios with a rare outcome (or risk ratios with a common outcome), we can convert these to obtain a proportion eliminated measure on the excess relative risk scale. In particular, if $OR^{CDE}(m)$ denotes the controlled direct effect on the odds ratio scale, when fixing the mediator to level $M = m$, and OR^{TE} denotes the total effect on the odds ratio scale, then assuming a rare outcome, the proportion eliminated on an excess relative risk scale can be calculated by

$$PE(m) = \frac{OR^{TE} - OR^{CDE}(m)}{OR^{TE} - 1}$$

Importantly, this is somewhat different from the proportion eliminated on the risk difference scale since the denominators of OR^{TE} and $OR^{CDE}(m)$ differ (cf. Suzuki et al., 2014). We will return to the use of proportion eliminated measures when using odds ratio or risk ratio estimates, along with the causal interpretation of such proportion eliminated measures, in Chapter 14.

An example where the proportion mediated and proportion eliminated measures do diverge is in the analysis of genetic epidemiology above examining the effects of chromosome 15q25 genetic variants on lung cancer with cigarettes smoked per day as an intermediate: Each variant allele on 15q25 increases the risk of lung cancer

by about 1.3-fold; however, a small proportion (perhaps no more than $5-10\%$) of this effect is mediated by increasing cigarettes per day. However, there is strong interaction between these variants and cigarettes per day in their effects on lung cancer, and there may in fact be a "pure" interaction such that the variants have no effect on lung cancer for those who do not smoke (Li et al., 2010a). The variants may operate by increasing the nicotine and toxins extracted per cigarettes smoked. In this case, the controlled direct effect if we were able to fix cigarettes per day to 0 for everyone would be $CDE(m=0)=0$ and thus the proportion of the effect eliminated would be $PE(m=0)=[TE-CDE(m=0)]/TE=[TE-0]/TE=100\%$. By eliminating smoking, we eliminate all of the effect of the variants. The proportion mediated (by cigarettes per day) is small because the variants do not increase cigarettes per day all that substantially (only by about 1 cigarette per day); but the proportion eliminated by fixing cigarettes per day to 0 is large because the variants do not seem to affect lung cancer without smoking.

The two measures, proportion eliminated and proportion mediated, have differing interpretations. The proportion eliminated is in general the more relevant policy measure. It captures how much of the effect of the exposure on the outcome we could eliminate by intervening on the intermediate. The proportion mediated captures how much of the effect of the exposure on the outcome is because of the effect of the exposure on the intermediate. It gives insight into the role of different pathways but not necessarily on what would happen if we were to intervene on particular intermediates. The proportion eliminated measure is attractive because it concerns the effect of actual potential policy interventions and because it requires only the estimation of controlled direct effects, which, as we have seen, can be identified under somewhat weaker assumptions than natural direct and indirect effects. If effect decomposition and evaluation of the operation of various pathways are of interest, natural direct and indirect effects and the proportion mediated measure may still be of interest. But for policy purposes, the proportion eliminated is often the more relevant measure.

2.15. STUDY DESIGN AND MEDIATION ANALYSIS

We discussed above our four assumptions about confounding that were important in giving causal interpretations to the direct and indirect effect estimates. Our assumptions were that control had been made for [assumption (A2.1)] exposure–outcome confounding, [assumption (A2.2)] mediator–outcome confounding, and [assumption (A2.3)] exposure–mediator confounding and that [assumption (A2.4)] none of the mediator–outcome confounders were themselves affected by the exposure. We also discussed how these assumptions essentially then required also an assumption about temporality: that the exposure preceded the mediator and that the mediator preceded the outcome. If such temporality did not hold, then associations between the exposure, mediator, and outcome would not reflect causal effects, implying violations of the confounding assumptions.

Several important implications follow from this concerning study design for mediation analysis. First, these issues of temporality and confounding have

important implications for the timing and measurement of variables. Studies should be designed and data collected in such a way that it is ensured, to the greatest extent possible, that the exposure precedes the mediator and that the mediator precedes the outcome. This will often require that data are collected on at least two, and often even three, different time points. It will also often rule out the use of cross-sectional designs. The trouble with cross-sectional designs is that it will often be difficult to know the direction of causality. If the exposure and mediator are associated in a cross-sectional design, it will be difficult to know if this is because the exposure affects the mediator, or if it is rather the case that the mediator affects the exposure, or if both are the case. In other words, with cross-sectional data, it will be difficult to rule out feedback and reverse causation. For example, if cognitive behavioral therapy were taken as the exposure, antidepressant use as the mediator and depressive symptom scores as the outcome, but cross-sectional data are used, then it will in general be difficult to know if an association between therapy and antidepressant use are because therapy encourages patients to comply with antidepressants, or whether it is rather the case that those who take anti-depressant medications are more functional and therefore more likely to attend therapy sessions. We cannot distinguish between these two explanations for associations with cross-sectional data or assess their relative contributions. Likewise, if we found that the supposed mediator, antidepressant use, were associated with the outcome, depressive symptom scores, then, with cross-sectional data, we would not know to what extent such associations reflected an actual effect of the antidepressant on depressive symptoms and to what extent the associations resulted from, or were muted by, those with more severe depression being more motivated to use an antidepressant. Again, with cross-sectional data, we cannot in general determine the direction of causality or, in the case of feedback, the relative magnitudes of the two possible directions that causality may operate.

To assess causality, and thus mediation, we will in general need designs that capture data at different points in time. If, because of the design of the study and the timing of the measurements, we know that the exposure precedes the mediator and that the mediator precedes the outcome, then these issues of reverse causation will be less of a concern. The most straightforward way to design a study to ensure temporality is to measure the exposure, the mediator, and the outcome used in the analysis at three different times. However, even with data collected at a single point in time, it is at least sometimes possible to have the temporal ordering clear. For example, in the lung cancer case–control study considered above, the exposure was a genetic variant, the mediator cigarettes per day, and the outcome lung cancer. Even if the data were collected at a single point in time, it is clear that the genetic variant precedes the mediator and the outcome (since the genetic variant is fixed at conception); and likewise, if the cigarettes per day measure is a self-reported average measure of cigarettes per day during the time preceding the assessment of lung cancer status, then it will also be clear that the mediator precedes the outcome. Thus, in some cases, it will be possible to establish the temporal ordering of variables even if all of the actual data collection takes place at a single point in time. However, in general, with cross-sectional data in which temporal ordering is not

clear and in which causality may occur in both directions, we cannot reliably draw conclusions about mediation.

Conceived of in another way, we might say that, with cross-sectional data, we cannot in general distinguish between mediation and confounding. If the variable A precedes M and M affects Y, then M may be a mediator of the effect of A on Y; but if M precedes A and M affects Y, then M may be a confounder of the effect of A on Y. No statistical techniques will distinguish between these two possibilities if we do not know the temporal ordering or something further about causal relationships between these variables. Establishing whether a variable potentially plays the role of a mediator or a confounder can only be done with substantive knowledge about the nature of the variables, the design of the study, and the timing of measurements.

Of course, in some cases, a variable may be both a mediator and a confounder. Therapy may affect subsequent antidepressant use (and thus antidepressant use may mediate some of the effect of therapy on the depression outcome), but it may also be the case that antidepressant use may affect whether a patient shows up for subsequent therapy (and thus prior antidepressant use might be a confounder for the effect of subsequent therapy). Once again, with cross-sectional data, it is not in general possible to tease apart these issues and distinguish between mediation and confounding when there is such feedback. We will revisit this issue of feedback between variables in the context of mediation in Chapter 6 in the setting of time-varying exposures, mediators, and confounders, but the possibility of feedback is relevant even when we have a single exposure, a single mediator, and a single outcome. Essentially this is because, even if we have measured an exposure at time t, a mediator at time $t + 1$, and an outcome at time $t + 2$, and thus know the temporal ordering of these variables, it is possible that the prior values of the exposure, the mediator, and the outcome could in fact serve as the most important confounding variables. The prior value of the mediator at time $t - 1$, say, may affect both the exposure at time t and the outcome at time $t + 2$ and thus serve as a confounder. Because of this potential for feedback and reverse causation, the strongest study designs for providing evidence for mediation will in general be those in which the exposure precedes the mediator, which precedes the outcome, and in which previous values of the exposure, mediator, and outcome can also be included in the set of baseline confounders (i.e., the values of the exposure, mediator and outcome, measured prior to the exposure measurement used as the actual exposure in the analysis, are included in the set of confounders C). Control for prior values of the exposure, mediator, and outcome helps ensure that the confounding assumptions, required for a causal interpretation of the direct and indirect effects, are more likely to be plausible.

In some settings, prior values of a particular exposure (or mediator or outcome) do not exist, and such considerations are then rendered irrelevant. For example, in the genetics example considered in Section 2.11, the exposure was a genetic variant and this is fixed at conception and has no prior value (though, even here, controlling for the genotype of the parents can help control for confounding). Likewise, in some examples later in the book, we will take mortality as the outcome and, assuming that everyone is alive at study entry, there is no "prior outcome" to

control for. Similarly again, if the exposure is an intervention that no one has previously received, there will be no "prior exposure" to control for; control effectively is made for it automatically since no one previously had the exposure. However, in general, when the exposure, mediator, or outcome vary over time, control for prior values of these variables as confounders will render the confounding assumptions more plausible and help rule out the possibility of reverse causation. These issues are partially addressed if the exposure itself is a randomized. If the exposure is randomized, then there will be no exposure–outcome confounding and no exposure–mediator confounding. However, as we have already noted, even if the exposure is randomized, there may still be mediator–outcome confounding (since the mediator has not been randomized) and thus, in such circumstances, controlling for past values of the mediator and the outcome (values prior to the exposure's randomization) can help control for mediator–outcome confounding; even previous values of the exposure (if defined prior to randomization) may help control for such mediator–outcome confounding if such pre-randomization values of the exposure are thought to likewise affect the post-randomization values of both the mediator and the outcome. In Chapter 6 we will consider further designs and methods that make use of multiple measurements of the exposure, the mediator, and the outcome. However, these principles of study design are again relevant even to the methods considered in this chapter. If we have a single exposure, mediator, and outcome used in the analysis and employ the methods described above in this chapter, then the confounding assumptions (A2.1)–(A2.4) will often be more plausible if we can control for prior values of the exposure, mediator, and outcome (i.e., values measured before the measurement of the exposure variable used in the analysis). In the next chapter we will also consider some alternative designs in which both the exposure and the mediator can be randomized to help ensure that the no-confounding assumptions are satisfied (cf. Imai et al., 2013; Emsley and VanderWeele, 2014).

Our focus in this section thus far has been principally on assumptions (A2.1)–(A2.3)—that is, on control for exposure–outcome, mediator–outcome, and exposure–mediator- confounding. However, our fourth assumption [assumption (A2.4)], that none of the mediator–outcome confounders are themselves affected by the exposure, also has important implications for study design. We have touched on this already, but this fourth assumption requires that, of everything on the pathway from the exposure measure used in the analysis to the mediator measure used in the analysis, there is no intermediate on this pathway from exposure to mediator that itself affects the outcome (Robins, 2003). As we have already discussed, this may be more plausible if there is a short period of time between the exposure and the mediator, rather than if there is a longer period of time (VanderWeele and Vansteelandt, 2009). When designing a study, the confounding assumptions may thus be more plausible if the time between the exposure and mediator measurements used in the analysis can be kept relatively short. Of course, some questions concerning mediation will not conform to this idealized setting and, in some circumstances, our mediator of interest may occur considerably after our exposure of interest. Chapter 5 will deal at greater length with such settings and designs in

which our fourth assumption (A2.4) is violated; we will also revisit the issue further in Chapter 6 in the context of exposures and mediators that may vary over time.

The importance of study design is often emphasized in trying to draw causal inferences from observational data (Shaddish et al., 2002; Rothman et al., 2008; Rubin, 2011; Hernán and Robins, 2015). Within the context of assessing mediation and direct and indirect effects, issues of temporality and study design are even more complex and subtle as the effects of multiple variables are in view. Such issues need to be thought about carefully and taken seriously. We will return to these important issues of study design again in Chapter 3 and yet further in Chapters 5 and 6.

2.16. COUNTERFACTUAL NOTATION FOR NATURAL DIRECT AND INDIRECT EFFECTS

Thus far we have described in words the direct and indirect effects that we have been estimating. However, a very compact and concise mathematical notation using counterfactual or potential outcomes (Neyman, 1923; Rubin, 1974) is also available for describing these effects (Robins and Greenland, 1992; Pearl, 2001). We will briefly describe this notation here, although it will not be necessary for a reader to grasp this notation to follow most of the remaining part of the chapter or subsequent chapters. Nonetheless, familiarity with this notation is important for reading through the various papers that have developed the methods described in this book. In addition, some sections or subsections in the chapters that follow will necessitate the use of this notation, though the reader will be alerted to this when it is so.

As before, let A denote the exposure of interest; Y, the outcome of interest; and M, the potential mediator of interest. For example, in our genetic epidemiology example, A might denote the presence of the genetic variant (for now, for simplicity, let us suppose that this is binary: present or absent; we will generalize this later), Y might denote lung cancer, and M might denote cigarettes per day. We now introduce counterfactual or potential outcomes (Neyman, 1923; Rubin, 1974, 1978) which will be used to formally define the direct and indirect effects of interest. Let Y_a denote a subject's outcome if treatment A were set, possibly contrary to fact, to a. The variables Y_a are referred to as counterfactuals or potential outcomes. If A is binary, then an individual will have two potential outcomes: Y_0, what the outcome would have been for the subject if exposure had been set, possibly contrary to fact, to 0, and, Y_1, what the outcome would have been for the subject if exposure had been set, possibly contrary to fact to level 1. In the context of the genetic example (assuming for now that the genetic variant is binary), we would consider for each individual whether they would have had lung cancer had the variant been present, Y_1, and whether they would have had lung cancer had the genetic variant been absent, Y_0. For each individual, we only get to observe one of these two potential outcomes, Y_1 if the individual actually had the exposure ($A = 1$) and Y_0 if the individual did not have the exposure ($A = 0$). However, at least in theory we could conceive of both hypothetical outcomes. If the two were different, then we would say that the exposure would have an effect on that individual. The causal

effect for the exposure on the outcome for that individual would be $Y_1 - Y_0$. Of course, since we do not get to observe both potential outcomes for any person, we cannot in general estimate the causal effect for a particular individual. At best we can hope to be able to do so on average for a population. We thus define the average causal effect for a population to be $\mathbb{E}[Y_1 - Y_0]$. We could also consider the average causal effect for subpopulation with particular covariates $C = c$. This would be the conditional causal effect, $\mathbb{E}[Y_1 - Y_0|c]$. However, even to estimate these average or conditional causal effects, we would need to adjust for variables that suffice to control for confounding of the exposure–outcome relationship.

In the context of mediation there will also be potential outcomes for the mediator variable as well. We thus let M_a denote a subject's counterfactual value of the mediator M if exposure A were set to the value a. For a binary exposure we would again have two potential outcomes for the mediator: M_1 and M_0, what the mediator would have been with or without the exposure respectively. Once again we only get to observe one of these two potential outcomes for each individual. In the context of the genetics example, M_1 would be how many cigarettes were smoked per day on average if the individual had had the genetic variant and M_0 would be the number of cigarettes smoked per day on average if the individual had not had the genetic variant.

Finally, we will also consider potential outcomes for Y under hypothetical interventions on both the exposure and the mediator. We let Y_{am} denote a subject's counterfactual value for Y if A were set to a and M were set to m. Then, for each individual, we have a potential outcome for each setting of the exposure and the mediator. In the context of the genetic example, $Y_{a=1,m=10}$ would be the outcome we would have observed had the individual had the genetic variant and smoked 10 cigarettes per day on average. Likewise, $Y_{a=0,m=20}$ would be the outcome we would have observed had the individual had not had the genetic variant and smoked 20 cigarettes per day on average. We have many different potential outcomes of this form Y_{am}, but once again we only get to observe one of them.

Using counterfactuals of this form, Robins and Greenland (1992) and Pearl (2001) gave then the following definitions for what are now often called "controlled direct effects" and "natural direct and indirect effects." The controlled direct effect of treatment A on outcome Y comparing $A = 1$ with $A = 0$ and setting M to m is defined by $Y_{1m} - Y_{0m}$ and measures the effect of A on Y not mediated through M—that is, the effect of A on Y after intervening to fix the mediator to some value m. The controlled direct effect then represents the effect of treatment on the outcome intervening to fix the intermediate variable to some particular level. In the genetics example, on a difference scale, the controlled direct effect, $Y_{1m} - Y_{0m}$, would denote the effect on lung cancer, comparing the presence and the absence of the genetic variant with cigarettes per day fixed at level m. Note that this direct effect may vary with m. There may, for example, be individuals for whom there would be no direct effect when cigarettes per day is fixed to $M = 0$ [i.e., we might have $CDE(0) = Y_{1,0} - Y_{0,0} = 0$], while there would be a direct effect if cigarettes per day were fixed to $M = 20$ [i.e., we might have $CDE(20) = Y_{1,20} - Y_{0,20} = 1$]. Once again, we cannot in general hope to obtain these effects for an individual; but, under assumptions given in Section 2.3, namely assumptions (A2.1) and (A2.2),

no-unmeasured confounding for the exposure–outcome and mediator–outcome relationships, we could identify these on average. The average controlled direct effect for a population is then denoted by $\mathbb{E}[Y_{1m} - Y_{0m}]$ and the average controlled direct effect, conditional on covariates $C = c$, is denoted by $\mathbb{E}[Y_{1m} - Y_{0m}|c]$. This latter expression was what we had called the average controlled direct effect conditional on the covariates, $CDE(m)$, in Sections 2.2 and 2.3 and which our regression-based approach estimated under assumptions (A2.1) and (A2.2).

In contrast to controlled direct effects, which fix the mediator to one specific value, natural direct effects fix the intermediate variable for each individual to the level it would have naturally been in the absence of exposure. The natural direct effect of exposure A on outcome Y comparing $A = 1$ with $A = 0$ intervening to set M to what it would have been if exposure had been $A = 0$ is formally defined by $Y_{1M_0} - Y_{0M_0}$. Essentially the natural direct effect assumes that the intermediate M is set to M_0, the level it would have been for each individual had exposure been 0, and then compares the direct effect of treatment (with the intermediate set to this level M_0). In the context of the genetics example, the natural direct effect thus captures the effect of the exposure, comparing the variant being present to that being absent, on the lung cancer outcome, intervening to set the mediator, cigarettes per day, to the level it would have been in the absence of the exposure level—that is, if the genetic variant had been absent. We can likewise define the average natural direct effect for the population as $\mathbb{E}[Y_{1M_0} - Y_{0M_0}]$ or conditional on covariates $C = c$, $\mathbb{E}[Y_{1M_0} - Y_{0M_0}|c]$. Again this latter effect was what we had called the average natural direct effect conditional on the covariates, NDE, in Sections 2.2 and 2.3 and which our regression-based approach estimated under assumptions (A2.1)–(A2.4): that conditional on covariates C there is no unmeasured [assumption (A2.1)] exposure–outcome confounding, [assumption (A2.2)] mediator–outcome confounding, and [assumption (A2.3)] exposure–mediator confounding and that [assumption (A2.4)] there are mediator-outcome confounders affected by the exposure.

Corresponding to a natural direct effect is a natural indirect effect. The natural indirect effect comparing fixing the mediator to M_1 versus M_0 and intervening to set exposure to $A = 1$ is formally defined by $Y_{1M_1} - Y_{1M_0}$. The natural indirect effect assumes that the exposure is set to level $A = 1$ and then compares what would have happened if the mediator were set to what it would have been if exposure had been $A = 1$ versus what would have happened if the mediator were set to what it would have been if exposure had been $A = 0$. In the context of the genetics example, the natural indirect effect thus captures the effect on lung cancer comparing what would happen if the genetic variant were present but we fixed cigarettes per day to the level it would have been had the variant been present versus absent. Note that for the natural indirect effect to be nonzero, M_1 and M_0 must be different from each other; otherwise the two counterfactuals in the contrast, $Y_{1M_1} - Y_{1M_0}$, would be the same and their difference would be zero. In other words the exposure has to have some effect on the mediator. Moreover, this change in the mediator from M_0 to M_1 must itself go on to change the outcome. So for the natural indirect effect to be nonzero, the exposure must change the mediator and then that change in the mediator must go on to change the outcome—this is essentially what

we mean by mediation. This is why the natural indirect effect captures, formally using counterfactuals, our notion of mediation. We can likewise define the average natural indirect effect for the population as $\mathbb{E}[Y_{1M_1} - Y_{1M_0}]$ or conditional on covariates $C = c$, $\mathbb{E}[Y_{1M_1} - Y_{1M_0}|c]$. Again this latter effect was what we had called the average natural direct effect conditional on the covariates, *NIE*, in Sections 2.2 and 2.3 and which our regression-based approach estimated under assumptions (A2.1)–(A2.4).

With these counterfactual definitions we can also see why a total effect can be decomposed into a natural direct and indirect effect. For example, with a binary treatment the total effect $Y_1 - Y_0$ can be written as $Y_1 - Y_0 = Y_{1M_1} - Y_{0M_0} = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$, where the final equality is obtained simply by adding and subtract the Y_{1M_0} term. The first expression in the decomposition is then the natural indirect or mediated effect and the second expression is the natural direct effect. Note that because we have defined effects in terms of counterfactuals, neither our definitions nor our decomposition presume any specific model or functional form, nor do they make assumptions about interaction. It is by using these counterfactual-based notions that we are able to achieve the generality we have considered in this chapter. We can also easily extend these counterfactual definitions to settings without a binary exposure as considered in previous sections of this chapter. For a general exposure A, if we are comparing two levels of exposure a and a^*, then we can define the effects by simply replacing "1" by "a" and "0" by "a*" in the definitions above. That is, we can define the controlled direct effects by $Y_{am} - Y_{a^*m}$, natural direct by $Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$, and natural indirect effects by $Y_{aM_a} - Y_{aM_{a^*}}$. We can decompose the total effect $Y_a - Y_{a^*}$ as $Y_a - Y_{a^*} = Y_{aM_a} - Y_{a^*M_{a^*}} = (Y_{aM_a} - Y_{aM_{a^*}}) + (Y_{aM_{a^*}} - Y_{a^*M_{a^*}})$, where the first expression in the sum is again the natural indirect effect and the second expression is the natural direct effect. We could similarly define population average, or conditional average versions of these effects. And we can also define these effects on the risk ratio or odds ratio scale as well (see Appendix for the formal counterfactual definitions on an odds ratio or risk ratio scale).

Of course, as we have seen, fairly strong assumptions are needed to identify these effects from the data, even the average of these effects for a particular population. When assumptions (A2.1)–(A2.4) hold, then the average controlled direct effect and average natural direct and indirect effects, conditional on covariates $C = c$, are identified from the observed data by

$$\begin{aligned}
\mathbb{E}[Y_{am} - Y_{a^*m}|c] &= \mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] \\
\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] &= \sum\nolimits_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c) \\
\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] &= \sum\nolimits_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\} \qquad \textbf{(2.6)}
\end{aligned}$$

The proofs for these equalities are given in the Appendix (cf. Pearl, 2001). In each case the left-hand side of the equation is a counterfactual quantity. The right-hand side of the equation is something we can estimate from the observed data. Under our no-unmeasured confounding assumptions the left-hand side and the right-hand side are equal and we can estimate the direct and indirect effects from data. The

expression for the controlled direct effect requires only assumptions (A2.1) and (A2.2); the expressions for the natural direct and indirect effects require assumptions (A2.1)–(A2.4). What the methods and models considered in the previous section do is essentially specify a model for the outcome $\mathbb{E}[Y|a,m,c]$ and for the mediator $P(m|a,c)$ and then calculate analytically the expression on the right-hand side of the equations above. If we change the models, then we have to go through a new calculation. However, because the expressions above are very general and do not presuppose any particular statistical model, they are sometimes said to be "nonparametric." This counterfactual approach is thus completely general in terms of the models that it can accommodate. Once we specify any particular model for the outcome and the mediator, we can proceed with calculating the direct and indirect effects. We could specify different models involving quadratic terms, or other interactions or further nonlinearities, and we could still use the formulae in (2.6) to derive the direct and indirect effects. However, each time a different functional form for the model is considered, we would need to do a new derivation. In Sections 2.2, 2.4, and 2.5, we considered a number of standard cases involving continuous or dichotomous mediators, with continuous, dichotomous, or count outcomes, and allowing for exposure–mediator interactions. Because of the generality, the expression on the right-hand side for the natural indirect effect is sometimes referred to as the mediation formula: $\sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}$ (Pearl, 2012a). Both the regression-based analytic approach described above and the simulation-based approach described below rely on this mediation formula in the estimation of natural indirect effects.

2.17. AN ALTERNATIVE REGRESSION-BASED ESTIMATION APPROACH USING SIMULATIONS

As noted above, the approach we have been describing specifies a model for the outcome and a model for the mediator and then uses the empirical expressions above to obtain the natural direct and indirect effect estimates. A downside of this approach is that each time the model is changed, a new formula has to be analytically obtained. An alternative approach is to use simulations to estimate the effects in (2.6) above. This approach was proposed by Imai et al. (2010a) and can accommodate a number of very flexible models. The approach has this flexibility because it is not necessary to derive a new formula each time the model changes since the effects are estimated using simulations instead. The approach essentially consists of fitting models for the outcome $P(y|a,m,c)$ and for the mediator $P(m|a,c)$ and then simulating potential outcomes under the equivalent of assumptions (A2.1)–(A2.4) above (cf. Imai et al., 2010a; Shpitser and VanderWeele, 2011). The models that are fit for the outcome $P(y|a,m,c)$ and for the mediator $P(m|a,c)$ can be linear, or logistic, or count outcomes; they could involve quadratic terms, arbitrary interactions between the exposure and mediators and covariates; they could also involve splines. The no-confounding assumptions that are required for the simulation-based approach and the regression-based approach are essentially the same—that is, assumptions (A2.1)–(A2.4). These assumptions can be

formally stated in slightly different ways (see Chapter 7), but they all essentially require that control has been made for [assumption (A2.1)] exposure–outcome confounding, [assumption (A2.2)] mediator–outcome confounding, and [assumption (A2.3)] exposure–mediator confounding and [assumption (A2.4)] that there is no mediator–outcome confounder that is itself affected by the exposure.

Two simulation approaches, with varying degrees of computational demands, are possible. The "parametric" approach consists first of fitting models for the mean outcome $P(y|a,m,c)$ and for the mediator $P(m|a,c)$. Once the models are fit to the data, the approach takes a random draw of the parameters for these two models using their sampling distribution once the models have been fit. For each individual, with their own baseline covariate C, some number K copies of the mediator are then simulated under each possible exposure a and a^* being compared, based on the parameters of the model from the random draw. Then based on these simulated values, values of the counterfactual outcomes for Y are simulated for each treatment exposure level under each of the simulated mediators. These are averaged across the K copies and all of the individuals in the sample to obtain an estimate of the natural direct and indirect effect. Then a new random draw of the parameters for the mediator and outcome models is taken using their sampling distribution, and the process repeats itself until some number J estimates of the natural direct and indirect effects are obtained. The mean of these J estimates can be used as the actual estimate of the natural direct and indirect effects, and the 2.5th and 97.5th percentiles of these values could be used as a 95% confidence interval for the natural direct and indirect effects. See Imai et al. (2010a) for further details.

A second, but more computationally involved, "nonparametric" approach is also possible. This approach has the advantage of not needing to know the sampling distribution of the parameters of the models for the outcome $P(y|a,m,c)$ and for the mediator $P(m|a,c)$; it can thus be used if semiparametric or nonparametric models are used for the mediator and the outcome. This approach involves drawing J bootstrapped samples with replacement from the original data. For each bootstrapped sample, the models for the outcome $P(y|a,m,c)$ and for the mediator $P(m|a,c)$ are fit to the data from that bootstrapped sample. Then once these models are fit, for each individual, with their own baseline covariate C, some number K copies of the mediator are then simulated under each possible exposure a and a^* being compared, using the models that have been fit to the data. Then based on these simulated values, values of the counterfactual outcomes for Y are simulated for each exposure level under each of the simulated mediators. These are averaged across the K copies and all of the individuals in the sample to obtain an estimate of the natural direct and indirect effect. This is done with each of the J bootstrapped samples. The mean of these estimates from the J bootstrapped samples can be used as the actual estimate of the natural direct and indirect effects and the 2.5th and 97.5th percentiles of these values could be used as a 95% confidence interval for the natural direct and indirect effects. The parametric approach and the nonparametric approach are thus quite similar but account for sampling variability in different ways. The parametric approach does so by sampling parameters J times from their fitted distribution once the models have been fit to the actual data. The nonparametric approach accounts for sampling variability by fitting the models to J different

bootstrapped samples. The parametric approach has the advantage that the models only need to be fit to the data once; the nonparametric approach has the advantage that it is not necessary to know the sampling distribution of the parameter estimators, and thus it can be used on a wider class of models. See Imai et al. (2010a) for further details.

The advantage of this simulation-based approach over the analytic regression-based approach that we have been considering is that the simulation-based approach can handle a wider range of models for the mediator and the outcome and more flexibly and easily accommodate quadratic terms, exposure–covariate interaction, or mediator–covariate interactions or other nonlinearities. For the analytic regression-based approach considered in previous sections, each time the model is changed for the mediator or the outcome from the basic forms considered above, new formulas have to be derived. The simulation-based approach circumvents this issue by not relying on analytic formulas but rather on simulations. The downside of the simulation approach, however, is that because it relies on simulations, it can be very computationally intensive. It can work quite well with smaller datasets, but with larger datasets the time required to obtain estimates and confidence intervals could be prohibitive. For example, in analysis of Ananth and VanderWeele (2011) that we will consider in the next chapter, 26 million observations (all US births between 1995 and 2002) were used in the analysis. For a dataset of this size, the analytic regression-based approach can still be employed, but the computational time for the simulation-based approach would probably amount to months or even years.

The simulation-based approach could in principle be used to estimate direct and indirect effects with both difference and ratio measures, but its current implementation (described in the following section) focuses on the difference scale, even with binary outcomes.

2.18. CODE FOR THE SIMULATION-BASED APPROACH IN R

Quite general software is now available to implement this simulation-based approach in R (Imai et al., 2010b; Tingley et al., 2014) under a wide range of models. Software in Stata (Hicks and Tingley, 2011) is also available to implement the simulation-based approach; however, at the time of the writing of this book, the functionality of the Stata version was more limited than that of the R version of the software. See also Daniel et al. (2011) for similar software in Stata.

To use the R package, it is first installed in the R system:

```
install.packages(''mediation'')
```

Two libraries are used in the code below which can be loaded by

```
library(mediation)
library(sandwich)
```

Following Imai et al. (2010b), suppose we were interested in assessing the extent to which the effect of the way a particular immigration issue was framed in a video (the exposure) on the participants' agreeing to write a letter to their congressmen was mediated by the level of anxiety that the video produced. Suppose we had a dataset named "framing" with an exposure variable "treat" that was binary, a mediator variable "emo" that was continuous, and an outcome variable "cong_mesg" that was binary as well as four covariates "age," "educ," "gender," and "income." We could first fit a linear regression for mediator model in R using

```
med.fit {<}- lm(emo \symbol{126}treat + age + educ + gender
  + income, data = framing)
```

We could then fit a probit model say for the outcome using

```
out.fit {<}- glm(cong\_mesg \symbol{126}emo + treat + age +
  educ + gender + income, data = framing, family
  = binomial(‘‘probit’’))
```

Note that the probit model is not covered in the SAS and SPSS macros described above whereas it (and numerous others) can be used via the simulation-based approach in R. The simulation-based approach to mediation can then be implemented by using the "mediate" command from the mediation package:

```
med.out {<}- mediate(med.fit, out.fit, treat = ‘‘treat’’,
  mediator = ‘‘emo’’, robustSE = TRUE)
```

Note that the inputs to the "mediate" command are the mediator model (med.fit), the outcome model (out.fit), and the names of the treatment and mediator variables. The final option "robustSE = TRUE" tells R to use robust standard errors to protect against heteroskedasticity (i.e., heterogeneous variance of the error term).

The results of the mediation analysis can then be obtained using the command

```
summary(med.out)
```

The output then gives the (pure and total) natural indirect effects (along with the proportion mediated under each, and the averages of these effects), the (pure and total, and their average) natural direct effects, and the total effect, along with the sample sized used and the number of simulations used. See the SAS macro section for the description of pure versus total direct and indirect effects. For example, using the dataset described above, the results would look as follows:

```
Estimate 95% CI Lower 95% CI Upper p-value
Mediation Effect_0 0.0826 0.0336 0.1427 0.01
Mediation Effect_1 0.0830 0.0341 0.1423 0.00
Direct Effect_0 0.0120 -0.1041 0.1383 0.80
Direct Effect_1 0.0123 -0.1137 0.1501 0.81
Total Effect 0.0950 -0.0377 0.2372 0.16
Proportion via Mediation_0 0.7895 -4.3005 5.3631 0.37
Proportion via Mediation_1 0.8078 -3.9146 4.9649 0.33
Mediation Effect (Ave.) 0.0828 0.0336 0.1412 0.00
Direct Effect (Ave.) 0.0122 -0.1089 0.1452 0.80
Proportion via Mediation (Ave.) 0.7987 -4.1202 5.1650 0.35
Sample Size Used: 265
Simulations: 1000
```

The number of simulations can be changed in the software options. Also the software has the flexibility of accommodating types of models other than the linear regression model used for the mediator and the probit model used for the outcome above. Various interactions or quadratic terms could also be included in these models. The user simply needs to change the form of these models in the R code above. The software also offers some functionality for sensitivity analysis for unmeasured confounding, a topic we turn to in the next chapter. For further information and examples using the software, the reader is referred to Imai et al. (2010b) and Tingley et al. (2014) and, for more updated information, to the documentation at the website: http://cran.r-project.org/web/packages/mediation/vignettes/mediation.pdf.

2.19. DISCUSSION

This chapter has described several regression-based approaches to estimating direct and indirect effects either using analytic formulae or using a simulation-based approach. The approach using analytic formulae is applicable to binary, continuous, and count outcomes and binary and continuous mediators; it can also allow for exposure–mediator interaction, extending traditional approaches in the social sciences. The simulation-based approach can flexibly handle an even broader class of models allowing for interactions with covariates, for quadratic terms, or for other nonlinearities. In subsequent chapters we will consider other types of methods, including methods for time-to-event outcomes in Chapter 4 and methods for multiple mediators in Chapter 5. The concepts of this chapter, however, will lay the foundation for these subsequent developments.

We have seen that to estimate direct and indirect effects, several strong assumptions are needed about confounding. When mediation is of interest, special attention in particular needs to be given in controlling for mediator–outcome confounding. The estimates from the product method or difference method or the methods from the causal inference literature considered here will all be biased if control is not made for these variables. Mediator–outcome confounding can be present even if the exposure is randomized (since the mediator is not randomized). Unfortunately,

this point was not made in the popular Baron and Kenny (1986) paper, though it was made by Judd and Kenny (1981) five years earlier and it has now been emphasized and clarified in the causal inference literature and is being emphasized again in the psychology literature. Psychologists, social scientists, and biomedical researchers need to take this assumption seriously if they hope to obtain valid conclusions about direct and indirect effects. If the investigator thinks that unmeasured confounding may be present, sensitivity analysis should be used (VanderWeele, 2010b; Imai et al. 2010a), the topic we now turn to in the next chapter.

3

Sensitivity Analysis for Mediation

Unmeasured or uncontrolled confounding is a common problem in observational studies in the biomedical and social sciences. Unmeasured confounding is a challenge to observational research even in the analysis of total effects, when we are not specifically interested in assessing pathways and direct and indirect effects. In the analysis of total effects, attempts are made to collect data on and control for as many pre-exposure covariates as possible that are related to the exposure and the outcome of interest. Often, however, one or more important covariates will remain unmeasured that might confound the relationship between the exposure and the outcome. Such unmeasured confounding variables can bias estimates of the effect of the exposure on the outcome. Moreover, as was seen in the previous chapter, when we are interested in pathways and direct and indirect effects, the assumptions about confounding that are needed to identify these direct and indirect effects are even stronger than for total effects. We might often, perhaps almost always, be worried that these assumptions are violated and that our estimates are biased.

Sensitivity analysis techniques can help assess how robust results are to violations in the assumptions being made. In the case of sensitivity analysis for unmeasured confounding, these techniques assess the extent to which an unmeasured variable (or variables) would have to affect both the exposure and the outcome in order for the observed associations between the two to be attributable solely to confounding rather than a causal effect of the exposure on the outcome. Sensitivity analysis can also be useful in assessing a plausible range of values for the causal effect of the exposure on the outcome corresponding to a plausible range of assumptions concerning the relationship between the unmeasured confounder and the exposure and outcome.

An early application of this approach was the work of Cornfield et al. (1959), who showed that the association between smoking and lung-cancer was unlikely to be entirely due to unmeasured confounding by unknown genetic factors, as had been proposed by Fisher (1958). As another example, a pregnant mother's receiving adequate prenatal care during pregnancy, is associated with higher birth weight of an infant, but we might wonder whether this association might be explained by the fact that mothers with adequate prenatal care are often

of a higher socioeconomic background or perhaps more conscientious generally, rather than that there is a actual causal effect of prenatal care itself.

Sensitivity analysis techniques can also be helpful in the analysis of direct and indirect effects, where the confounding assumptions are even stronger. The purpose of this chapter is to present some simple sensitivity analysis techniques that can be used to assess how robust results are to various biases. The term "sensitivity analysis" is also used outside the context of unmeasured confounding bias. The term and a similar sort of approach is often used to address other types of bias such as selection bias and measurement error bias. The term "bias analysis" has been used more recently to describe this whole range of techniques. We will begin with sensitivity analysis for unmeasured confounding for total effects; we will then discuss analogues of these techniques for controlled direct effects and natural direct and indirect effects. Toward the end of the chapter we will also consider some simple sensitivity analysis techniques that can be used for direct and indirect effects when measurement error in the mediator may be present.

3.1. SENSITIVITY ANALYSIS FOR UNMEASURED CONFOUNDING FOR TOTAL EFFECTS

In this section we will consider sensitivity analysis techniques for total effects. In the following sections we will extend our discussion to direct and indirect effects. One approach to sensitivity analysis is to assume an measured confounder U that affects both the exposure A and the outcome Y such that if control were made for both the measured covariates C and the unmeasured confounder U, then this would suffice to control for confounding. However, since no data are available on U, the observed effect estimates are subject to confounding bias. Such an unmeasured confounder is represented in Figure 3.1.

The basic idea of sensitivity analysis is to specify parameters corresponding to the relationships between U and Y and between U and A and from these, along with the observed data, to obtain "corrected" effect estimates corresponding to what would have been obtained had control been made for U and not only C. One may not believe any particular specification of these parameters, but the parameters themselves could be varied across a plausible range of values to see how the "corrected" estimate of the effect varies and to obtain a plausible range of estimates

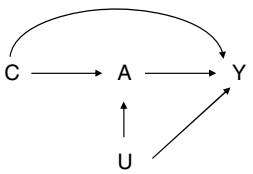

Figure 3.1 An unmeasured confounder U that affects exposure A and outcome Y.

for the causal effect. If all of this range indicates a substantial effect, we may be somewhat more confident that the results are not driven solely by confounding.

Many sensitivity analysis techniques for unmeasured confounding for total effects have been developed (Kitagawa, 1955; Cornfield et al., 1959; Bross, 1966; Schlesselman, 1978; Breslow and Day, 1980; Rosenbaum and Rubin, 1983a; Yanagawa, 1984; Gail et al., 1988; Flanders and Khoury, 1990; Copas and Li, 1997; Lin et al., 1998; Scharfstein et al., 1999; Robins et al., 2000b; Rosenbaum, 2002; Imbens, 2003; Greenland, 2003, 2005; Stürmer et al., 2005; McCandless et al., 2007; VanderWeele and Arah, 2011). Here we will consider an easy-to-use technique under simplifying assumptions. The results essentially just compare (i) what one obtains adjusting only for measured covariates C with (ii) what one would have obtained had it been possible to adjust for measured covariates C and unmeasured covariate(s) U. If it is thought that adjusting for C and U together would suffice to control for confounding, then we may also interpret the results as comparing (i) the effect estimate that is obtained adjusting only for measured covariates C versus (ii) the true causal effect.

3.1.1. Continuous Outcomes

We will first consider results for a continuous outcome Y and then afterwards for a dichotomous outcome Y. Suppose then we have obtained an estimate of the effect of the exposure A on the outcome Y conditional on measured covariates C using regression or propensity score analysis (Rosenbaum and Rubin, 1983b) or some other technique. We will define the bias factor $B_{add}(c)$ on the additive scale as the difference between (i) the expected differences in outcomes comparing $A = a$ and $A = a^*$ conditional on covariates $C = c$ and (ii) what we would have obtained had we been able to adjust for U as well. If the exposure is binary, then we simply have $a = 1$ and $a^* = 0$. A very simple approach to sensitivity analysis is possible if we are willing to assume that [assumption (A3.1)] U is binary and [assumption (A3.2)] that the effect of U (on the additive scale) is the same for those with exposure level $A = a$ and exposure level $A = a^*$ (i.e., no $U \times A$ interaction). If these assumptions hold, let γ be the effect of U on Y conditional on A and C, that is,

$$\gamma = \mathbb{E}(Y|a, c, U = 1) - \mathbb{E}(Y|a, c, U = 0)$$

Note that by assumption (A3.2), $\gamma = \mathbb{E}(Y|a, c, U = 1) - \mathbb{E}(Y|a, c, U = 0)$ is the same for both levels of the exposure of interest. Note also that γ is the effect of U on Y already having adjusted for C; that is, in some sense the effect of U on Y not through C (VanderWeele and Arah, 2011). Now let δ denote the difference in the prevalence of the unmeasured confounder U for those with $A = a$ versus those with $A = a^*$, that is,

$$\delta = P(U = 1|a, c) - P(U = 1|a^*, c)$$

Under assumptions (A3.1) and (A3.2), the bias factor is simply given by the product of these two sensitivity analysis parameters:

$$B_{add}(c) = \gamma\,\delta \tag{3.1}$$

Thus to calculate the bias factor we only need to specify the effect of U on Y and the prevalence difference of U between the two exposure groups and then take the product of these two parameters. Once we have calculated the bias term $B_{add}(c)$, we can simply estimate our causal effect conditional on C and then subtract the bias factor to get the "corrected estimate"—that is, what we would have obtained if we had controlled for C and U. Under these simplifying assumptions (A3.1) and (A3.2), we can also get adjusted confidence intervals by simply subtracting $\gamma\,\delta$ from both limits of the estimated confidence intervals.

We may not believe any particular specification of the parameters γ and δ, but we could vary these parameters over a range of plausible values to obtain what were thought to be a plausible range of corrected estimates. The range of the values over which the sensitivity analysis parameters are varied could be determined by substantive knowledge or other prior studies that may have reported estimates of the associations of the covariates with the outcome. Using this technique, we could also examine how substantial the confounding would have to be to explain away an effect (we could do this for the estimate and confidence interval).

The simple result in (3.1) that $B_{add}(c) = \gamma\,\delta$ is given in a number of places in the literature (e.g., Cochran, 1938; Draper and Smith, 1981; Lin et al., 1998) and was shown to hold if the initial estimates were obtained using regression with main effects for A and C in the regression model. In fact, it can be shown that the result holds much more generally and applies irrespective of whether regression, propensity score analysis, inverse probability weighting or some other technique were used to obtain the initial estimate (VanderWeele and Arah, 2011). It can be used irrespective of the method or model used to produce the initial effect estimate conditional on C, provided that assumptions (A3.1) and (A3.2) hold. Two further extensions are also worth noting. First, assumption (A3.1) that U is binary can be relaxed somewhat; if U is assumed continuous, then the bias formula in (3.1) $B_{add}(c) = \gamma\,\delta$ will still hold if the sensitivity analysis parameter $\delta = P(U = 1|a, c) - P(U = 1|a^*, c)$ for a binary unmeasured confounder is replaced by $\delta = \mathbb{E}(U|a, c) - \mathbb{E}(U|a^*, c)$ for a continuous unmeasured confounder—that is, the difference in means of U for those with exposure levels $A = a$ versus $A = a^*$. Second, the bias formula above and the simple approach of subtracting $B_{add}(c) = \gamma\,\delta$ from the estimate and both limits of the confidence interval apply to effect estimates conditional on a particular value of the covariates $C = c$. This is what is obtained from a regression analysis. Often, we might be interested in effects averaging over the covariates. If there is no interaction between the exposure A and the covariates C in a linear regression, then these conditional and marginal effects are one and the same. Otherwise, we may need to, for the conditional effect estimates, average over the distribution of the covariates C. When we use the sensitivity analysis technique above for marginal effects (averaged over the covariates), if we specify the same parameters γ and δ for each strata of the covariates, then the exact same approach as that described above still

applies: We can simply subtract $\gamma\,\delta$ from the estimate and both limits of the confidence interval to obtain corrected estimates and confidence intervals. If we want to specify different values of the sensitivity parameters, γ and δ, for different levels of the covariates, then an alternative approach, described below, is needed. First, however, we will illustrate in the next subsection the current approach that can be employed when only conditional effects are being estimated or when we specify the sensitivity analysis parameters as being the same across all strata of the covariates.

A similar approach for sensitivity analysis also works when the effect of the exposure on the exposed (also called the effect of treatment on the treated) is of interest, rather than the effect of exposure in the overall population. When conducting sensitivity analysis for the effect of the exposure on the exposed, one does not need the assumption that the effect of U on the outcome Y is equal for both exposure groups (VanderWeele and Arah, 2011). For the effect of the exposure on the exposed, the sensitivity analysis parameter γ is simply specified as the effect of U on Y among the unexposed [i.e., $\gamma = \mathbb{E}(Y|A = 0, c, U = 1) - \mathbb{E}(Y|A = 0, c, U = 0)$], and then the same approach as described above can be used in sensitivity analysis with the parameter γ so defined. Likewise, a similar approach can be used in sensitivity analysis for the effect of the exposure among those who are actually unexposed, by redefining the sensitivity analysis parameter γ as the effect of U on Y among the exposed [i.e., $\gamma = \mathbb{E}(Y|A = 1, c, U = 1) - \mathbb{E}(Y|A = 1, c, U = 0)$].

3.1.2. Example of Sensitivity Analysis for Continuous Outcomes

Consider a study by Reinisch et al. (1995) that examined the effect of in utero exposure to phenobarbital on intelligence in men. Subjects were selected from the largest hospital in Copenhagen. The exposure group consisted of those who had been exposed in utero to phenobarbital ($A = 1$) and the control group of those who had not ($A = 0$). Propensity-score matching and regression techniques (Rosenbaum and Rubin, 1985) were used to adjust for background characteristics in making intelligence comparisons using the Danish Military Board Intelligence Test (Y) taken by the exposed and unexposed men when they had reached their early twenties. The background characteristics (C) for which adjustment was made included: family socioeconomic status; breadwinner's education; sibling position; whether the pregnancy was wanted; whether the mother attempted an abortion; maternal marital status; predisposing risk score indicating that conditions were less than optimal for conception; mother's age; father's age; gestational length; birth weight; birth length; number of cigarettes per day in the third trimester; maternal weight gain divided by height cubed; and the maternal complaint score.

Subjects exposed to phenobarbital were found to have significantly lower scores on the Danish Military Board Intelligence Test than they would have had they not been exposed. Specifically, under the assumption that the effect of exposure is unconfounded given C, Reinisch et al. (1995) obtained an estimate of the effect of the exposure on the exposed of -4.77 (95% CI $=-7.96$ to -1.58). These authors suggest that parental intelligence, which was not measured in the study, may partially confound the analysis. They reason informally, without a quantitative analysis,

to argue that it is unlikely that parental intelligence, rather than drug effects, are responsible for the observed intelligence deficits. Using the sensitivity analysis technique above, if we hypothesize an unmeasured confounding variable U of, say, the average of maternal and paternal intelligence measured by the Danish Military Board Intelligence Test, and we assume that if unexposed, a one-point increase in U would on average result in a 0.3 point increase in Y so that $\gamma = (0.3)$, then it follows from equation (3.1) above that $B_{add}(c) = (0.3)\delta$. It would then require a difference in parental intelligence of $\delta = -4.77/(0.3) = -15.9$ between the parents of the exposed and unexposed on the Danish Military Board Intelligence Test to completely explain away the estimated deficit. A standard deviation for a national sample of subjects taking the Danish Military Board Intelligence Test was 11.38. A 1.3-standard-deviation difference in parental intelligence between the exposed and unexposed, although not entirely implausible, may seem unlikely. In spite of the possibility of unmeasured confounding, it seems that there is still some evidence for an effect.

3.1.3. Sensitivity Analysis for a Continuous Outcome with Different Sensitivity Analysis Parameters for Different Covariate Values

Suppose now that instead of focusing on effects conditional on a particular covariate value $C = c$ or specifying the sensitivity analysis parameters γ and δ to be the same for each covariate C, we were interested in the overall marginal effect averaged over the covariates and we wanted to specify different sensitivity analysis parameters for different covariate levels. Suppose then for each level of the covariates of interest $C = c$ we specified a value for the effect of U on Y [i.e., $\gamma(c) = \mathbb{E}(Y|a, c, U = 1) - \mathbb{E}(Y|a, c, U = 0)$] and also a value for the prevalence difference of U between those with exposure status $A = a$ and $A = a^*$ and covariates $C = c$ [i.e., $\delta(c) = P(U = 1|a, c) - P(U = 1|a^*, c)$]. We could then obtain an overall bias factor, B_{add}, by taking the product of the bias factors in each strata of C and then averaging these over C, weighting each strata of C according to what proportion of the sample was in that strata. The overall bias factor is then

$$B_{add} = \sum\nolimits_c \{\gamma(c)\delta(c)\}P(C = c)$$

We could then subtract this overall bias factor from our estimate adjusted only for C to obtain a corrected estimate. In this case, however, we can now longer simply subtract the bias factor from both limits of the confidence intervals because this does not take into account the variability in our estimates of the proportion of the sample in each strata of the covariates $P(C = c)$. Corrected confidence intervals could instead be obtained by bootstrapping.

EXAMPLE: PERINATAL EPIDEMIOLOGY

We illustrate the approach of specifying different sensitivity analysis parameters for different covariate values using an example from perinatal epidemiology. We use data from the National Center for Health Statistics (NCHS) Birth Certificate Files for year 2000 to consider the effect of adequate prenatal care on birth weight. Prenatal care was classified as adequate versus inadequate, as defined by a modification

of the Adequacy of Prenatal Care index (Kotelchuck, 1994; VanderWeele et al., 2009). Baseline covariates in the NCHS data include maternal age, race, place of birth, place of residence, education, marital status, plurality, gravidity, prior preterm birth, prior birth greater than 4000 grams, alcohol consumption, and tobacco use. Socioeconomic status might serve as an unmeasured confounding variable for the relationship between prenatal care and birth weight. Although education provides one measure of socioeconomic status, there are likely aspects of socioeconomic status that education does not capture. The relationships between socioeconomic status and birthweight, as well as between socioeconomic status and adequate prenatal care, may vary by age. For example, the relationship between socioeconomic status and adequate prenatal care may be weaker for those age 19 years or under at the time of childbirth than for those at older ages, due to State Children's Health Insurance Programs (SCHIP). On the other hand, the birthweight of infants of younger mothers may be more sensitive to adverse socioeconomic circumstances. Estimates of the effect of prenatal care on birth weight, stratified by age ($\leq$ 19 versus $>$ 19 years), were obtained using NCHS data for year 2000, controlling for the aforementioned measured covariates. The estimated effect for younger mothers was 82.4 grams (95% CI = 78.2 to 86.7) and for older mothers was 78.3 grams (95% CI= 76.3 to 80.3); 12.2% of the mothers were age 19 years or under, and 87.8% were older than 19; and thus the overall estimate of the effect is $(0.122)(82.4) + (0.878)(78.3) = 78.8$ grams (95% CI = 76.4 to 81.2). Let U denote adverse versus adequate socioeconomic status. Suppose we hypothesize that the average effect of adverse socioeconomic status for younger mothers is 120 grams and the average effect for older mothers is 80 grams whereas the difference in the likelihood of adverse socioeconomic status, comparing those with adequate versus inadequate prenatal care, was only 20% for young mothers (due to SCHIP); however, for older mothers, comparing those receiving adequate versus inadequate prenatal care, the difference was 50%. Our bias factor for younger mothers is $(120)(0.2) = 24$ and the bias factor for the older mothers is $(80)(0.5) = 40$. The overall bias factor is then $24(.122) + 40(.878) = 38.0$. Subtracting this from our initial value, we would have a corrected estimate of effect of 40.8 grams (95% CI = 38.3 to 43.2). In this case the sample size is very large (approximately 4 million births) and the confidence interval obtained from bootstrapping versus simply subtracting the bias factor from the observed limits of the confidence interval will almost coincide.

Other values for the sensitivity analysis parameters could similarly be considered. For example, if the effect sizes for the unmeasured confounder were changed to 280 grams and 160 grams for younger and older mothers, respectively, this would bring the confidence interval down to include zero, with a corrected estimate of 1.2 grams (95% CI = -0.8 to 3.2). Here, in contrast to the first example in Section 3.1.2, the values of the sensitivity analysis parameters needed to completely explain away the effect are perhaps not as implausible.

Here we have allowed the effect of the unmeasured confounder and prevalence of the unmeasured confounder to vary with age. We could also allow it to vary with other covariates; the sensitivity analysis becomes more complicated and we have to specify more parameters but the same principles apply. Assumptions (A3.1) and

(A3.2) could also be relaxed; for example, we could allow the effect of U on Y to be different for different exposure groups, but this requires specifying an even larger number of sensitivity analysis parameters. The interested reader is referred elsewhere for these techniques (VanderWeele and Arah, 2011). In general there is tension between a sensitivity analysis that is (i) simple and easy to use and interpret without requiring too many parameters and one that is (ii) general without making many assumptions. In observational studies, unmeasured confounding is frequently a problem and at least the simple approach described above should be considered. Depending on the problem and how sensitive conclusions seem to be, a more nuanced approach involving more sensitivity analysis parameters can also be used.

3.1.4. Sensitivity Analysis for Unmeasured Confounding for a Total Effect with a Binary Outcome

For a binary outcome, the same approach as described above could be employed if risk differences were estimated. Often, however, when an outcome is binary, a logistic regression model is fit to the data and effect measures on the odds ratio scale are estimated. In this subsection we will describe a sensitivity analysis approach that can be used to assess the impact of unmeasured confounding when the odds ratio scale is used for a binary outcome and the outcome is rare. The techniques described here will apply to both risk ratio and odds ratios with a rare outcome (odds ratios then approximate risk ratios). In general if an outcome is common, the odds ratio scale is often best avoided because the odds ratio is often, inappropriately, interpreted as a risk ratio. The approach we describe is for odds ratios or risk ratio estimates of the effect of the exposure A on outcome Y, conditional on the covariates C. This is what is obtained from a logistic regression model.

We will now define the bias factor $B_{mult}(c)$ on the multiplicative scale as the *ratio* of (i) the risk ratio (or odds ratio, with a rare outcome) comparing $A = a$ and $A = a^*$ conditional on covariates $C = c$ and (ii) what we would have obtained as the risk ratio (or odds ratio) had we been able to condition on both C and U. We now make the simplifying assumptions that [assumption (A3.1)] U is binary and that [assumption (A3.2b)] the effect of U (on the risk ratio scale) is the same for those with exposure level $A = a$ and exposure level $A = a^*$ (i.e., no $U \times A$ interaction on the risk ratio scale). If these assumptions hold, we will let γ be the effect of U on Y conditional on A and C on the risk ratio scale, that is,

$$\gamma = \frac{P(Y = 1|a, c, U = 1)}{P(Y = 1|a, c, U = 0)}$$

Note that by assumption (A3.2b), $\gamma = \frac{P(Y=1|a,c,U=1)}{P(Y=1|a,c,U=0)}$ is the same for both levels of the exposure. Again this is the effect of U on Y having already adjusted for C; in some sense this is the effect of U on Y not through C. Under assumptions (A3.1) and (A3.2b), the bias factor on the multiplicative scale is simply given by

$$B_{mult}(c) = \frac{1 + (\gamma - 1)P(U = 1|a, c)}{1 + (\gamma - 1)P(U = 1|a^*, c)} \tag{3.2}$$

We can thus use the bias formula by specifying the effect of U on Y on the risk ratio scale and the prevalence of U amongst those with exposure levels $A = a$ and $A = a^*$. The result above is given by Schlesselman (1978), building on the work of Bross (1966). Once we have calculated the bias term $B_{mult}(c)$, we can simply estimate our risk ratio controlling only for C (e.g., if the outcome is rare, we just fit a logistic regression) and we *divide* our estimate by $B_{mult}(c)$ to get the corrected estimate for risk ratio—that is, what we would have obtained if we had adjusted for U as well. Under the simplifying assumptions of (A3.1) and (A3.2b), we can also obtain corrected confidence intervals by dividing both limits of the confidence interval by $B_{mult}(c)$. Assumptions (A3.1) and (A3.2b) can be relaxed (VanderWeele and Arah, 2011), but then the resulting formulae for the bias factor become more complicated. Note that to use the bias factor in (3.2), we must now specify the prevalence of the unmeasured confounder in both exposure groups [i.e., $P(U = 1|a, c)$ and $P(U = 1|a^*, c)$], not just the difference between these two prevalences as with the bias factor in (3.1) for outcomes on the additive scale.

EXAMPLE: BREASTFEEDING

We illustrate this approach for binary outcomes using analyses presented by Moorman et al. (2008), who report on associations between ovarian cancer and breastfeeding. For premenopausal women, with no breastfeeding as a reference group, Moorman et al. (2008) find an odds ratio for ovarian cancer of 0.5 (95% CI: 0.3, 0.8) for those who women who breastfed 6–12 months versus not at all. Moorman et al. (2008) control for a number of covariates but do not control for socioeconomic status (SES), which is often thought to be a confounder in associations between breastfeeding and health outcomes. Let $U = 1$ denote low SES (versus high SES as the reference). Suppose we thought that low SES increased the risk of ovarian cancer by 1.5-fold and that 30% of the 6- to 12-month breastfeeding group was low SES but 70% of the reference group (no breast-feeding) was low SES. From (3.2), we would obtain a bias factor of $B_{mult}(c) = [1 + (1.5 - 1)(0.3)]/[1 + (1.5 - 1)(0.7)] = 0.85$. If we divide the observed estimate and the confidence interval by this bias factor, we obtain a corrected estimate of 0.6 (95% CI: 0.4, 0.9). Suppose instead we thought that low SES increased the risk of ovarian cancer by 2.5-fold and that 30% of the 6- to 12-month breastfeeding group was low SES but 70% of the reference group (no breastfeeding) was low SES. The bias factor would then be $B_{mult}(c) = [1 + (2.5 - 1)(0.3)]/[1 + (2.5 - 1)(0.7)] = 0.71$; and if we divide the observed estimate and confidence interval by this bias factor, we obtain a corrected estimate of 0.7 (95% CI: 0.4, 1.1). The estimate still appears protective, but the confidence interval now does contain 1.

3.1.5. Different Approaches to Sensitivity Analysis

An objection is sometimes made to sensitivity analysis that the exercise itself is too subjective. It is too subject to the values of the sensitivity analysis parameters that an investigator specifies. This is certainly the case to some degree, and the value of the sensitivity analysis does in part depend on the investigators' level of integrity and

commitment to adequately assessing robustness. Nonetheless, several approaches to sensitivity analysis can help give the sensitivity analysis techniques a somewhat more objective character. First, an investigator can give sensitivity analysis parameter values that would suffice to explain away the effect estimate (i.e., reduce it to zero). A reader can then judge whether he or she thinks it plausible that an unmeasured confounder of that magnitude is likely present. Second, an investigator might consider producing a table of "corrected" estimates using a wide range of all sensitivity analysis parameters, including values which the investigator believes are unreasonably large. Then, even if a reader might disagree with the investigator as to what range of the parameters is reasonable, the "corrected" estimates corresponding to the values of the sensitivity analysis parameters that the reader believes plausible would still be in the table and the reader could judge for himself or herself how sensitive the conclusions of the study in fact are to unmeasured confounding. Third, an investigator could consider what the sensitivity analysis parameters are for the most important *measured* confounder. This could be done by carrying out the analysis multiple times in each instance omitting a single covariate and identifying the covariate for which the effect estimate changes the most. The effect of this covariate on the outcome and how the distribution of this covariate differs for the exposed and unexposed could be assessed. Those same sensitivity analysis parameter values could then be used in sensitivity analysis to assess how an unmeasured confounder of similar importance to the most important measured confounder would change estimates. If with the sensitivity analysis parameters so chosen the estimate and its confidence interval still indicate an effect, then the investigator could at least claim that an unmeasured confounder would have to be stronger than the most important measured confounder to explain away the effect estimate. The adequacy of this third approach of course depends in part on how satisfactory the set of measured confounders in fact is. Note that all three of the approaches above do not necessarily require the investigator to know what the unmeasured confounder is. Finally, in settings in which it is clear what the unmeasured confounding variable is, it may be possible to obtain estimates or ranges of the sensitivity analysis parameters from other published papers that did collect data on the unmeasured confounder or from expert opinion. There is still some subjectivity in this fourth approach as the parameters may still depend on which papers an investigator happens to look at. There is also still some subjectivity in all of the approaches in attempting to judge whether unmeasured confounding of a particular magnitude is or is not plausible, but these various approaches can help give the sensitivity analysis a more objective character as compared with simply having the investigator choose a few parameters that conveniently still give a significant corrected estimate and then claim "see—there is still an effect!" An investigator could potentially use all four of these approaches in carrying out and discussing the results of sensitivity analysis. At a minimum, it may be useful to present (i) the sensitivity analysis parameters that would suffice to completely explain away an effect and also (ii) the sensitivity analysis parameters that would be required to shift the confidence interval to just include the null. Such information can be presented in a paper using only a couple of sentences and would thus certainly be possible even in journals with fairly strict word limits.

3.2. SENSITIVITY ANALYSIS FOR UNMEASURED CONFOUNDING FOR CONTROLLED DIRECT EFFECTS

Our discussion up until now was concerned with sensitivity analysis for unmeasured confounding for total effects. But similar issues arise for controlled direct effect and natural direct and indirect effects. As was noted in Chapter 2, even if the exposure is randomized or if we controlled for all exposure–outcome confounding, if there is an unmeasured confounder of the mediator–outcome relationship, then estimates of direct and indirect effects will be biased. This will be a common problem in many studies in which questions of mediation and of direct effects are of interest. In this section we will consider sensitivity analysis techniques for controlled direct effects. In the two sections that follow, we will turn to natural direct and indirect effects.

3.2.1. Sensitivity Analysis for Controlled Direct Effects for a Continuous Outcome

We will first consider results for an additive scale and then for a risk ratio scale (or odds ratio with rare outcome). Suppose controlling for (C, U) would suffice to control for exposure–outcome and mediator–outcome confounding [assumptions (A2.1) and (A2.2)] but that no data are available on U and that U confounds the mediator–outcome relationship. These relationships are depicted in Figure 3.2. Similar formulae also hold if U also affects A; see VanderWeele (2010a) for further discussion. If we have not adjusted for U, then our estimates controlling only for C will be biased. We will consider estimating the controlled direct effect, $CDE(m)$, with the mediator fixed to m conditional on the covariates $C = c$. Let $B_{add}^{CDE}(m|c)$ denote the difference between (i) the estimate of the controlled direct effect conditional on C (e.g., using the methods described in Chapter 2) and (ii) what would have been obtained had adjustment been made for U as well. As with total effects, we will be able to use a very simple formula for sensitivity analysis for controlled direct effects under some simplifying assumptions. Suppose then that [assumption (A3.1)] U is binary and [assumption (A3.2c)] the effect of U on Y on the additive scale, conditional on exposure, mediator, and covariates, (A, M, C), is the same for both exposure levels $A = a$ and $A = a^*$. Let γ_m be the effect of U on Y conditional on A, C, and $M = m$, that is,

$$\gamma_m = \mathbb{E}(Y|a, c, m, U = 1) - \mathbb{E}(Y|a, c, m, U = 0)$$

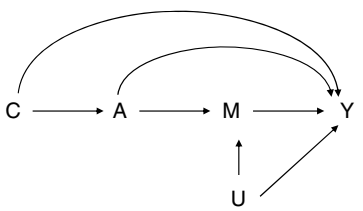

Figure 3.2 An unmeasured mediator–outcome confounder U that affects mediator M and outcome Y.

Note that by assumption (A3.2c), $\gamma_m = \mathbb{E}(Y|a,c,m,U=1) - \mathbb{E}(Y|a,c,m,U=0)$ is the same for both levels of the exposure. Let δ_m be the difference in the prevalence of the unmeasured confounder for those with $A=a$ versus those with $A=a^*$ conditional on $M=m$ and $C=c$, that is,

$$\delta_m = P(U=1|a,m,c) - P(U=1|a^*,m,c)$$

Under assumptions (A3.1) and (A3.2c), the bias factor is simply given by the product of these two sensitivity-analysis parameters (VanderWeele, 2010a):

$$B_{add}^{CDE}(m|c) = \delta_m \gamma_m. \tag{3.3}$$

Formula (3.3) states that under assumptions (A3.1) and (A3.2c) the bias factor $B_{add}^{CDE}(m|c)$ for the controlled direct effect $CDE(m)$ is simply given by the product $\delta_m \gamma_m$. Under these simplifying assumptions, this gives rise to a particularly simple sensitivity analysis technique for assessing the sensitivity of estimates of a controlled direct effect to an unmeasured mediator–outcome confounder. We can hypothesize a binary unmeasured mediator–outcome confounding variable U such that the difference in expected outcome Y comparing $U=1$ and $U=0$ is γ_m across strata of A conditional on $M=m, C=c$ and such that the difference in the prevalence of U, comparing exposure levels a and a^* (e.g., comparing the exposed and unexposed), is δ_m conditional on $M=m, C=c$. We consider the interpretation of this parameter in greater detail below. For such an unmeasured mediator–outcome confounding variable, the bias of our estimate of the controlled direct effect controlling just for C is given simply by $\delta_m \gamma_m$. We can assess sensitivity to the presence of such an unmeasured confounding variable by varying γ_m (which is essentially the direct effect of U on Y) and by varying δ_m, interpreted as the prevalence difference of U, comparing exposure levels a and a^* conditional on $M=m$ and $C=c$. We can subtract the bias factor $B_{add}^{CDE}(m|c) = \delta_m \gamma_m$ from the observed estimate to obtain a corrected estimate of the effect (i.e., what we would have obtained had it been possible to adjust for U as well). Under the simplifying assumptions (A3.1) and (A3.2c), we could also subtract this bias factor from both limits of a confidence interval to obtain a corrected confidence interval. Note that the controlled direct effect, $CDE(m)$, may vary with m, and so for different values of m we will likely want to consider different specifications of the values δ_m and γ_m in the sensitivity analysis. As discussed below, if there is no interaction between the effects of A and M on Y, then this simple sensitivity analysis technique based on using formula (3.3) will also be applicable to natural direct effects as well.

It is important to note that δ_m is the prevalence difference conditional on $M=m$ and $C=c$ and not the unconditional prevalence difference. To see the importance of this distinction, let us assume that A and M are binary and note that U is assumed to be a cause of M. In specifying the conditional prevalence difference $P(U=1|a,m,c) - P(U=1|a^*,m,c)$, the variable U might be (unconditionally) of equal or greater prevalence comparing exposure levels $A=1$ and $A=0$, but it might still be conditionally less prevalent comparing $A=1$ and $A=0$ say, given $M=1$; this is because $M=1$ might occur if either $U=1$ or $A=1$ and thus conditional on $M=1$; if $A=0$, then we would know that $U=1$ because $M=1$ only when

either $U = 1$ or $A = 1$. The prevalence of U, conditional on $M = m$, in exposure levels $A = 1$ and $A = 0$ will depend on the unconditional prevalence of U in exposure levels and also on the information that conditioning on $M = m$ gives about the prevalence of U in exposure levels $A = 1$ and $A = 0$. We will return to this point below in the examples and illustrate this issue of conditioning on M in the interpretation of the sensitivity analysis parameters. When U, A, M are all binary, some intuition can be given concerning on how such conditioning (i.e., conditioning on a common effect, M, of U and A) influences associations. In general, if the mechanism for M is an "and" mechanism (so that M occurs if A and U occur), then this will generally induce positive correlation between A and U, conditional on M; if the mechanism for M is an "or" mechanism (so that M occurs if A or U occurs), then this will generally induce negative correlation between A and U, conditional on M; however, exceptions can arise (VanderWeele and Robins, 2007a, 2009a).

3.2.2. Sensitivity Analysis for Controlled Direct Effects for a Binary Outcome

The sensitivity analysis technique above could also be employed with binary outcomes on a risk difference scale. However, often with binary outcomes, risk ratios or odds ratios are used as obtained from logistic regression analysis. Again we suppose that controlling for (C, U) would suffice to control for exposure–outcome and mediator–outcome confounding but U confounds the mediator–outcome relationship and that no data are available on U. If we have not adjusted for U, then our estimates controlling only for C will be biased. We will consider estimating the controlled direct effect odds ratio from Chapter 2, $OR^{CDE}(m)$, with the mediator fixed at level m, conditional on the covariates $C = c$. This approach will assume a rare outcome but can also be used for risk ratios with a common outcome. Let $B^{CDE}_{mult}(m|c)$ denote the *ratio* of (i) the estimate of the controlled direct effect conditional on C (e.g., using the methods described in Chapter 2) and (ii) what would have been obtained had adjustment been made for U as well. Suppose that [assumption (A3.1)] U is binary and that [assumption (A3.2d)] the effect of U on Y on the ratio scale, conditional on exposure, mediator, and covariates (A, M, C), is the same for both exposure levels $A = a$ and $A = a^*$. Let γ_m be the effect of U on Y conditional on A, C, and $M = m$, that is,

$$\gamma_m = \frac{P(Y = 1|a, c, m, U = 1)}{P(Y = 1|a, c, m, U = 0)}$$

Note that by assumption (A3.2d), γ_m is the same for both levels of the exposure of interest. Under assumptions (A3.1) and (A3.2d), the bias factor on the multiplicative scale is simply given by

$$B^{CDE}_{mult}(m|c) = \frac{1 + (\gamma_m - 1)P(U = 1|a, m, c)}{1 + (\gamma_m - 1)P(U = 1|a^*, m, c)} \tag{3.4}$$

Once we have calculated the bias term $B^{CDE}_{mult}(m|c)$, we can simply estimate our controlled direct effect risk ratio controlling only for C (e.g., if the outcome is rare,

we just fit a logistic regression) and we *divide* our estimate and confidence intervals by the bias factor $B^{CDE}_{mult}(m|c)$ to get the corrected estimate for controlled direct effect risk ratio and its confidence interval—that is, what we would have obtained if we had adjusted for U a well. Assumptions (A3.1) and (A3.2d) can be relaxed, but then the resulting formulae for the bias factor become more complicated to use and the interested reader is referred elsewhere (VanderWeele, 2010a). Note here we have to specify the two prevalences of U, namely $P(U = 1|a, m, c)$ and $P(U = 1|a^*, m, c)$, in the different exposure groups conditional on M and C. As with controlled direct effects on an additive scale, the issue of conditioning on M in the interpretation of these prevalences is important, as we illustrate in the examples that follow.

3.2.3. Some Examples of Sensitivity Analysis for Controlled Direct Effects

We illustrate these sensitivity analysis techniques for controlled direct effects using two examples from perinatal epidemiology. Empirical studies have found that when controlling for birth weight, M, for the group of infants with the lowest birth weight (less than 2000 grams say, denoted by $M = 1$) maternal smoking, A, is associated with a lower risk of infant mortality, Y, seemingly suggesting a protective effect for maternal smoking amongst infants weighing the least. This somewhat puzzling finding is commonly referred to as the "birth weight paradox" (Yerushalmy, 1971; Wilcox, 1993; Hernández-Díaz et al., 2006). Hernández-Díaz et al. (2006) point out that although analyses that document this association control for a number of maternal demographic factors, C, the analyses do not in general control for birth defects or malnutrition, U (i.e., other causes of low birth weight), which would serve as a confounder of the birth weight (mediator)–mortality (outcome) relationship. Estimates of the controlled direct effect are thus biased because control is not made for such mediator–outcome confounders. Essentially, infants might be low birth weight either because of smoking or because of, say, a birth defect or malnutrition. If an infant is not low birth weight because of smoking, it is more likely that the low birth weight is because of a birth defect or malnutrition or some other cause. Thus, if control is not made for these other causes of low birth weight and if comparison is made between the groups with and without maternal smoking, then it looks as if the effect of smoking is protective for the infants of lowest birth weight; this is simply because for this group of low-birth-weight infants, control is not made for other causes of low birth weight and thus no smoking and low birth weight together is likely indicative of the presence of a birth defect or malnutrition. Using the sensitivity analysis techniques, we can assess the degree of confounding required to completely explain away the birth weight paradox. Hernández-Díaz et al. (2006) use 1991 US linked birth/infant-death datasets from the National Center for Health Statistics and they define the lowest birth weight category ($M = 1$) as birth weight less than 2000 g. They control for maternal age, gravidity, education, marital status, race/ethnicity, and prenatal care (denoted by C). They find that if a naive estimate is used of the controlled direct effect risk ratio, one obtains

$\frac{P(Y=1|A=1,M=1,c)}{P(Y=1|A=0,M=1,c)} = 0.79$, suggesting a protective effect of maternal smoking on infant mortality for the low-birth-weight infants. The outcome is relatively rare in this case. Suppose, for sensitivity analysis, that we take U to be other causes of low birth weight, then, using the bias formulas for risk ratios for controlled direct effect in (3.4), we find that if $U = 1$ were to conditionally increase the risk of infant mortality three-and-a-half-fold so that $\frac{P(Y|a,M=1,c,U=1)}{P(Y|a,M=1,c,U=0)} = 3.5$ and if the prevalence of U for low-birth-weight infants whose mothers smoke is 0.025 but the prevalence of U for low-birth-weight infants whose mothers do not smoke is 0.14 (smoking is ruled out as an explanation of low birth weight, rendering other causes more likely), then this would indicate the bias produced by the unmeasured mediator–outcome confounder U, namely $B^{CDE}_{mult}(m|c) = \frac{1+(3.5-1)(0.03)}{1+(3.5-1)(0.14)} = 0.79$, which is sufficient to completely explain the apparent protective effect. Note that in specifying the prevalence of the unmeasured confounder for the exposed and the unexposed, we had to take into account that we were conditioning on the strata of $M = 1$ in which the infants were low birth weight and that low birth weight might be caused by either the maternal smoking exposure or by our unmeasured confounder U (e.g., birth defect or malnutrition). We thus specified a higher prevalence of U among those low-birth-weight infants whose mothers did not smoke (because there must then be some other cause of low birth weight) than among those whose mothers did smoke (because smoking was then likely the cause of low birth weight).

As a second example, Mann et al. (2011) used 122,476 mother–infant pairs in the South Carolina Medicaid program between 1996 and 2002 to study associations between pre-eclampsia (A) and cerebral palsy (Y) by preterm birth status (M with $M = 1$ denoting preterm birth). They control for a number of sociodemographic variables and obtain odds ratios stratified by preterm birth status of 0.73 (95% CI: 0.42, 1.26) for preterm infants and 3.46 (95% CI: 1.42, 8.41) for term infants. Pre-eclampsia appears to have a protective effect for preterm infants. This might once again be due to unmeasured mediator–outcome confounding. For example, Mann et al. (2011) do not control for intrauterine infection, which could affect both preterm birth and cerebral palsy. Let U then denote the presence of a common cause of preterm birth and cerebral palsy. The outcome is relatively rare in this case. Suppose for preterm infants the prevalence of U when pre-eclampsia is present is relatively low, say $P(U = 1|A = 1, M = 1, c) = 0.05$, because it is likely that pre-eclampsia was the cause of preterm birth. Suppose that for preterm infants the prevalence of U when pre-eclampsia is absent is much higher, say $P(U = 1|A = 0, M = 1, c) = 0.50$ (because if pre-eclampsia was not the cause of preterm birth, this renders other causes more likely present). If the effect of U on cerebral palsy were a two-fold increase, we would have $B^{CDE}_{mult}(m|c) = 0.70$; and by dividing the initial estimates by this bias factor, we obtain a corrected odds ratio of 1.04 (95% CI: 0.60, 1.80). If the effect of U were a four-fold increase, we would have $B^{CDE}_{mult}(m|c) = 0.46$ and a corrected odds ratio of 1.59 (95% CI: 0.91, 2.74). Even under the more moderate confounding scenario, there appears to be a detrimental effect of pre-eclampsia for preterm infants. Similar arguments suggest that the odds ratio 3.46 for term infants cannot be explained away even by substantial confounding (cf. VanderWeele and Hernandez-Diaz, 2011). There appears to

be detrimental direct effect of pre-eclampsia on cerebral palsy for both term and preterm infants.

In Chapter 2 we had discussed two limitations or problems with some of the more standard approaches to mediation analysis, such as the difference method. First, there was the possibility of having substantial interaction and ignoring it, and second there was the problem of unmeasured confounding. Both seem to be present here; it seems that there may likely be unmeasured confounding, by intrauterine infection, and there was also evidence for interaction. We had obtained initial odds ratio estimates of 0.73 (95% CI: $0.42, 1.26$) for preterm infants and 3.46 (95% CI: $1.42, 8.41$) for term infants. Here these two issues of unmeasured confounding and potential interaction would seem to combine in a rather bizarre way if we were to use the difference method. The naïve analysis suggests a negative direct effect for preterm infants and a positive direct effect for term infants; because these supposed effects are in different directions, ignoring the interaction might give us something not too far from a null controlled direct effect. The standard conclusion would then be that all of the effect is mediated (once the mediator is in the model, the effect of exposure is null). This would be the opposite of our more reasoned conclusion of a strong direct effect; the two problems of interaction and no unmeasured confounding essentially would combine here to give precisely the wrong conclusion! Sensitivity analysis, and allowing for interaction, helps us to be able to assess these problems.

3.3. SENSITIVITY ANALYSIS FOR UNMEASURED CONFOUNDING FOR NATURAL DIRECT AND INDIRECT EFFECTS

Thus far we have considered sensitivity analysis techniques for total direct effects and controlled direct effects. In the next two sections we will consider some sensitivity analysis techniques for natural direct and indirect effects.

3.3.1. Sensitivity Analysis for Natural Direct and Indirect Effects in the Absence of Exposure-Mediator Interaction

One simple setting in which we can very easily employ sensitivity analysis for natural direct and indirect effects is when natural direct effects and controlled direct effects coincide. This occurs when our four confounding assumptions from Chapter 2 are satisfied [assumptions (A2.1)–(A2.4)] and there is no exposure–mediator interaction in the statistical model. In this case, we can essentially just proceed with the sensitivity analysis techniques that we have already described for controlled direct effects. If we assume an unmeasured mediator–outcome confounder U as in Figure 3.2, we can use these same techniques and the same parameters to do sensitivity analysis for natural direct effects as well.

For natural indirect effects, we can proceed by making use of the decomposition property of the total effect. Suppose we have unmeasured mediator–outcome

confounding as in Figure 3.2. It can be shown (see the Appendix) that even if this mediator–outcome confounding is present and our estimates of the natural direct and indirect effects are biased, the natural direct and indirect effects themselves will still combine to the correct total effect. Intuitively, because a mediator–outcome confounder does not confound the exposure outcome relationship, we can still obtain valid estimates of the total effect. And, it turns out that the combination of the direct and indirect effects do constitute a consistent estimator of the total effect, even though the direct and indirect effects estimators will themselves be biased for the true natural direct and indirect effects.

Knowing that the direct and indirect effect estimates combine to a valid estimate of the total effect then allows us to also be able to essentially employ the sensitivity analysis techniques for controlled direct effects for natural indirect effects as well. To do so, we use the negation (on the additive scale) or the inverse (on the multiplicative ratio scale) of the bias formulas that we used for controlled direct effects (and natural direct effects). Thus on the additive scale, for a continuous outcome for example, our bias factor for the natural direct effect would simply be the negation of the formula in (3.3) (that is, it would be $-\delta_m \gamma_m$) and we could subtract this from the estimate and both limits of the confidence interval to obtain a corrected estimate and confidence interval for the natural indirect effect. For a binary outcome, on the odds ratio scale with rare outcome or risk ratio scale with common outcome, our bias factor for the natural indirect effect would be the inverse of that in (3.4); that is, it would be $\frac{1+(\gamma_m-1)P(U=1|a^*,m,c)}{1+(\gamma_m-1)P(U=1|a,m,c)}$ and we could divide our natural indirect effect estimates and its confidence interval by this bias factor to obtain a corrected estimate and confidence interval. When there is no exposure–mediator interaction, sensitivity analysis for direct and indirect effects could proceed in this relatively simple manner.

Example: Environmental Epidemiology

As an example of such a sensitivity analysis approach, we will consider analyses of Caffo et al. (2008), who studied the extent to which the effect of cumulative lead dose, A, on cognitive function, Y, is mediated by brain volumes, M. Prior research had indicated that occupational exposure to organic and inorganic lead was longitudinally associated both with cognitive decline (Schwartz et al., 2000) and with a decrease in the volume of brain structures as measured by MRI (Stewart et al., 2006). Caffo et al. (2008) hypothesized that the effect of lead exposure on cognitive function would be at least partially mediated by brain volumes; to examine this, they used data for 2001–2003 from a study of 513 former organolead manufacturing workers in their analyses. Brain volumes were measured using magnetic resonance imaging, which captures only difference in brain volume and not more subtle neurobiologic changes to brain structure. Caffo et al. control for a number of covariates, C, including age, education, smoking, and alcohol consumption; in one of their analyses they consider an executive cognitive functioning test score outcome with white matter volume as the mediator and use a linear regression model of Y on A, M, C without an $A \times M$ product term. Under the assumption that the regression model is correctly specified and that there is no unmeasured

exposure–outcome or mediator–outcome confounding, they obtain estimates of a direct effect of 3.79 point decline (95% CI: $-7.40, -0.18$) in executive functioning cognitive test scores per 1-μg/g increase in peak tibia lead exposure, controlling for white matter in brain regions associated with lead; the indirect effect of lead exposure as mediated through white matter brain volume was a 1.21 ($P = 0.01$) point decline in executive functioning cognitive test scores per 1-μg/g increase in peak tibia lead exposure; the total effect of lead exposure is thus a 5.00 point decline (95% CI = -8.57 to -1.42) in executive functioning cognitive test scores per 1-μg/g increase in peak tibia lead exposure.

Suppose now that there is an unmeasured confounding variable U affecting both white matter brain volume and cognitive function; U might denote an unknown genetic factor. We could employ the approach above to assess the extent to which U might change our conclusions about the direct and indirect effect. To do so, we are making the simplifying assumptions that U is only a mediator–outcome confounder (i.e., it does not affect cumulative lead dose), that U is binary, and also that U does not interact with the exposure, cumulative lead dose, in its effects on cognitive function. Suppose then that individuals with the genetic factor present had on average 7-point-lower executive functioning test scores in that $\gamma_m = \mathbb{E}[Y|a,m,c,U=1] - \mathbb{E}[Y|a,m,c,U=0] = -7$. Using the sensitivity analysis approach, we have that if, conditional on white matter brain volume, individuals with a 1-μg/g increase in peak tibia lead had a 0.54 higher probability of having the genotype present (i.e., $\delta_m = P(U=1|a+1,m,c) - P(U=1|a,m,c) = 0.54$), then the true direct effect might in fact be $-3.79 - (-7)(0.54) = 0$ (95% CI: $-3.61, 3.61$). Relatively large parameters are required to explain away the effect. Figure 3.3 in fact gives the values of $\gamma = \mathbb{E}[Y|a,m,c,U=1] - \mathbb{E}[Y|a,m,c,U=0]$ and $\delta = P(U=1|a+1,m,c) - P(U=1|a,m,c)$ that would be required to completely eliminate the direct effect; values of γ and δ that lie below the curve would reverse the sign of the direct effect point estimate. Using the approach above for natural indirect effects, we also have that if individuals with the genetic factor had on average 7-point-lower executive functioning test scores and if, conditional on white matter brain volume, individuals with a 1-μg/g increase in peak tibia lead had a 0.17 lower probability of having the genotype present, then the true indirect effect might in fact be $-1.21 + (-7)(-0.17) = 0$. This is arguably somewhat less implausible than the degree of confounding needed to explain away the direct effect. Overall, the presence of at least the direct effect and possibly the indirect effect in the analyses of Caffo et al. (2008) seem unlikely to be due simply to such mediator–outcome confounding.

3.3.2. Sensitivity Analysis for Natural Direct and Indirect Effects with a Binary Mediator and Binary Outcome

In this subsection we will describe a sensitivity analysis technique that can be used for natural direct and indirect effects even in the presence of interaction, provided that the mediator and outcome are both binary. Consider the case with binary outcome and binary mediator. As in the last chapter, Section 2.5, we could fit two

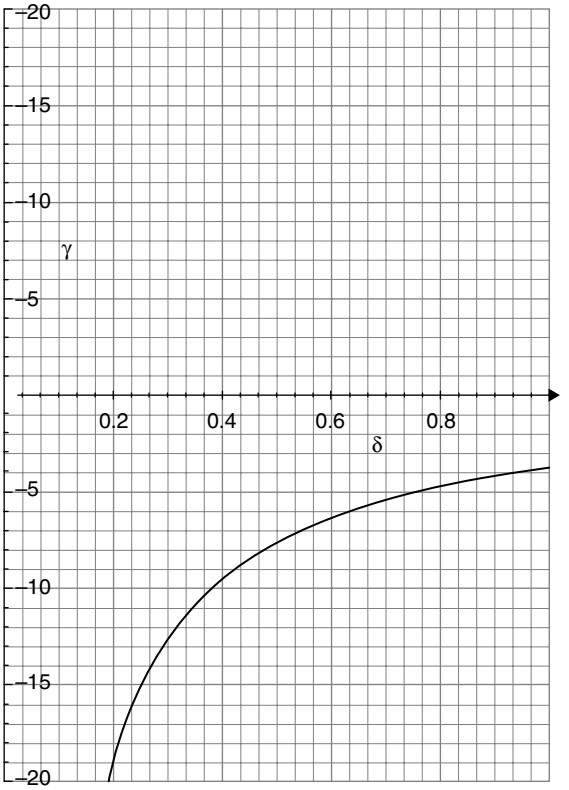

Figure 3.3 Sensitivity analysis plot: values of γ (the effect of U on the outcome) and δ $(= P(U = 1|a + 1, m, c) - P(U = 1|a, m, c))$ that lie below the curve would reverse the sign of the direct effect point estimate

logistic regression models, one for the mediator and one for the outcome:

$$\text{logit}\{P(Y = 1|a, m, c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M = 1|a, c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

If we thought our four assumptions about no-unmeasured confounding held—that is, no unmeasured exposure–outcome, mediator–outcome, and exposure–outcome confounding conditional on C and no mediator–outcome confounders affected by the exposure [assumptions (A2.1)–(A2.4)]—we could use the formulas given in Section 2.5 to obtain our estimates of the natural direct and indirect effects. Suppose now that there were an unmeasured binary U that confounded only the mediator–outcome relationship. And suppose we made the assumption that (A3.2d) the effect of U on Y on the ratio scale was the same for both exposure groups and both mediator levels and we specify this effect of U on Y as one of our sensitivity analysis parameters:

$$\gamma = \frac{P(Y = 1|a, m, c, U = 1)}{P(Y = 1|a, m, c, U = 0)}$$

Suppose as additional sensitivity analysis parameters we specified the prevalence of U in each of the four categories defined by our exposure and mediator—that is, $P(U=1|a,M=0,c)$, $P(U=1|a,M=1,c)$, $P(U=1|a^*,M=0,c)$, and $P(U=1|a^*,M=1,c)$. Using these prevalences and our other sensitivity analysis parameter γ, we could calculate the following three quantities:

$$B_0 = \frac{1+(\gamma-1)P(U=1|a,M=0,c)}{1+(\gamma-1)P(U=1|a^*,M=0,c)}$$
$$B_1 = \frac{1+(\gamma-1)P(U=1|a,M=1,c)}{1+(\gamma-1)P(U=1|a^*,M=1,c)}$$
$$B_2 = \frac{1+(\gamma-1)P(U=1|a^*,M=1,c)}{1+(\gamma-1)P(U=1|a^*,M=0,c)}$$

If we then let

$$\theta_1^\dagger = \theta_1 - \log(B_0)$$
$$\theta_2^\dagger = \theta_2 - \log(B_2)$$
$$\theta_3^\dagger = \theta_3 - \log(B_1) + \log(B_0)$$

and if we substitute $(\theta_1,\theta_2,\theta_3)$ with $(\theta_1^\dagger,\theta_2^\dagger,\theta_3^\dagger)$ in the formulas in Section 2.5, we would obtain corrected effects. Corrected standard errors and confidence intervals are no longer as easy to obtain as in the other sensitivity analysis techniques considered in this chapter. They can be obtained via the delta method (see the Appendix) or by bootstrapping.

EXAMPLE: PERINATAL EPIDEMIOLOGY

As an illustration of this technique, we consider an analysis of Ananth and VanderWeele (2011) examining the extent to which the effect of abruption, A, on early neonatal mortality (0–6 days), Y, is mediated by preterm birth, M. Placental abruption is defined as the premature separation of a normally implanted placenta from its attachment to the uterus wall prior to delivery of the fetus. Although the overall incidence of abruption is only about 1%, its effects on neonatal mortality are very substantial. Data come from the U.S. National Center for Health Statistics Period Linked Birth Certificate Infant Mortality files 1995–2002. Adjustment was made for a number of covariates, C, including age, education, gravidity, smoking, and race. The approach described above for a binary outcome and binary mediator is used but with log-linear regression for the outcome, rather than logistic regression (since the outcome is not rare when abruption is present). The estimates obtained were: $\beta_0 = -2.71$, $\beta_1 = 1.44$, $\theta_1 = 2.39$, $\theta_2 = 3.15$, and $\theta_3 = -1.42$. The estimates of the natural direct and indirect effect risk ratios were: $RR^{NDE} = 5.59$ (95% CI:5.19−6.02) and $RR^{NDE} = 1.61$ (95% CI: 1.55−1.67). Suppose now that there were an unmeasured binary mediator–outcome confounding variable. We then need to specify our sensitivity analysis parameters. Suppose that the prevalence of U amongst term pregnancies was $P(U=1|A=1,M=0,c) = P(U=$

$1|A=0,M=0,c)=5\%$, with and without abruption, and that the prevalence of U was $P(U=1|A=1,M=1,c)=10\%$ amongst preterm deliveries with abruption but $P(U=1|A=1,M=1,c)=50\%$ amongst preterm deliveries without abruption (the prevalence of U among preterm infants without abruption is specified higher than those with abruption because in the presence of placental abruption it is placental abruption itself that is likely the cause of preterm birth, whereas without abruption some other cause of preterm birth was likely present). Suppose U increased the likelihood of the mortality outcome by a factor of $\gamma = 1.5$. We could then use these sensitivity analysis parameters to calculate (B_0,B_1,B_2) and $(\theta_1^\dagger,\theta_2^\dagger,\theta_3^\dagger)$ using the formulas above and by replacing $(\theta_1,\theta_2,\theta_3)$ by $(\theta_1^\dagger,\theta_2^\dagger,\theta_3^\dagger)$ we obtain corrected parameters estimates of $(\theta_1^\dagger,\theta_2^\dagger,\theta_3^\dagger)=(2.39,2.95,-1.24)$ and corrected natural direct and indirect effect risk ratio estimates of: $RR^{NDE}=6.28$ $(95\%\ CI{:}5.82-6.76)$ and $RR^{NIE}=1.59$ $(95\%\ CI{:}\ 1.53-1.66)$. If U increased the likelihood of the mortality outcome by a factor of $\gamma = 6$, then we would obtain corrected parameters estimates of $(\theta_1^\dagger,\theta_2^\dagger,\theta_3^\dagger)=(2.39,2.12,-0.57)$ and corrected natural direct and indirect effect risk ratio estimates of: $RR^{NDE}=9.07$ $(95\%\ CI{:}8.41-9.77)$ and $RR^{NDE}=1.51$ $(95\%\ CI{:}\ 1.45-1.57)$. The estimates do change as the sensitivity analysis parameters vary, but the general conclusions that a substantial portion of the effect is direct but that some of the effect of abruption is mediated by preterm birth seems fairly robust. See Ananth and VanderWeele (2011) for a fuller analysis of these data and for a substantive investigation.

A downside of the approach presented here is that the sensitivity analysis parameters for the prevalence of the unmeasured confounder within strata of the exposure and the mediator, $P(U=1|a,m,c)$, have to be specified, and this can be difficult to do directly since the unmeasured confounder affects the mediator and not vice versa. An alternative parameterization was proposed by Hafeman (2011) for the setting in which the exposure, mediator, and outcome were all binary with a binary unmeasured confounder U. In addition to specifying the equivalent of the parameter γ above for the effect of U on Y, her approach specifies another parameter corresponding to the effect of the unmeasured confounder on the mediator—for example, $\gamma = \frac{P(M=1|a,c,U=1)}{P(M=1|a,c,U=0)}$. This may be easier to specify in practice than the prevalence of the unmeasured confounder within strata of the exposure and the mediator, $P(U=1|a,m,c)$. This approach can be implemented on the risk difference or risk ratio scales. The web appendices of Hafeman provide SAS code to carry out this approach. Her approach has fewer parameters but makes stronger assumptions than that presented above (e.g., not only that the effect of U on Y is constant across strata of A, but also that the effect of U on M is constant across strata of A). However, the implementation of Hafeman (2011) currently only applies to settings in which the exposure, mediator, and outcome are all binary and there are no measured covariates, only a single binary unmeasured covariate U. Thus her approach may perhaps be most applicable in a randomized trial, so that measured covariates are not needed to control for exposure–outcome confounding, in which both the mediator and outcome are binary. However, her parameterization could perhaps be adapted to more general settings in future research.

3.3.3. A Correlated Errors Approach to Sensitivity Analysis for Natural Direct and Indirect Effects

The approaches we have been considering have hypothesized an unmeasured confounder U that affects both the mediator and the outcome. We have considered specifying parameters corresponding to the effect of U on M and the effect of U and Y and varying these in sensitivity analysis. An alternative way to conceptualize unmeasured confounding is to suppose that the errors terms in the regression models for M and Y are correlated and then to specify the correlation of these error terms as a sensitivity analysis parameter. Imai et al. (2010a,b) develop such an approach. Suppose then, for example, that the variables follow the regression models

$$Y = \phi_0 + \phi_1 A + \phi_4' C + \epsilon_1$$
$$M = \beta_0 + \beta_1 A + \beta_2' C + \epsilon_2$$
$$Y = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4' C + \epsilon_3$$

but that the errors terms ϵ_2 and ϵ_3 are correlated. Imai et al. (2010a) show that if one specifies the correlation between the error terms in the second two regressions for M and Y as ρ, then the natural indirect effect for example is given by

$$NIE = \frac{\beta_1 \sigma_1}{\sigma_2}(\tilde{\rho} - \rho\sqrt{\{1-\tilde{\rho}^2\}/(1-\rho^2)})$$

where $\sigma_1 = Var(\epsilon_1|A=1)$, $\sigma_2 = Var(\epsilon_2|A=1)$, $\tilde{\rho} = Cov(\epsilon_1, \epsilon_2|A=1)$, and β_1 can all be estimated from the data. The parameter ρ could then be varied in a sensitivity analysis to examine how strongly the errors terms, ϵ_2 and ϵ_3, in the regressions for M and for Y, would have to be correlated to eliminate the evidence of a mediated effect. Imai et al. (2010a) also provide a software implement for this approach for both continuous outcomes and binary outcomes (using a probit model), and this software will also give corrected confidence intervals.

3.4. SENSITIVITY ANALYSIS USING TWO TRIALS

In this section we will consider an alternative sensitivity analysis approach for natural direct and indirect effects that uses data from two different trials, one of which randomizes the exposure and the other of which randomizes the mediator. Under some assumptions about the absence of interaction described below, if the two trials are run in the same population, then it will be possible to identify natural direct and indirect effects. However, a far more common setting might involve (a) having a primary trial in which just the exposure is randomized and (b) access to the results from other trials in other populations in which the mediator was randomized. In this case one might have a range of plausible values for the effect of the mediator on the outcome from previously published results in different populations. This range could then be used in sensitivity analysis to assess a range of plausible values for the

natural direct and indirect effects in the primary study with the exposure randomized. The approach here will assume that there are unmeasured mediator–outcome confounders and thus that the approach to estimating natural direct and indirect effects presented in Chapter 2 would be biased. The approach here will get around this problem by using the results of a different trial, with either the same or a different population, which randomizes the mediator. We will first consider the setting in which both trials are conducted in the same population and then explain how trials in different populations might be used in a sensitivity analysis.

Specifically, let us assume that exposure A is randomized and that our mediator is binary. We can then assess the effect of the exposure A on the outcome Y simply by taking the difference in sample averages of the outcome in the two exposure groups: $\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]$. We can likewise assess the effect of the exposure A on the mediator M simply by taking the difference in sample averages of the mediator in the two exposure groups: $\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]$. Suppose now that assumptions (A2.1)–(A2.4) hold for some possibly unmeasured and unknown set of covariates W. Since the exposure is randomized, these would essentially have to be the mediator–outcome confounders. Because W does not have to be measured or known, we can simply define W to be the set of all mediator–outcome confounders and the only substantive assumption here that we are effectively making is that [assumption (A2.4)] none of the mediator–outcome confounders are affected by the exposure. We discuss violations of this assumption in Chapter 5. Suppose now that in a second trial in the same population we randomize the mediator. We can then assess the effect of the mediator on the outcome by taking the difference in sample average in the outcome across mediator groups: $\mathbb{E}[Y|M=1] - \mathbb{E}[Y|M=0]$. Let's call this effect $\gamma = \mathbb{E}[Y|M=1] - \mathbb{E}[Y|M=0]$. It can be shown (Emsley and VanderWeele, 2014; cf. Appendix) that provided that the mediator does not interact with the exposure or covariates W in its effect on the outcome on the additive scale, then the natural indirect effect is given by

$$NIE = \gamma \{\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]\}$$

and the natural direct effect is given by

$$NDE = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] - \gamma \{\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]\}$$

We can obtain both the natural direct effect and the natural indirect effect using the data from our two trials. To do this, the assumptions we had to make were first that [assumption (A2.4)] there were no mediator–outcome confounders affected by the exposure and second that the mediator does not interact with the exposure or covariates W in its effect on the outcome on the additive scale. There can, however, be interactions between the exposure and the covariates W in their effects on the outcome or on the mediator, and the approach above still applies. Note that we do not have to collect data on the mediator–outcome confounders (or even necessarily know what they are) to compute the natural direct and indirect effects. We get around the need for collecting data on the mediator–outcome confounders by randomizing the mediator in a second trial.

Now consider a perhaps more common setting in which we only have data on a primary trial in which just the exposure is randomized. While we may not be able to know what the effect of the mediator on the outcome is in the trial population, there may be estimates of the effect of the mediator on the outcome in other trials in other populations in which the mediator was randomized. We could potentially use the results from these other trials to conduct a sensitivity analysis for the natural direct and indirect effects in the trial of interest. In particular, now instead of letting γ denote the effect of the mediator on the outcome in the primary population, we will let γ denote a sensitivity analysis parameter for this effect. We might inform our range of plausible sensitivity analysis parameter values for γ, the effect of the mediator on the outcome, by using results from published trials randomizing just the mediator M. Note that we would not need to have access to the data in these published trials; we would only need the estimates (or confidence intervals) in these trials for the effect of the randomized mediator on the outcome. Then, once again, provided that we assume [assumption (A2.4)] that there are no mediator–outcome confounders affected by the exposure and that the mediator does not interact with the exposure or with (possibly unmeasured or unknown) mediator–outcome confounders W in its effect on the outcome on the additive scale, then we could again use the formulas above for the natural direct and indirect effect. In this case, however, we could compute a range of values for the natural direct and indirect effects by varying γ. We would compute $\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]$ and $\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]$ using the primary trial data, and we would vary γ according to the range thought plausible in the sensitivity analysis informed by trials that had randomized the mediator.

Example: Randomized Trial of Cognitive Behavioral Therapy

In the SMaRT trial (Strong et al., 2008), a randomized cognitive behavioral therapy intervention was found to have a beneficial effect on depression symptoms after 3 months follow-up using the SCL-20 depression scale (a scale from 0 to 4). However, it was also noted that the intervention also had an effect on the use of antidepressants. Those who were randomized to the cognitive behavioral therapy arm were more likely to use antidepressants during follow-up. This led to questions with regard to whether the cognitive behavioral therapy intervention had a beneficial effect on depressive symptoms simply because it led to higher antidepressant use, or whether the intervention affected depressive symptoms through other pathways—for example, by changing the thought and behavioral patterns of the participants. If the intervention were only beneficial because of higher use of antidepressants, then the cognitive-behavioral aspects of the intervention could perhaps be abandoned without much loss and a more cost-effective intervention just focusing on antidepressant adherence could be developed. Alternatively, it may be the case that the intervention was effective both because of increased antidepressant use and because of cognitive-behavioral changes.

In the trial, the effect of the intervention on depressive symptoms (scale of 0–4) at 3 months was $\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] = -0.342$ (95% CI: -0.55, -0.13). The effect of the intervention on antidepressant use was $\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0] = 0.27$; those in the intervention arm were more likely to use an

antidepressant. These are intent-to-treat effects and by randomization not subject to confounding. We would like to estimate the direct effect of the intervention on depressive symptoms, not through antidepressant use. A naive way to assess this direct effect, which ignores mediator–outcome confounding, would be to simply put the mediator in the regression model. If we do this, using the methods described in Chapter 2 and ignoring mediator–outcome confounding, we obtain an estimate of the controlled or natural direct effect of -0.368 and an estimate of the natural indirect effect of 0.026. Using this naive approach, ignoring mediator–outcome confounding, we get somewhat counterintuitive results. It looks as though the direct effect is slightly larger than the total effect, and it looks like the mediated effect makes depression worse by increasing antidepressant use! In fact, the coefficient for antidepressant use in the regression of depressive symptoms here is positive—it looks like antidepressant use itself increases depression! This is almost certainly due to unmeasured mediator–outcome confounding. Antidepressant use here is not randomized. It is thus likely that those in more difficult or negative situations are those who both use antidepressants and who are more depressed. The actual effect of antidepressant use on depressive symptoms is likely to decrease depressive symptom scores, but we do not see this because antidepressant use is not randomized.

To get around these problems, we can use the sensitivity analysis approach described above. We have estimates of the effect of the intervention on depressive symptoms and on antidepressant use. Now we simply need reliable estimates of the effect of antidepressant use on depressive symptoms scores. Two studies, Morrow et al. (2003) and Kroenke et al. (2001), ran randomized trials of antidepressant use and measured depressive symptoms using the SCL-20 depression scale. The two trials were conducted in different populations: Morrow et al. (2003) examined the effect of antidepressants for cancer patients and Kroenke et al. (2001) examined the effect of antidepressants for primary care patients. After standardizing effect measures, Morrow et al. (2003) obtained an effect estimate of -0.219; Kroenke et al. (2001) obtained an effect estimate of -0.698. We can use these two estimates to inform the range of our sensitivity analysis parameter γ for the effect of an antidepressant on depressive symptom scores. In particular, if we use a value of $\gamma = -0.698$ from Kroenke et al. (2001) as a "worst-case scenario" (for the direct effect), then for the indirect effect and direct effects we obtain estimates of

$$\begin{aligned} NIE &= \gamma \{\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]\} \\ &= -0.698\{0.27\} \\ &= -0.188 \ (95\% \text{ CI: } -0.279, -0.096) \end{aligned}$$

and

$$\begin{aligned} NDE &= \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] - \gamma \{\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]\} \\ &= -0.342 - (-.188) \\ &= -0.155 \ (95\% \text{ CI: } -0.383, 0.073) \end{aligned}$$

If we use a value of $\gamma = -0.219$ from Morrow et al. (2003) as a best-case scenario, then for the direct effect and indirect effects we obtain estimates of

$$\begin{aligned} NIE &= \gamma \{\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]\} \\ &= -0.219\{0.27\} \\ &= -0.059 \text{ (95\% CI: } -0.088, -0.030) \end{aligned}$$

and

$$\begin{aligned} NDE &= \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] - \gamma \{\mathbb{E}[M|A=1] - \mathbb{E}[M|A=0]\} \\ &= -0.342 - (-.059) \\ &= -0.283 \text{ (95\% CI: } -0.504, -0.063) \end{aligned}$$

In both cases the estimates indicate evidence that there is a direct effect of the therapy on depressive symptoms not through antidepressant use, though in the former "worst-case" scenario the 95% confidence interval for the direct effect does include 0.

Here we have considered a setting in which we randomized the exposure and either were able to randomize the mediator in a second trial or had access to the results of other studies that had randomized the mediator. Imai et al. (2013) consider other more complex designs involving randomization of the exposure and the mediator which can allow for inference for natural direct and indirect effects under weaker assumptions. In particular, one design that they consider is what they call a "parallel design." In this design, the sample is randomly split into two. In the first half-sample, just the exposure is randomized; in the second half-sample, both the exposure and the mediator are randomized. Imai et al. (2013) note that under the assumption that there is no interaction between the exposure and the mediator at the individual counterfactual level on the additive scale, the natural direct effect is given by $NDE = \mathbb{E}[Y|A=1,M=1] - \mathbb{E}[Y|A=0,M=1]$, where $\mathbb{E}[Y|A=1,M=1] - \mathbb{E}[Y|A=0,M=1]$ is calculated in the second half-sample with both the exposure and the mediator randomized and the natural indirect effect is given by *NIE*, the difference between (i) $\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]$, computed in the first half-sample with just the exposure randomized, and (ii) $\mathbb{E}[Y|A=1,M=1] - \mathbb{E}[Y|A=0,M=1]$, computed in the second half-sample with both the exposure and the mediator randomized. For both the natural direct effect and the natural indirect effect, the expression $\mathbb{E}[Y|A=1,M=1] - \mathbb{E}[Y|A=0,M=1]$ computed in the second half-sample could be replaced by $\mathbb{E}[Y|A=1,M=0] - \mathbb{E}[Y|A=0,M=0]$, likewise computed in the second half-sample. This is because under the assumption of no interaction between the exposure and the mediator in their effects on the outcome on the additive scale, these two quantities will be equal. We can use these expressions because (as noted earlier and in the Appendix, cf. Robins, 2003) if there is no interaction between the exposure and the mediator at the individual counterfactual level on the additive scale, then the natural direct effect is equal to the controlled direct effect; and if we have randomized the exposure and the

mediator, we can estimate the controlled direct effect. We can then use the difference between the total effect (computed in the first half-sample) and the natural direct effect (computed in the second half sample) as our estimate of the natural indirect effect. Note that here we did not have to make assumptions about the absence of mediator–outcome confounders affected by the exposure, but we did have to make an assumption, at the individual level, about no interaction between the exposure and the mediator. Imai et al. (2013) also discuss bounds for the natural direct and indirect effects in this parallel design when the no interaction assumption is not made. Sometimes the bounds suffice to identify at least the direction of the natural direct and indirect effect without making any assumptions at all beyond randomization. Note, however, that here, in contrast to the approaches described earlier in this section, we have to randomize the exposure and the mediator in the same study.

Imai et al. (2013) also discuss inference for natural direct and indirect effects in other types of designs as well which involve multiple randomized interventions. They discuss a crossover design in which it is possible to randomize the exposure and observe the value of the mediator and the outcome and then, for that same individual, re-randomize the exposure and randomize the mediator and observe the outcome once again. To use such a design, one must assume that the outcome can be observed multiple times (e.g., the outcome could not be death) and also that when we assign the exposure the first time and observe the mediator and outcome, there is no carry-over effect on the outcome when it is observed for that individual the second time after randomizing the exposure and the mediator. These are strong assumptions. Imai et al. (2013) also consider designs for natural direct and indirect effects, analogous to the parallel design and the crossover design, in which it is not possible to randomize the mediator directly but only to implement an intervention that will change the probability of (e.g., encourage) the mediator occurring; they derive bounds for natural direct and indirect effects under such "encouragement" designs. See Imai et al. (2013) for further details and see also Mattei and Mealli (2011) for discussion of similar designs used to estimate principal stratum direct effects, a topic we will discuss in Chapter 8.

3.5. SENSITIVITY ANALYSIS FOR DIRECT AND INDIRECT EFFECTS IN THE PRESENCE OF MEASUREMENT ERROR

3.5.1. Measurement Error Correction Techniques with a Continuous Mediator Subject to Nondifferential Measurement Error

Bias for direct and indirect effect estimates can arise not only due to unmeasured confounding but perhaps also due to measurement error. If the mediator is measured with error, this could potentially affect the regression coefficient estimates from both the mediator and the outcome regressions, thereby biasing estimates of direct and indirect effects. In this section we will describe some simple approaches that can be used to correct for such measurement error when the mediator is

continuous and the outcome is either binary or continuous and when there is no exposure–mediator interaction. Correcting for measurement error in these settings is relatively simple. However, correcting for measurement error in other contexts can be considerably more complicated. We will briefly summarize the existing literature on these more advanced correction methods.

We will suppose that a mediator is subject to nondifferential measurement error or misclassification (that is to say, the error does not depend on the exposure or outcome conditional on the true mediator and covariates). Intuitively we might expect that such measurement error will weaken the association between the mediator and the outcome and will therefore perhaps bias estimates of mediated effects toward the null and bias estimates of direct effects away from the null. An important question is under what conditions this intuition holds.

Suppose then that the mediator is continuous and the outcome is either binary or continuous and that there is no exposure-mediator interaction in the statistical models and suppose we employ the methods of the Chapter 2 to estimate direct and indirect effects. If there is no exposure mediator interaction then we would ordinarily use $NDE = \theta_1(a - a^*)$ and $NIE = \beta_1\theta_2(a - a^*)$ for the natural direct and indirect effects for a continuous outcome and $OR^{NDE} = \exp\{\theta_1(a - a^*)\}$ and $OR^{NIE} = \exp\{\beta_1\theta_2(a - a^*)\}$ for the natural direct and indirect effects on an odds ratio scale for a dichotomous outcome.

Suppose we have a mismeasured mediator, $\widetilde{M}$, where $\widetilde{M} = M + \varepsilon$ and ε is normally distributed with mean 0 and independent of M. Let λ denote the proportion of the variance in $\widetilde{M}$ explained by M, conditional on A and C. In obtaining direct and indirect effects estimates, we have two regressions to consider: our mediator regression and our outcome regression. Suppose both were fit using the mismeasured $\widetilde{M}$ rather than the true mediator M. With a continuous mediator subject to measurement error of the form above, our linear regression model for the mediator using $\widetilde{M}$ instead of M will still give valid estimates of the mediator regression coefficients (Carroll et al., 2006). Essentially, all that using $\widetilde{M}$ instead of M will do is introduce additional random variability in the dependent variable for this regression, but we can still obtain valid estimates.

The regression estimates for the model for Y do, however, need to be corrected to take into account the measurement error. Suppose then that we fit either the model

$$\mathbb{E}(Y|A = a, \widetilde{M} = \widetilde{m}, C = c) = \widetilde{\theta}_0 + \widetilde{\theta}_1 a + \widetilde{\theta}_2 \widetilde{m} + \widetilde{\theta}_4' c$$

if the outcome were continuous or

$$\text{logit}[P(Y = 1|A = a, \widetilde{M} = \widetilde{m}, C = c)] = \widetilde{\theta}_0 + \widetilde{\theta}_1 a + \widetilde{\theta}_2 m + \widetilde{\theta}_4' c$$

if the outcome were binary—that is, we use the analogues of models (2.2) and (2.4) in Chapter 2 but using the mismeasured mediator rather than the true mediator—and without interactions. Under the assumptions made above about measurement error, the relationships between the coefficients from the regression with the mismeasured mediator $\widetilde{M}$ and the true mediator M are given by (Carroll

et al., 2006)

$$\theta_1 = \widetilde{\theta}_1 - \widetilde{\theta}_2(\frac{1}{\lambda} - 1)\beta_1 \qquad (3.5)$$
$$\theta_2 = \frac{\widetilde{\theta}_2}{\lambda}$$

We can then use these relations to obtain corrected estimates of direct and indirect effects (le Cessie et al., 2012; VanderWeele et al., 2012e). We could use the outcome model with the mismeasured mediator to estimate $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$. Once we specify λ (the proportion of the variance of $\widetilde{M}$ explained by M, conditional on A and C), then we could use λ, $\widetilde{\theta}_1$, and, $\widetilde{\theta}_2$, and the equations in (3.5) to obtain corrected estimates of θ_1, θ_2. We could then use the formulas for controlled direct effects and natural direct and indirect effects from Chapter 2 but using the corrected coefficients instead of the ones from the mismeasured regression. Corrected confidence intervals have to take into account that β_1 is estimated (since β_1 appears in the correction formula for θ_1; this can be done with the delta method), or they could be obtained by bootstrapping.

3.5.2. Predicting the Direction of the Bias

The results also have interesting implications for the direction of bias of these effects. If we ignored measurement error and used $NIE = \beta_1\widetilde{\theta}_2(a - a^*)$ or $OR^{NIE} = \exp\{\beta_1\widetilde{\theta}_2(a - a^*)\}$ as the estimate of the natural indirect effect on the difference or odds ratio scale respectively (i.e., if we ignored the measurement error), then our estimate of the natural indirect effect would be biased toward the null since $\widetilde{\theta}_2 = \lambda\theta_2$ and $\lambda < 1$. Similarly, it follows from equation (3.5) that if the direct and indirect effects are in the same direction, and if the exposure has a positive effect on the mediator ($\beta_1 > 0$) and if we ignored measurement error and used or $NDE = \widetilde{\theta}_1(a - a^*)$ or $OR^{NDE} = \exp\{\widetilde{\theta}_1(a - a^*)\}$ as a measure of the controlled direct effect or natural direct effect, then this will be biased away from the null (i.e., if the true direct effect is positive, then the biased estimate will be even larger; if the true direct effect is negative, then the biased estimate will be even more negative). Thus under classical non-differential measurement error with a normally distributed mediator, under the regression models with no exposure–mediator interaction, the bias of the mediated effect is always toward the null; and if the effect of the exposure on the mediator is positive and the direct and indirect effects are in the same direction, then the bias of direct effect is away from the null.

We might then wonder whether nondifferential measurement error of the mediator will always result in a similar pattern of biases with the natural indirect effect biased toward the null (and under certain conditions at least the direct effect biased away from the null). This pattern is in some sense intuitive insofar as if we have measurement error in the mediator we might expect this to weaken the association between the observed mediator and outcome, which might then in general lead to

underestimating the indirect effect. It can be shown (Ogburn and VanderWeele, 2012) that if a binary mediator is subject to nondifferential misclassification, then once again the bias of the mediated effect is toward the null and the bias of direct effect is away from the null. Unfortunately, however, nondifferential misclassification of a polytomous mediator (e.g., a mediator with more than two levels) will not always result in biases that follow these patterns. It is possible to construct examples of a nondifferentially misclassified mediator with three levels such that the bias of the mediated effect is away from the null and the bias of the direct effect is toward the null. It is even possible to construct examples in which the direct and mediated effect estimates, when ignoring measurement error, lie on the wrong side of the null. The intuition on the consequences of measurement error for direct and indirect effects will often hold but not always.

One final point is of interest before moving on. Using the definitions of natural direct and indirect effects given in the causal inference literature, the total effect will always decompose into the sum of the natural direct and indirect effects on a difference scale, and a total effect on the odds ratio scale will always decompose into the product of the natural direct and indirect effects odds ratios. We saw above that in the presence of measurement error for the mediator, the standard estimators for the direct and indirect effects will be biased. However, even if we use these biased direct and indirect measures and take their product on the odds ratio scale (or sum on the difference scale), we will still get an unbiased estimate of the total effect. As with the analogous result for confounding, in some ways this is intuitive. Even if we have measurement error of the mediator, we should still be able to obtain valid estimates of total effects by simply ignoring the mediator. What may be surprising is that even if we use the mismeasured mediator to estimate biased direct and indirect effects, their combination is still unbiased for the total effect. In fact, this property holds not simply for nondifferential measurement error of the mediator, but also for any form of measurement error of the mediator (VanderWeele et al., 2012e).

3.5.3. Other Measurement Error Correction Techniques

Our discussion thus far has focused on nondifferential measurement error of the mediator in the absence of exposure–mediator interaction. Other forms of measurement error may be more difficult to correct for. In a recent paper, le Cessie et al. (2012) discuss using measurement error methods (Carroll et al., 2006) to correct estimates of controlled direct effects for different forms of measurement error in the mediator including differential measurement error with the exposure or outcome affecting the mediator measurement, differential or nondifferential intra-individual variation over time, and various trigger mechanisms. They restrict their attention to the controlled direct effect in settings without exposure–mediator interaction. In the absence of such exposure–mediator interaction, we could also potentially use these approaches to reason about natural direct and indirect effects. We could do this by using the approaches in le Cessie et al. (2012) to get corrected estimates of the controlled direct effect. If the four no-confounding assumptions in Chapter 2

hold and there is no exposure–mediator interaction, then this will also give a corrected estimate of the natural direct effect. We could then estimate the total effect by simply ignoring data on the mediator. Measurement error of the mediator will then not affect estimates of the total effect. Finally we could take the difference between the total effect and the corrected natural direct effect (or on the odds ratio scale, we could take the ratio of our total effect odds ratio and the measurement-error-corrected direct effect odds ratio) to obtain a measurement-error-corrected natural indirect effect. Standard errors could be obtained by bootstrapping. This approach could be employed whenever the mediator measurement error is such that we are able to obtain measurement-error-corrected estimates of the direct effect.

Of course all of our discussion here has presupposed that the models were correctly specified. One advantage of the approach to mediation that has developed within the causal inference literature is that it has allowed for the definition and estimation of direct and indirect effects even in the presence of exposure–mediator interactions. However, correcting for measurement error in such cases is more difficult; some results are available for both binary and continuous mediators and for binary, continuous, and count outcomes (Valeri et al., 2014a; Valeri and VanderWeele, 2014), but easy-to-use software remains to be developed. In settings with a binary mediator subject to non-differential misclassification, correction for measurement error is also more complicated because the measurement error then affects the coefficients for both the outcome regression for Y and the mediator regression for M, and corrections need to be made to both regression models. Some techniques have been proposed (Valeri and VanderWeele, 2014), but again no simple implementation is yet available. Here at least, with a binary mediator, as noted above, we can predict the direction of the bias of direct and indirect effects. Other work in the social sciences addresses mediator measurement error by utilizing data on multiple measurements or on variables related to the mediator. See Bollen (1989) and MacKinnon (2008) for further details. Tchetgen Tchetgen and Lin (2013) have also recently proposed a three-stage least squares approach to addressing measurement error for a continuous mediator subject to nondifferential classical measurement error if some of the covariates C affect exposure A and mediator M but do not affect the outcome Y. More research is needed on approaches to handle measurement error in mediation analysis.

Recent work has also investigated the biases of natural direct and indirect effect estimators in the presence of nondifferential measurement error of the exposure or the outcome (Jiang and VanderWeele, 2015a,b). For nondifferential measurement error of the outcome, both direct and indirect effects are unbiased for continuous outcomes, and both are biased toward the null for dichotomous outcomes (Jiang and VanderWeele, 2015a). For nondifferential measurement error of the exposure, in the absences of exposure–mediator interaction, the natural direct effect is biased toward the null, but the natural indirect effect can be biased in either direction (Jiang and VanderWeele, 2015b). Jiang and VanderWeele (2015a,b) also developed correction methods for direct and indirect effects estimators in the presence of nondifferential measurement error of the exposure and outcome.

3.6. DISCUSSION

As was clear in Chapter 2, the assumptions required to identify direct and indirect effects from observational data are quite strong. In many settings, these assumptions will be violated. It is therefore important to assess the extent to which conclusions being drawn about direct and indirect effects are robust to violations of these assumptions. The sensitivity analysis tools in this chapter provide several different approaches to assess, at least to a certain extent and in some settings, how strong violations in assumptions would need to be to invalidate results. We can use these sensitivity analysis techniques to assess the extent to which unmeasured confounding might explain away an estimate; however, we can also use a variety of different specifications of the sensitivity analysis parameters to attempt to come up with a range of plausible values. The sensitivity analysis techniques themselves make assumptions but still allow an investigator to get a handle on the degree of confounding or measurement error that would be necessary to explain away an effect estimate. An alternative approach to sensitivity analysis, which often makes no or minimal assumptions, is to find bounds for total effects or direct effects (Manski, 1990, 1997, 2003; Cai et al., 2008; Kaufman et al., 2005, 2009; Sjölander, 2009; Imai et al., 2010c; Robins and Richardson, 2010; VanderWeele, 2011a). Such bounds generally do not make assumptions on the magnitude of unmeasured confounding at all. Unfortunately, these bounds are often not very informative and almost always include the null hypothesis of no effect. They essentially consider very extreme scenarios and thus, in most cases, may be of limited use. Sensitivity analysis techniques consider less extreme scenarios and allow the investigator to specify the range of sensitivity analysis parameters that are thought to be plausible. Because estimation of direct and indirect effects make strong assumptions, they are best accompanied by sensitivity analysis. The chapter here has presented a number of different techniques, but much further methodological research needs to be done to expand the range of settings that these techniques can handle and to develop simple software implementations. We will also return to this topic of sensitivity analysis in the next two chapters, where we consider survival outcomes (in Chapter 4) and settings with mediator–outcome confounders that are affected by the exposure (in Chapter 5).

4

Mediation Analysis with Survival Data

In some social and biomedical research the outcome of interest is neither binary nor continuous but rather the time to an event. In Chapters 2 and 3 we considered methods for binary and continuous outcomes. In this chapter we will consider analogous methods for time-to-event outcomes. This chapter will presume familiarity with survival analysis models such as the proportional hazards model and the accelerated failure time models. Readers without such familiarity can skip this chapter and move directly on to Chapter 5. Nothing in subsequent chapters will require knowledge of the contents of this chapter. The present chapter also describes the use of the additive hazards models for mediation, a weighting approach to mediation with time-to-event outcomes, and sensitivity analysis techniques for mediation with time-to-event outcomes.

4.1. EARLIER LITERATURE ON MEDIATION ANALYSIS WITH SURVIVAL MODELS

As in previous chapters, we will let A denote an exposure of interest, M a mediator, and C a set of covariates. Now, however, we will use T to denote a time-to-event outcome. For a time-to-event outcome T we will let $S_T(t)$ denote the survival function at time t—that is, $S_T(t) = P(T > t)$; the survival function conditional on covariates $C = c$ can likewise be defined as $S_T(t|c) = P(T > t|c)$. We will use $\lambda_T(t)$ and $\lambda_T(t|c)$ for the hazard or conditional hazard at time t, that is the instantaneous rate of the event conditional on $T \geq t$. We will assume, as before, that our four assumptions in Chapter 2 about confounding hold—that is, that the covariates C suffice to control for [assumption (A2.1)] exposure–outcome, [assumption (A2.2)] mediator–outcome, and [assumption (A2.3)] exposure–mediator confounding and that [assumption (A2.4)] none of the mediator–outcome confounders are themselves affected by the exposure. We will also assume throughout that if there is censoring, then it is uninformative.

The survival analysis models most frequently employed in the epidemiologic and social science literatures are probably, first, the proportional hazards model and, second, accelerated failure time models. The possibility of conducting mediation analysis with survival data under both models was in fact considered in a paper by Tein and MacKinnon (2003) in the social science literature some years ago. As noted in Chapter 2, there have traditionally been two methods for undertaking mediation analysis. The "difference method," which is more common in epidemiology, considers an outcome model both with and without the mediator and takes the difference in the coefficients for the exposure in the models without and with the mediator as the measure of the indirect or mediated effect. The "product method," more common in the social sciences, takes as a measure of the indirect effect the product of the coefficient for the exposure in the model for the mediator and the coefficient for the mediator in the model for the outcome. If the outcome and mediator are continuous and there are no interactions in the model for the outcome, then the two methods coincide (MacKinnon et al., 1995). However, with binary outcomes the two methods may diverge (MacKinnon and Dwyer, 1993; MacKinnon, 2008); they will approximately coincide when the binary outcome is rare (VanderWeele and Vansteelandt, 2010).

Tein and MacKinnon (2003) consider whether the two approaches coincide with proportional hazards and accelerated failure time models. They effectively use a linear regression model for the mediator as was done in Chapter 2:

$$\mathbb{E}(M|A=a,C=c)=\beta_0+\beta_1 a+\beta_2^{'}c \tag{4.1}$$

and then for the outcome, use either a proportional hazards model of the form

$$\lambda_T(t|a,m,c)=\lambda_T(t|0,0,0)e^{\gamma_1 a+\gamma_2 m+\gamma_4^{'}c} \tag{4.2}$$

or an accelerated failure time model of the form

$$\log(T)=\theta_0+\theta_1 A+\theta_2 M+\theta_4^{'}c+\nu\varepsilon \tag{4.3}$$

where ε is a random variable following an extreme value distribution and ν is a scale parameter so that T follows a Weibull distribution. Using simulations, Tein and MacKinnon find that the difference method and product method give different results for the proportional hazards model but the same results for the accelerated failure time model. Their results raise the question of whether either of these methods for either of the models has a clear causal interpretation. We will begin with the interpretation of the product and difference methods in accelerated failure time models and discuss also incorporating exposure–mediator interactions in the analysis of direct and indirect effects with accelerated failure time models. We will then turn to similar considerations with the proportional hazards model. Later we will also consider another class of models, additive hazards models, and their use in the analysis of direct and indirect effects. Finally we will consider an alternative

approach to the analysis of direct and indirect effects in survival models based on weighting.

4.2. MEDIATION ANALYSIS WITH AN ACCELERATED FAILURE TIME MODEL

Let us first consider the accelerated failure time model. We note first that it is no coincidence that the product and difference methods coincided in the simulations of Tein and MacKinnon (2003) for the accelerated failure time model in (4.2). It can in fact be shown analytically that, provided that the models are correctly specified and there are no interactions in model (4.3), these two approaches will indeed coincide (VanderWeele, 2011ba). The result holds for arbitrary distributions of ε in model (4.3)—that is, not just Weibull models. Not only do these two approaches coincide, but it is also the case that provided that the models are correctly specified and that there are no exposure–mediator interactions in the model and that our four assumptions about confounding hold, both the product and the difference method will yield estimates that can be interpreted as measures of natural direct and indirect effects on the log mean survival time scale (see the Appendix for greater formality). The natural direct effect on the log mean survival time scale is equal to $\theta_1(a-a^*)$ and the natural indirect effect on the log mean survival time scale is equal to $\beta_1\theta_2(a-a^*)$. In other words, we once again obtain the result that the exposure coefficient in model (4.3) for the outcome is a measure of the direct effect, and the product of the exposure coefficient in model for the mediator times the mediator coefficient in model (4.3) for the outcome is a measure of the indirect effect. If we exponentiate the direct effect, we get $\exp\{\theta_1(a-a^*)\}$, which can be interpreted as the ratio by which the direct effect increases the mean survival time. If we exponentiate the indirect effect, we get $\exp\{\beta_1\theta_2(a-a^*)\}$, which can be interpreted as the ratio by which the indirect effect increases the mean survival time.

For the accelerated failure time model, these analytic expressions can also be extended so as to allow for exposure–mediator interaction in model (4.3). Suppose we extend model (4.3) to allow for such interaction:

$$\log(T) = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4' c + \nu\varepsilon \qquad (\mathbf{4.4})$$

If model (4.4) holds for the outcome and our linear regression model (4.1) holds for the mediator, and our four confounding assumptions (A2.1)–(A2.4) hold, then the natural direct and indirect effects on the mean survival time ratio scale conditional on $C = c$ are given by (VanderWeele, 2011b)

$$\begin{aligned} NIE^{AFT} &= \exp[(\theta_2\beta_1 + \theta_3\beta_1 a)(a-a^*)] \\ NDE^{AFT} &= \exp[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a-a^*) \\ &\quad + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})] \end{aligned}$$

where the first expression is the natural indirect effect and the second expression is the natural direct effect, and where σ^2 is the variance of the error term in regression

model for the mediator. These results hold for arbitrary distributions for ε in model (4.4) but do require a normally distributed mediator in the mediator regression model (4.1). Note that when there is no interaction ($\theta_3 = 0$), the expressions reduce to those given above and considered by Tein and MacKinnon (2003). The expressions given here for the accelerated failure time model are analogous to those given in Chapter 2 for odds ratios for mediation analysis for a dichotomous outcome. Expressions for standard errors for these direct and indirect effects could likewise be adapted from those given in Chapter 2 (see Appendix).

Similarly, under (A2.1)–(A2.4), if the mediator is binary and follows a logistic regression model

$$\text{logit}\{P(M = 1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c \qquad \textbf{(4.5)}$$

and if the time-to-event outcomes follows the accelerated failure time model in (4.4), then natural direct and indirect effects on the mean survival time ratio scale conditional on $C = c$ are given by expressions similar to those in Chapter 2 for a binary mediator and binary outcome:

$$NIE^{AFT} = \frac{\{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2' c)\}}{\{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}$$

$$NDE^{AFT} = \frac{\exp(\theta_1 a)\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}{\exp(\theta_1 a^*)\{1 + \exp(\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c)\}}$$

Again the results hold for arbitrary distributions for ε in model (4.4) and expressions for standard errors for these direct and indirect effects could likewise be adapted from those for a binary outcome and binary mediator (see Appendix).

4.3. MEDIATION ANALYSIS WITH A PROPORTIONAL HAZARDS MODEL

Let us now turn to the proportional hazards model in (4.2). With the proportional hazards model, somewhat analogous results can be obtained, but only when the outcome is rare. Specifically, consider an extension to model (4.2) which allows for exposure–mediator interaction:

$$\lambda_T(t|a,m,c) = \lambda_T(t|0,0,0)e^{\gamma_1 a + \gamma_2 m + \gamma_3 am + \gamma_4' c} \qquad \textbf{(4.6)}$$

If model (4.6) holds for the outcome and the linear regression model (4.1) for the mediator is correctly specified, and our four confounding assumptions (A2.1)–(A2.4) hold, then provided the outcome is rare, natural direct and indirect effects on the hazard ratio scale are given by (VanderWeele, 2011a,b)

$$\begin{aligned} NIE^{PH} &= \exp[(\gamma_2\beta_1 + \gamma_3\beta_1 a)(a - a^*)] \\ NDE^{PH} &= \exp[\{\gamma_1 + \gamma_3(\beta_0 + \beta_1 a^* + \beta_2' c + \gamma_2\sigma^2)\}(a - a^*) \\ &\quad + 0.5\gamma_3^2\sigma^2(a^2 - a^{*2})] \end{aligned}$$

where σ^2 is again the variance of the error term in regression model for the mediator. The expressions are likewise analogous to those given in Chapter 2 for a dichotomous outcome, but these expressions only apply for a rare outcome. Natural indirect and direct effect hazard ratios (rather than log hazard ratios) can be obtained by exponentiating the right-hand side of the equalities. It is also the case that when there is no exposure–mediator interaction as in model (4.2) and when the outcome is rare, then the product and difference methods will coincide approximately (VanderWeele, 2011b); however, as shown by Tein and MacKinnon (2003) the product and difference methods may diverge when the outcome is common.

In the general setting of a proportional hazards model with non-rare outcome, unfortunately, neither the product method or the difference method for the proportional hazards model have any sort of clear causal interpretation as a measure of effect. Tein and MacKinnon (2003) show that not only can the product and difference methods diverge, but they may even be of opposite signs! It is, however, the case that even if the outcome is common, the product method using the linear regression model for the mediator and proportional hazards model (4.2) or (4.6) for the outcome will at least provide a valid test for whether there is any mediated effect, provided that the models are correctly specified and our four assumptions about confounding hold (VanderWeele, 2011b). With the proportional hazards model and a common outcome, the product method can thus be useful at least in testing the hypothesis of any mediated effect. However, neither the product method nor the difference method should itself in general be used as a measure of an indirect effect. In Section 4.5 we discuss an approach that can be used to estimate direct and indirect effects in the proportional hazards model with a common outcome.

When the outcome is rare and we have a binary mediator, we can also proceed with the estimation of natural direct and indirect effects. Suppose the mediator is binary and follows a logistic regression model

$$\text{logit}\{P(M = 1|a, c)\} = \beta_0 + \beta_1 a + \beta_2' c \tag{4.7}$$

If the time-to-event outcomes follow the proportional hazards model in (4.6), then natural direct and indirect effects on the hazard ratio scale conditional on $C = c$ are are given by expressions similar to those in Chapter 2 for a binary mediator and binary outcome:

$$NIE^{PH} = \frac{\{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)\}\{1 + \exp(\gamma_2 + \gamma_3 a + \beta_0 + \beta_1 a + \beta_2' c)\}}{\{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c)\}\{1 + \exp(\gamma_2 + \gamma_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}$$

$$NDE^{PH} = \frac{\exp(\gamma_1 a)\{1 + \exp(\gamma_2 + \gamma_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}{\exp(\gamma_1 a^*)\{1 + \exp(\gamma_2 + \gamma_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c)\}}$$

and again expressions for standard errors for these direct and indirect effects could likewise be adapted from those for a binary outcome and binary mediator (see Appendix). SAS macros to implement these analyses using accelerated failure time or proportional hazards models was recently provided by Valeri and VanderWeele (2015).

4.4. MEDIATION WITH AN ADDITIVE HAZARD MODEL

4.4.1. Analytic Results with the Additive Hazards Model

Lange and Hansen (2011) present an approach to mediation analysis with survival data using an additive hazard model. In the most basic form they consider, the model for the time-to-event outcome can be written as

$$\lambda_T(t|a,m,c) = \lambda_0 + \lambda_1 a + \lambda_2 m + \lambda_4' c \tag{4.7}$$

and they use the same linear regression model (4.1) for the mediator:

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

They show that on the hazard difference scale, if model (4.7) for the outcome and (4.1) for the mediator are correctly specified with a normally distributed mediator, and if the four confounding assumptions (A2.1)–(A2.4) hold, then natural direct and indirect effects are given by

$$NIE^{AH} = \beta_1 \lambda_2 (a - a^*)$$
$$NDE^{AH} = \lambda_1 (a - a^*)$$

where the first expression is the indirect effect and the second expression is the direct effect on the hazard difference scale.

However, the additive hazard model approach described by Lange and Hansen (2011) also has additional flexibility. The approach can allow the hazard functions to vary over time, and can be extended to incorporate exposure–mediator interactions as well, though in the latter case no analytic expressions are yet available. Lange and Hansen (2011) do, however, provide an R package to implement this additive hazard approach, and the interested reader is referred to that paper for details. The generality of the approach proposed is impressive, and the methodology and software provided will certainly be of use for causal mediation analysis within a survival context.

4.4.2. Example of Mediation Analysis Using the Additive Hazards Model. Socioeconomic Status and Work Conditions

Lange and Hansen (2011) use this approach to examine the extent to which the effect of socioeconomic status on the onset of long-term absence from work due to illness is mediated by how demanding the physical work environment is. They analyze a random sample of 11,437 people collected as part of the Danish Work Environment Cohort Study. They analyze men and women separately. We will present here their analyses of the men. Long-term sickness absence is defined as 3 weeks of consecutive sickness absence. Follow-up in the study was 18 months from 1 January 2006 to 30 June 2007. The physical work environment was measured as a score between 1 and 100 (with 100 corresponding to the most physically

demanding), taken as an average of response to five questions. Lange and Hansen used a log transformation of this score because it was more normally distributed. Socioeconomic status was measured in five categories, with I indicating the highest SES group and V indicating the lowest SES group.

They first fit a regression of the transformed physical environment score on SES adjusted for age and family status as confounders. They then fit the additive hazard model to the onset of long-term sickness absence with age, family status, SES, and physical work score as covariates. Their analyses suggested that there was no evidence of time-dependent hazards or interactions. They found that in comparing men in SES group V versus I, those in group V (the lowest SES group) had an average log physical work score 1.83 units higher than the men in SES group I (the highest SES group) after adjustment for age and family status. Likewise, the number of long-term sickness absences was 10.3 (95% CI: 6.5, 14.1) persons per week per 10,000 men at risk, greater for men in group V than for men in group I. Using the additive hazard approach for natural direct and indirect effects, they found a natural indirect effect of 4.3 (95% CI: 2.4, 6.2) additional persons per week per 10,000 men at risk. The natural direct effect in this case was estimated to be 6.01 (95% CI: 1.8, 10.2) additional persons with long-term absences per week per 10,000 men at risk. According to these estimates, 43% (95% CI: 22%, 74%) of the increased rate of absences can be attributed to the pathway through physical work environment. In other words, because there is no interaction and the natural and controlled direct effects coincide if an intervention could improve the physical work environment of the low SES group (group V) to the level of the high SES group (group I) without affecting other aspects of social deprivation, according to this estimate 43% (95% CI: 22%, 74%) of the effect of socioeconomic status on long-term sickness absence could be eliminated.

4.5. A WEIGHTING APPROACH TO DIRECT AND INDIRECT EFFECTS WITH SURVIVAL OUTCOMES

4.5.1. Description of the Weighting Approach

As noted above, with the proportional hazard model, the product and difference methods for mediation analysis will only be valid if the outcome is rare. The proportional hazards model is, however, used in many settings with common outcomes. Fortunately, rather than using a regression-based approach to the estimation of direct and indirect effects in a survival context, an alternative approach can be used for direct and indirect effects with the proportional hazards model that employs weighting (Lange et al., 2012) that can be used with a common outcome. The approach proceeds by fitting what is sometimes called a "natural effects model" (Lange et al., 2012; Vansteelandt et al., 2012a). Technical details are given in the Appendix, but here we describe how the approach can be implemented to obtain direct and indirect effects with the proportional hazards model with a common outcome.

The basic analytic approach is to fit a proportional hazards model to the data but with two modifications: First, the model is fit with a set of weights that are described below; second, the dataset is modified to have two copies of each individual with an additional variable included as well. For simplicity here we will describe this approach with a binary exposure, but the approach can be extended to cover categorical and continuous exposures as well (Lange et al., 2012).

We first create a new dataset that has two copies, instead of one, of each individual, as well as an additional new variable. This additional new variable we will call A^*. This variable A^* will be equal to the A for the first copy of the individual and will be equal to $1 - A$ for the second copy. To carry out the weighting, two sets of weights are needed, one for the exposure and one for the mediator. For each individual i in the sample, the exposure weight is calculated by

$$w_i^A = \frac{P(A = a_i)}{P(A = a_i | C = c_i)}$$

where a_i and c_i are the actual values of the exposure and the covariates for individual i. In the denominator of the weight we have the probability of receiving the treatment the individual in fact received conditional on the covariates C taking the value c_i; that is, this is an individual's predicted probability of actually having the treatment that they received, given they had their covariate values. For a binary exposure, this probability could be estimated, for example, by using a logistic regression model (e.g., see code below). The numerator of the weight for the exposure is just the overall proportion of those with the exposure in the population. The ratio of these two is our weight for the exposure. We also similarly obtain a weight for the mediator as follows. For each individual i in the sample, the mediator weight is calculated by

$$w_i^M = \frac{P(M = m_i | A = a_i^*, C = c_i)}{P(M = m_i | A = a_i, C = c_i)}$$

where m_i is likewise the actual value of the mediator for individual i. The denominator of this weight w_i^M is the probability of having the value of the mediator that the individual in fact had conditional on the values of the exposure and covariates, $A = a_i$ and $C = c_i$, that the individual actually had. For a binary mediator, this predicted probability could likewise be obtained by a logistic regression. The numerator probability is the predicted probability of having the value of the mediator that the individual in fact had if the individual had the values of the covariates, $C = c_i$, but had a value of the exposure which was in fact equal to $A = a_i^*$ (i.e., equal to what the individual actually had for the first replication and to 1 minus this value for the second replication). Once again, for a binary mediator, this predicted probability could be also be obtained by logistic regression with appropriate coding. We provide an example of such code in SAS below.

Once these weights are estimated, we can obtain an overall weight for the individual by taking a product of these weights:

$$w_i = w_i^A \times w_i^M$$

We can then obtain natural direct and indirect effects estimates on the log hazards scale by fitting the following proportional hazards model:

$$\lambda_T(t|a,a^*) = \lambda_T(t|0,0)e^{\kappa_1 a + \kappa_2 a^*}$$

but where each individual is weighted by the overall weight $w_i = w_i^A \times w_i^M$; that is, we just use a proportional hazards model in which our variables A and A^* are the only covariates included in the model. Provided that the no unmeasured confounding assumptions (A2.1)–(A2.4) hold and our models for the predicted probabilities are correctly specified, the coefficient κ_1 for A in this weighted proportional hazards model will give us an estimate of the natural direct effect on the log hazards scale and the coefficient κ_2 for A^* in this weighted proportional hazards model will give us an estimate of the natural indirect effect on the log hazards scale. Note that the covariates C are not included in the regression model that we fit for Y. We regress Y only on A and A^*; control for confounding by the covariates C is essentially done by weighting. Contrary to what is claimed in Lange et al. (2012), "robust" or "sandwich" standard errors for κ_1 and κ_2 will not necessarily be valid for the natural direct and indirect effects. Bootstrapping must be used.

The approach can also accommodate potential interaction. To do so, the final proportional hazards model that is fit is instead

$$\lambda_T(t|a,a^*) = \lambda_T(t|0,0)e^{\kappa_1 a + \kappa_2 a^* + \kappa_3 a a^*}$$

and this is again weighted by the overall weight $w_i = w_i^A \times w_i^M$. In this case, the natural direct effect is given by $\kappa_1 + \kappa_3 a^*$ and the natural indirect effect is given by $\kappa_1 + \kappa_3 a$. As discussed in the Appendix, this can essentially then accommodate exposure–mediator interaction.

4.5.2. Code for the Weighting Approach

Here we present some SAS code, adapted from Lange et al. (2012) for implementing the weighting approach above for the proportional hazards model. We will assume that the exposure is binary and the mediator is also binary. See Lange et al. (2012) for additional code for other settings in both SAS and R. Suppose that the name of the dataset was "mydata," the exposure variable was "a," the mediator variable was "m," a time to event outcome with record time was called "mytime," and an event indicator was called "event." Suppose also we have three baseline covariates "c1," "c2," and "c3." We can first construct the weights for the exposure using the following code. We will need to modify the dataset to obtain the mediator weights but we do not need to do so for the exposure weights.

```
proc logistic data=mydata descending;
   model a= ;
   output out=mydata predicted=pna;
run;

proc logistic data=mydata descending;
```

```
   model a= c1 c2 c3;
   output out=mydata predicted=pda;
run;
```

We now modify the dataset in the manner described above and we also estimate the mediator weights. We construct a new variable "astar." We replicate each observation twice: once with astar = 0 and once with astar = 1. Whenever astar = a, we also put Mtemp=m; thus Mtemp is missing for all "created" observations, but equal to the observed mediator for the actual observations. We do this so that when using Mtemp as outcome in the model for the mediator, the model fit will be done using only the actual data, but fitted probabilities are computed for both actual and "created" observations.

```
DATA newMyData;
    SET myData;
    astar = 0; Mtemp =.;
    IF a=astar then Mtemp = m;
    OUTPUT;
    astar = 1; Mtemp =.;
    IF a=astar then Mtemp = m;
    OUTPUT;
RUN;
```

Next we fit a logistic regression for the mediator (m) on exposure (a) and baseline confounders and compute predicted values. As we later want to change the values of the exposure when doing predictions, we use a copy of the exposure variable in the model fit.

```
DATA newMyData;
    SET newMyData;
    Atemp = a;
RUN;

PROC logistic data=newMyData descending;
    model Mtemp = Atemp c1 c2 c3;
    output out=newMyData predicted=pdm;
RUN;
```

The above code will give us the predicted probability for the denominator in the mediator weight. We now set Atemp = astar and repeat the logistic regression and predicted probabilities to get the numerator for the mediator weight.

```
DATA newMyData;
    SET newMyData;
    Atemp = astar
RUN;

PROC logistic data=newMyData descending;
    model Mtemp = Atemp c1 c2 c3;
    output out=newMyData predicted=pnm;
RUN;
```

Finally the weights can be computed by

```
DATA newMyData;
```

```
    SET newMyData;
    w = (pna*pnm)/(pda*pdm);
RUN;
```

We can then fit a proportional hazards model for direct and indirect effects weighted by the weights we have calculated.

```
PROC PHREG data=newMyData COVSANDWICH;
    CLASS a (param=ref ref=''1'') astar (param=ref ref=''1'') event;
    MODEL ttt*event(0) = a astar;
    WEIGHT w;
    ID id;
RUN;
```

The coefficient for "a" in the output of this weighted proportional hazards model regression will give an estimate for the natural direct effect. The coefficient for "astar" in the output of this weighted proportional hazards model regression will give an estimate for the natural indirect effect. Bootstrapping would have to be used to obtain valid standard errors.

The weighting approach described above could also be used for binary and continuous outcomes by simply changing the final weighted proportional hazards model to a weighted logistic or linear regression model (Lange et al., 2012). However, as discussed in Chapter 7, this approach is generally less efficient (i.e., it results in larger standard errors and wider confidence intervals) than the approach we described in Chapter 2. For other variants of this weighting approach using categorical or continuous exposures or mediators, see Lange et al. (2012).

4.6. SENSITIVITY ANALYSIS WITH SURVIVAL DATA

4.6.1. Sensitivity Analysis for Total Effects on the Hazard Difference and Hazard Ratio Scales

First, consider total effects on the hazard difference scale. Suppose we obtain estimates of such effects (e.g., using an additive hazards model), controlling for our measured covariates which we will denote by C. Suppose now that there is also an unmeasured covariate U and that we would have controlled for confounding for the effect of the exposure on the outcome if we had controlled for C and U but not simply by controlling for C alone. Under some simplifying assumptions, we can proceed with a very easy-to-use sensitivity analysis technique to assess what the estimate would have been had we been able to adjust for U as well. Specifically we will assume that the unmeasured variable U is binary and that the effect of U on the outcome on the hazard difference scale is the same for both exposure groups (i.e., no interaction between the effects of U and the exposure on the additive scale). These assumptions can be relaxed, and a more general approach is presented in the Appendix. We will also assume here, and in all of the results below, that the outcome is relatively rare. Under these assumptions we can carry out sensitivity analysis by specifying two sensitivity analysis parameters. We need to specify the effect of U

on the outcome on the hazard difference scale, conditional on the exposure and the covariates; we will call this parameter γ. We also need to specify the difference in the prevalence of U amongst the exposed and the unexposed; we will call this parameter δ. We can then calculate a bias factor by taking the product $\gamma\,\delta$. To obtain a "corrected"estimate of the effect on the hazard difference scale (i.e., what we would have obtained had we controlled for U as well), we can simply take our estimate from the observed data, controlling only for C, and subtract the bias factor $\gamma\,\delta$ from the estimate; under the simplifying assumptions above, we can also obtain a corrected confidence interval by subtracting the bias factor $\gamma\,\delta$ from both limits of the confidence interval. We could then vary the sensitivity analysis parameters according to values that were thought plausible or as informed by external information or expert opinion to see how sensitive estimates were to the possibility of unmeasured confounding.

A similar approach can be carried out on the hazard ratio scale. For the hazard ratio scale we will again assume a rare outcome and a binary unmeasured confounder U, but now we will assume that the effect of U on the outcome on the hazard ratio scale does not vary across exposure groups (i.e., no interaction between the effects of U and the exposure on the hazard ratio scale). Under these simplifying assumptions we will specify three parameters to carry out the sensitivity analysis. We will now let γ denote the effect of U on the outcome on the hazard ratio scale, conditional on the exposure and the covariates. And we will let π_1 and π_0 denote the prevalence of U amongst the exposed and unexposed respectively, conditional on measured covariates C. Once we have specified these parameters, we can obtain a bias factor on the hazard ratio scale by the formula $[1 + (\gamma - 1)\pi_1]/[1 + (\gamma - 1)\pi_0]$. We can then simply take our estimate of the hazard ratio from the observed data, controlling only for C, and now divide the estimate by this bias factor to obtain a corrected estimate (what we would have obtained had we controlled for U as well). Under the simplifying assumptions above, we can also obtain a corrected confidence interval by dividing both limits of the confidence interval by the bias factor.

The two formulae presented above for the additive hazard scale and the hazard ratio scale are exactly the same for those presented on the additive scale for binary or continuous outcomes or on the ratio scale for binary outcomes in Chapter 3.

4.6.2. Sensitivity Analysis for Direct and Indirect Effects on the Hazard Difference and Hazard Ratio Scales

We can likewise use simple sensitivity analysis formulae for direct and indirect effects (cf. VanderWeele, 2013c). The approach here for hazard differences or ratios is essentially the same as that in Chapter 3 for risk differences and ratios and merely replaces the sensitivity analysis parameter for the effect of the unmeasured confounder on the outcome which had been on the risk difference or ratio scale in Chapter 3 to one on the hazard difference or ratio scale here. More specifically, for controlled direct effects we need to control for both exposure–outcome and mediator–outcome confounding. Suppose that there were an unmeasured mediator–outcome confounder U but we controlled only for measured covariates C. Suppose that if we had controlled for both C and U we would have controlled for exposure–outcome and mediator–outcome confounding. We can once again

use sensitivity analysis to examine how such an unmeasured confounder might change estimates of the controlled direct effect on the hazard difference or hazard ratio scale.

For the hazard difference scale, we again specify two parameters. We will again assume that the outcome is rare and we assume a binary unmeasured confounder U such that the effect of U on the outcome on the hazard difference scale is the same for both exposure groups. The approach is very similar to that for total effects, but the interpretation of the parameters is somewhat different. Suppose controlling only for measured covariates C, we obtained (e.g., using an additive hazard model) an estimate of controlled direct effect with the mediator fixed to m on the hazard difference scale—that is, the hazard difference of the exposure on the outcome conditional on C and $M = m$. For the first parameter, we specify the effect of U on the outcome on the hazard difference scale, conditional on the exposure, the mediator, and the covariates; note that this is the effect of the unmeasured confounder U on the outcome not through the mediator M; we will call this parameter γ_m. We also need to specify the difference in the prevalence of U amongst the exposed conditional on $M = m$ and the prevalence of U among the unexposed, again now conditional on $M = m$; we will call this parameter δ_m. See Chapter 3 for further discussion of the interpretation of these prevalences of the unmeasured confounder U conditional on the mediator value $M = m$. We can then compute a bias factor for the controlled direct effect on the hazard difference scale by taking the product $\gamma_m\delta_m$. We can obtain a corrected estimate of the controlled direct effect on the hazard difference scale by subtracting the bias factor from the estimate and, under the simplifying assumptions above, we can also obtain a corrected confidence interval by subtracting the bias factor from both limits of the confidence interval.

For the hazard ratio scale, we assume again a rare outcome and that U is a binary unmeasured mediator–outcome confounder; we now assume that the effect of U on the outcome on the hazard ratio scale is the same for both exposure groups. Suppose, controlling only for measured covariates C, we had obtained (e.g., using a proportional hazards model) an estimate of the hazard ratio for the effect of the exposure on the outcome with the mediator fixed to value m. We specify three sensitivity analysis parameters. We now let γ_m denote the effect of U on the outcome on the hazard ratio scale, conditional on the exposure, the mediator, and the covariates; again this is the effect of U on the outcome not through the mediator. We let π_{1m} and π_{0m} denote the prevalence of U amongst the exposed conditional on $M = m$ and the prevalence of U among the unexposed, again conditional on $M = m$. We can obtain a bias factor on the hazard ratio scale by the formula $[1 + (\gamma_m - 1)\pi_{1m}]/[1 + (\gamma_m - 1)\pi_{0m}]$. We can take our estimate of the hazard ratio from the observed data, controlling only for C, and divide the estimate by this bias factor to obtain a corrected estimate and, under the simplifying assumptions above, we can also obtain a corrected confidence interval by dividing both limits of the confidence interval by the bias factor as well. More general expressions for controlled direct effects for both the hazard difference and hazard ratio scales, which do not require the simplifying assumptions, are given in the Appendix. The expressions given here are analogous to those given in Chapter 2 for binary and continuous variables and, as shown in the Appendix, they are applicable to time-to-event outcomes as well when the outcome is rare.

As noted above, for natural direct and indirect effects to be identified, we need to control for [assumption (A2.1)] exposure–outcome confounding, [assumption (A2.2)] mediator–outcome confounding, and [assumption (A2.3)] exposure–mediator confounding, and we need that [assumption (A2.4)] none of the mediator–outcome confounders are affected by the exposure. Under the further assumption that the exposure and mediator do not interact in their effects on the outcome, the controlled direct effect will equal the natural direct effect; and the natural indirect effect will equal the total effect minus the controlled direct effect. If we were concerned about unmeasured mediator–outcome confounding but were willing to assume absence of exposure–mediator interaction, then we could apply the techniques described above for controlled direct effects, under the assumptions described above, and use them for the natural direct effects. We could also then use the opposite of these formulas for the natural indirect effects; that is, for the natural indirect effect on the hazard difference scale we could add the bias factor $\gamma_m\delta_m$ to the natural indirect effect estimate (whereas we would subtract this from the natural direct effect), and for the natural indirect effect on the hazard ratio scale we could multiply the natural indirect effect estimate by $[1 + (\gamma_m - 1)\pi_{1m}]/[1 + (\gamma_m - 1)\pi_{0m}]$ (whereas for the natural direct effect we would divide the estimate by this expression). More general sensitivity analysis techniques for natural direct and indirect effects for the hazard difference scale which do not assume the absence of exposure–mediator interaction and which do not make the simplifying assumptions above are given in the Appendix.

4.7. DISCUSSION

The discussion above has provided expressions for natural direct and indirect effects for the accelerated failure time model, and for the proportional hazards model when the outcome is rare, and for additive hazard models. As was already noted in Chapter 2, a major contribution of the counterfactual approach to causal mediation analysis has been to clarify the no-confounding assumptions required for the identification of direct and indirect effects. Within the context of survival data, the counterfactual approach also clarifies when different methods for direct and indirect effects can be interpreted as measures of effects rather than simply as a test for a mediated effect. The causal inference approach clarifies further on what scale these measures apply when they can be so interpreted. The observations of Tein and MacKinnon (2003) can be given a more rigorous formulation, and the approach has been extended to allow for exposure–mediator interactions. The proportional hazards model is perhaps used most commonly in empirical research with survival data. As we have seen, a fairly simple approach is possible, analogous to that described in Chapter 2, if the outcome is rare. If the outcome is common, then the weighting approach can still be employed. However, if an investigator is willing to work with new models, the additive hazard model constitutes a very general alternative to mediation analysis with survival data. Other methods for the estimation of direct and indirect effects with survival data have also been proposed. Martinussen et al. (2011) consider the use of additive hazards model to estimate

controlled direct effects in the presence of exposure-induced mediator–outcome confounding (i.e., violations of the fourth assumption about confounding). This will be the topic of Chapter 5. Tchetgen Tchetgen (2011) consider a combination of the regression-based approach and weighting approaches to survival data that are more robust to model misspecification and will be discussed further in Section 7.1. The development of approaches for mediation analysis with survival data is still an active area of methodological research, and further developments and software tools are likely to be available in the years ahead.

5

Multiple Mediators

The methods for mediation that we have considered up until this point all allow for only one mediator. In many applications, several mediators may be of interest; we may be interested in assessing the extent to which the effect of an exposure on some outcome is mediated by several mediators considered together. Even when we were only interested in one mediator, we have also seen in previous chapters that to identify direct and indirect effects, we needed to make an assumption that there was no mediator–outcome confounder affected by the exposure. Such a variable is in fact, also a second mediator. If a variable is affected by the exposure and itself affects the outcome, then it too lies on the pathway from the exposure to the outcome and we are once again in a setting with multiple mediators. The focus of this chapter will be mediation analysis for several mediators. We will consider how to go about assessing the extent to which the effect of an exposure on some outcome is mediated by several mediators considered together, and we will also consider methods to handle situations in which only one mediator is of interest but there is an exposure-induced mediator confounder as described above.

When multiple mediators are of interest, one approach would be to consider the mediators one at a time. As will be seen in this chapter, however, this will in general require that the mediators do not affect one another. The methods in this chapter will instead allow an investigator to assess mediation with multiple mediators simultaneously and will also be able to accommodate cases in which the mediators affect one another. We will allow for potential exposure–mediator interactions as well as some settings with mediator–mediator interactions. When the ordering of the mediators is known, we will discuss how the approach can be applied sequentially. We give two different statistical techniques—one based on regression and one based on weighting—to estimate direct and indirect effects in these cases. We show how the approach is robust to unmeasured common causes of two or more mediators whereas handling the mediators one-by-one will fail in these cases. When mediators are handled one-by-one, the sum of the proportion mediated for the mediators can sometimes total more than 100%, even if the direction of mediation is the same for all mediators. As will be discussed in this

chapter, if this phenomenon arises, it is because the mediators in fact affect one another or because of mediator–mediator interactions. The methods described below can handle these situations. These various topics will be the content of the current chapter.

5.1. REGRESSION-BASED APPROACHES TO MULTIPLE MEDIATORS

5.1.1. Definitions and Assumptions for Multiple Mediators

Suppose as before that A is the exposure, Y the outcome, and C a set of pre-exposure covariates. Suppose also that there are now multiple mediators of interest, $\mathbf{M} = (M^{(1)}, \ldots, M^{(K)})$, and that we are interested in the effects mediated through $(M^{(1)}, \ldots, M^{(K)})$ jointly and the effects through pathways other than through $(M^{(1)}, \ldots, M^{(K)})$. We can define controlled direct effects and natural direct and indirect effects in a similar way as before by simply replacing our single mediator M with the entire vector of mediators $\mathbf{M} = (M^{(1)}, \ldots, M^{(K)})$. To proceed, we will also need the four assumptions about confounding [assumptions (A2.1)–(A2.4)] to hold but now with respect to the whole set of mediators $\mathbf{M} = (M^{(1)}, \ldots, M^{(K)})$. In other words, we need to control for all [assumption (A2.1)] exposure–outcome, [assumption (A2.2)] mediator–outcome, and [assumption (A2.3)] exposure–mediator confounders, but now the assumption about the mediator–outcome confounders [assumption (A2.2)] must hold for all of the mediators, not just one, and likewise the assumption about exposure–mediator confounders [assumption (A2.3)] must hold for all of the mediators, not just one. Our fourth [assumption (A2.4)] again requires that there be no effect of the exposure that confounds the mediator–outcome relationship, and this assumptions again must hold for all of the mediators. However, in some sense, we can now handle violations of the assumption because if there were such a variable, then to proceed we could include the variable as well in the mediator vector $\mathbf{M}$ and this fourth assumption would not be violated. Since the assumptions must hold for the whole vector of potential mediators, let us call these assumptions (A5.1)–(A5.4).

5.1.2. A Regression-Based Approach for Multiple Mediators with a Continuous Outcome

Under these assumptions, the natural direct and indirect effects can once again be estimated using a regression-based approach. We will use one regression for the outcome Y and a separate regression for each of the mediators. We will begin with the case of a continuous outcome with continuous mediators and no interactions and we will consider (a) extensions allowing for exposure–mediator interactions and (b) some settings with mediator–mediator interaction as well as binary mediators and binary outcomes below.

Suppose then that our confounding assumptions (A5.1)–(A5.4) held for the vector of mediators $\mathbf{M} = (M^{(1)}, \ldots, M^{(K)})$ and that the following regressions were fit to the data:

$$\mathbb{E}[Y|a, \mathbf{m}, c] = \theta_0 + \theta_1 a + \theta_2^{(1)} m^{(1)} + \theta_2^{(2)} m^{(2)} + \cdots + \theta_2^{(K)} m^{(K)} + \theta_4' c$$

$$\mathbb{E}[M^{(i)}|a, c] = \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c \quad \text{for } i = 1, \ldots, K$$

Natural direct and indirect effects are then given by (VanderWeele and Vansteelandt, 2013)

$$\begin{aligned} CDE(m) &= \theta_1(a - a^*) \\ NDE &= \theta_1(a - a^*) \\ NIE &= [\beta_1^{(1)}\theta_2^{(1)} + \cdots + \beta_1^{(K)}\theta_2^{(K)}](a - a^*) \end{aligned} \tag{5.1}$$

The direct effects are perhaps exactly what one might expect: simply the coefficient for the exposure, θ_1, in the model that contains all of the mediators. The natural indirect effect is equal to the sum over the various mediators, $M^{(1)}, \ldots, M^{(K)}$, of the product of the coefficient for the exposure in the model for the mediator ($\beta_1^{(k)}$ for the kth mediator) and the coefficient for the mediator ($\theta_2^{(k)}$ for the kth mediator) in the model for the outcome that has all the mediators. The indirect effect has this fairly intuitive form. Note, however, that this is different from applying the approach to mediation for a single mediator described in Chapter 2, one mediator at a time, and then summing up the indirect effects. This is because if the mediators were handled one at a time, then a different regression for Y would be fit for each mediator and only one mediator would be included in each of these regressions. The approach described in this section fits only a single regression for Y which includes all of the mediators under consideration.

In fact these two approaches will coincide if the mediators do not affect one another (or more precisely, if the mediators are independent of one another conditional on A and C), but they will diverge otherwise. They will diverge if the mediators affect one another because certain pathways will be counted twice if the mediation analysis is done one at a time. For example, if there are two mediators where $M^{(1)}$ affects $M^{(2)}$ as in Figure 5.1 and if the analysis were done one mediator at a time, then the path $A - M^{(1)} - M^{(2)} - Y$ would be included in the indirect effect both for the analysis for $M^{(1)}$ and for the analysis for $M^{(2)}$. If the two "indirect effects" were summed, the path would essentially be counted twice. The approach described in this section circumvents this difficulty by fitting only one regression for Y. When the mediators affect one another, the approach of handling one mediator at a time also suffers from another difficulty, which is that for the second (and potentially each subsequent) mediator, assumption (A5.4) will not hold if the mediators are considered one at a time. This is because $M^{(1)}$ may be affected by A and in turn

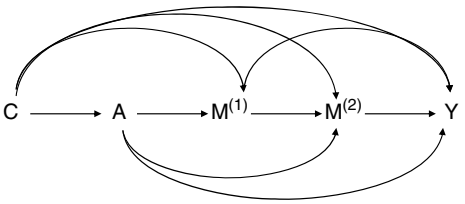

Figure 5.1 Mediation with two mediators of interest.

affect both $M^{(2)}$ and Y and thus would be a mediator–outcome confounder affected by the exposure when $M^{(2)}$ alone is considered as the mediator. The assumption about no exposure-induced mediator–outcome confounding may hold with regard to a whole collection of mediators $\mathbf{M} = (M^{(1)}, \ldots, M^{(K)})$ without holding for each mediator individually. When mediators are considered one mediator at a time, natural direct and indirect effects will thus often not be identified except under strong assumptions about the absence of interaction (cf. Robins, 2003). As will be seen in the next subsection, however, the approach described here will also be able to be used even in the presence of interaction. Moreover, later we will see that even if the mediators are in fact independent of one another, the approach described here will be robust to unmeasured common causes of two or more mediators, whereas the approach considering the mediators one at a time will not be robust to such unmeasured variables.

5.1.3. Exposure–Mediator Interactions, Binary Mediators, and Mediator–Mediator Interactions

In this subsection we will discuss how the simple approach above can be adapted to allow for exposure–mediator interactions, binary mediators, and to a certain extent mediator–mediator interactions. Because of the large number of possible variations on the types of interaction that can be employed, although it would be possible to give analytic standard errors for each of the variations below using the delta method, a new formula would have to be given in each case. Bootstrapping is therefore recommended in the estimation of standard errors. Suppose we wished to allow for an interaction between the exposure A and a mediator $M^{(i)}$ in the model for Y, for example, so that the outcome model became

$$\begin{aligned}\mathbb{E}[Y|a,\mathbf{m},c] = \theta_0 + \theta_1 a + \theta_2^{(1)} m^{(1)} + \theta_2^{(2)} m^{(2)} + \cdots \\ + \theta_2^{(K)} m^{(K)} + \theta_3^{(i)} a m^{(i)} + \theta_4' c\end{aligned}$$

The expressions for the controlled direct effect and natural direct and indirect effects in (5.1) are then modified by adding $\theta_3^{(i)} m^{(i)}(a - a^*)$ to the controlled direct effect, $\theta_3^{(i)}\{\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c\}(a - a^*)$ to the natural direct effect, and

$\theta_3^{(i)}\beta_1^{(i)}a(a-a^*)$ to the natural indirect effect so that the effects become

$$\begin{aligned} CDE(m) &= \{\theta_1 + \theta_3^{(i)}m^{(i)}\}(a-a^*) \\ NDE &= [\theta_1 + \theta_3^{(i)}\{\beta_0^{(i)} + \beta_1^{(i)}a^* + \beta_2^{(i)\prime}c\}](a-a^*) \\ NIE &= [\beta_1^{(1)}\theta_2^{(1)} + \cdots + \beta_1^{(K)}\theta_2^{(K)} + \theta_3^{(i)}\beta_1^{(i)}a](a-a^*) \end{aligned}$$

If a further interaction is thought present between the exposure and another of the mediators—for example, j—then the same terms could once again be added to these expressions: $\theta_3^{(j)}m^{(j)}(a-a^*)$ to the controlled direct effect, $\theta_3^{(j)}\{\beta_0^{(j)} + \beta_1^{(j)}a^* + \beta_2^{(j)\prime}c\}(a-a^*)$ to the natural direct effect, and $\theta_3^{(j)}\beta_1^{(j)}a(a-a^*)$ to the natural indirect effect—and similarly for other exposure–mediator interactions; any number of exposure–mediator interactions could be accommodated in this manner.

Thus far we have assumed all mediators are continuous. Suppose that one or more of the mediators is binary, say mediator j, and that we fit a logistic regression model for $M^{(j)}$ (instead of a linear regression model):

$$\text{logit}\{P[M^{(j)} = 1|a,c\} = \beta_0^{(j)} + \beta_1^{(j)}a + \beta_2^{(j)\prime}c$$

When all of the mediators were continuous and there were no exposure–mediator interactions, the direct and indirect effects were given by (5.1), namely,

$$\begin{aligned} CDE(m) &= \theta_1(a-a^*) \\ NDE &= \theta_1(a-a^*) \\ NIE &= [\beta_1^{(1)}\theta_2^{(1)} + \cdots + \beta_1^{(K)}\theta_2^{(K)}](a-a^*) \end{aligned} \qquad \textbf{(5.1)}$$

With one (or more) binary mediators, the expressions for the controlled direct effect and natural direct effect remain the same, but the expression for the natural indirect effect is modified. Instead of the term $\beta_1^{(j)}\theta_2^{(j)}(a-a^*)$ for the jth mediator, we would include in the natural indirect effect the term $\left(\frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c\}} - \frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}\right)\theta_2^{(j)}$; that is, we would replace $\beta_1^{(j)}(a-a^*)$ in the expression in (5.1) with $\left(\frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c\}} - \frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}\right)$ and we would likewise do this for each mediator that were binary. If we also wanted to allow for exposure–mediator interaction with a mediator that were binary, we would further add $\theta_3^{(j)}m^{(j)}(a-a^*)$ to the controlled direct effect, $\theta_3^{(j)}\frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}(a-a^*)$ to the natural direct effect,

and $\left(\frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c\}}-\frac{\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}{1+\exp\{\beta_0^{(j)}+\beta_1^{(j)}a^*+\beta_2^{(j)\prime}c\}}\right)\theta_3^{(j)}a$ to the natural indirect effect. We could once again do this for each mediator that were binary and for which we wanted to include an exposure–mediator interaction.

Finally, suppose that a binary exposure variable A were randomized and that no covariates were needed for assumptions (A5.1)–(A5.4) to hold. Suppose that we wanted to allow for mediator–mediator interaction between two mediators i and j as, for example, in the regression model:

$$\mathbb{E}[Y|a,\mathbf{m}]=\theta_0+\theta_1 a+\theta_2^{(1)}m^{(1)}+\theta_2^{(2)}m^{(2)}+\cdots+\theta_2^{(K)}m^{(K)}+\tau^{(ij)}m^{(j)}m^{(i)}$$

The controlled direct effect and natural direct effect will both be exactly the same as described above. However, the natural indirect effect needs to be modified further. In particular, we could fit a linear regression model for the product

$$\mathbb{E}[M^{(i)}M^{(j)}|a]=\beta_0^{(ij)}+\beta_1^{(ij)}a$$

We then would add the term $\tau^{(ij)}\beta_1^{(ij)}(a-a^*)$ to the natural indirect effect. Unfortunately, as discussed in the Appendix, if covariates C are included in the model, then this can lead to issues of model compatibility between the models for $M^{(i)}$ and $M^{(j)}$ and that for the product $M^{(i)}M^{(j)}$ (VanderWeele and Vansteelandt, 2013). In Section 5.2 we present an alternative weighting approach that circumvents this issue and is applicable to settings with mediator–mediator interactions.

5.1.4. A Regression-Based Approach for Multiple Mediators with a Binary Outcome

When the outcome is binary, rather than continuous, a similar approach to that described above, can also be employed; but as will be seen below, it is subject to some restrictions. First, the regression-based approach we describe below will only work when all of the mediators are continuous. Second, this regression-based approach will not allow for mediator–mediator interactions when the outcome is binary. In the next section we will also describe a weighting-based approach that can be used if the mediators are binary (or if some are binary and some are continuous) and that can also accommodate potential mediator–mediator interactions.

Suppose then that the outcome is binary and that all mediators are continuous and the following two models are fit to the data:

$$\text{logit}[P\{Y=1|a,\mathbf{m},c\}]=\theta_0+\theta_1 a+\theta_2^{(1)}m^{(1)}+\theta_2^{(2)}m^{(2)}+\cdots+\theta_2^{(K)}m^{(K)}+\theta_4'c$$

$$\mathbb{E}[M^{(i)}=1|a,c]=\beta_0^{(i)}+\beta_1^{(i)}a+\beta_2^{(i)\prime}c \text{ for } i=1,\ldots,K$$

Suppose also now that the mediators $M^{(1)},\ldots,M^{(K)}$ follow a multivariate normal distribution conditional on A and C. We show in the Appendix that when the models are correctly specified and when assumptions (A5.1)–(A5.4) hold and when the

outcome is rare (or the logistic regression model is replaced by a log-linear model for a common outcome with the effect measures interpreted as relative risks rather than odds ratios), then the log of the controlled direct effect and natural direct and indirect effect odds ratios are given approximately by

$$\log\{OR^{CDE}_{a,a^*|c}(m)\} = \theta_1(a - a^*)$$
$$\log\{OR^{NDE}_{a,a^*|c}(a^*)\} \approx \theta_1(a - a^*)$$
$$\log\{OR^{NIE}_{a,a^*|c}(a)\} \approx [\beta_1^{(1)}\theta_2^{(1)} + \cdots + \beta_1^{(K)}\theta_2^{(K)}](a - a^*)$$

These were the same expressions we had obtained in (5.1) for a continuous outcome. If we wish to allow for exposure–mediator interactions, we can simply add additional terms. Suppose we wished to allow for an interaction $\theta_3^{(i)} am^{(i)}$ between the exposure A and a mediator $M^{(i)}$ in the model for Y. The expressions for the controlled direct effect and natural indirect effects are then modified by adding $\theta_3^{(i)} m^{(i)}(a - a^*)$ to the controlled direct effect and adding $\theta_3^{(i)}\beta_1^{(i)} a(a - a^*)$ to the natural indirect effect; however, the expression for the natural direct effect is more complicated because it involves the correlation between the mediators; it is given in the Appendix.

If the data come from a case–control study with a rare outcome, then this same approach and the same expressions can be used but the regression models for the mediators are then fit only among the controls as described in Chapter 2. The same expressions can likewise be used for the ratio of expectations if the outcome is a count outcome and the logistic regression model is replaced by a log-linear model.

The approach described here for a binary or count outcomes will, as noted above, only apply if all of the mediators are continuous. If one or more of the mediators are binary, then an alternative approach will need to be used. Moreover, as we have seen, even if all of the mediators are continuous, the expressions for the natural direct effect become more complicated if there are exposure–mediator interactions. All of this motivates the alternative weighting approach, described below in Section 5.2, which can much more flexibly accommodate binary outcomes.

5.1.5. Assessing Mediators Sequentially

Before moving on, a few additional comments merit attention. First, the approach we have described so far does not necessarily require knowing the ordering of the mediators $M^{(1)}, \ldots, M^{(K)}$, though it does again require that there is no further variable that is affected by the exposure and that goes on to affect one of the mediators $M^{(1)}, \ldots, M^{(K)}$ and also the outcome. If there is such a variable, it needs to be included in the set $M^{(1)}, \ldots, M^{(K)}$. If the ordering of $M^{(1)}, \ldots, M^{(K)}$ is known, then some further progress can also be made. One could, for example, begin with the first mediator $M^{(1)}$ and use the approach described here to examine the portion of the effect mediated through $M^{(1)}$. One could then consider $M^{(1)}$ and $M^{(2)}$ jointly and use the approach described here to examine what proportion of the

effect is mediated through both $M^{(1)}$ and $M^{(2)}$ together. Doing so would allow one to assess the additional contribution of $M^{(2)}$ beyond $M^{(1)}$ alone. Note that the difference between the two will potentially be different than simply the effect mediated through $M^{(2)}$ itself because, for example, $M^{(1)}$ and $M^{(2)}$ may share common pathways (if, for example, $M^{(1)}$ affected $M^{(2)}$ or if, as discussed later in the paper, $M^{(1)}$ and $M^{(2)}$ interact in their effects). One could further then consider $(M^{(1)}, M^{(2)}, M^{(3)})$ and examine the proportion mediated by all three jointly along with the additional contribution of $M^{(3)}$ beyond $(M^{(1)}, M^{(2)})$. One could carry on this process, adding sequentially one mediator at a time, until all K mediators are included.

Undertaking this sequential approach does, however, place additional restrictions on the models being used. This is because for each group of mediators a different model is being fit for Y. In particular, as is discussed at greater length in the Appendix, for the various models for Y to be compatible with one another, it necessary that the exposure is binary and either (i) there are no exposure–mediator or mediator–mediator interactions or (ii) the models must be extended to allow for exposure–covariate interaction. See VanderWeele and Vansteelandt (2013) for further details.

5.1.6. Some Further Properties: Robustness to Mediator Confounding and Joint Versus Summed Proportion Mediated

As noted above, when multiple mediators are of interest, the approach of considering mediators one at a time will only be appropriate if the mediators do not affect one another. If one of the mediators of interest affects another, then assumption (A5.4) will be violated for one or more mediators. The approach we have described in this section can, however, still be used. The approach described here has other advantages, even if the mediators do not affect one another. Suppose, for example, the mediators do not affect each other but there is an unmeasured common cause U of two or more mediators as in Figure 5.2. In this case, the approach of considering the mediators $M^{(1)}$ and $M^{(2)}$ one-by-one will be biased because when $M^{(1)}$ alone is considered, U will be an unmeasured confounder for the effect of $M^{(1)}$ on Y (it affects Y through $M^{(2)}$) and when $M^{(2)}$ alone is considered, U will be

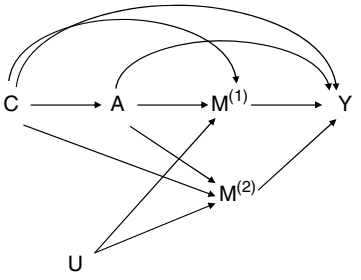

Figure 5.2 Two mediators with an unmeasured common cause.

an unmeasured confounder of the effect of $M^{(2)}$ on Y (it affects Y through $M^{(1)}$). However, when $M^{(1)}$ and $M^{(2)}$ are considered jointly as we have been doing, U no longer serves as a confounder for the joint effect of $(M^{(1)}, M^{(2)})$ on Y because U no longer affects Y except through $(M^{(1)}, M^{(2)})$.

When the mediators affect one another, then, as discussed above, we generally cannot estimate the natural direct and indirect effects for one or more mediators by just considering them one at a time. Even if we could, the sum of the proportion mediated can be more than 100%, even if all pathways affect the outcome in the same direction. This can occur because certain paths may be counted twice. In Figure 5.1, if the analysis were done one mediator at a time, then the path $A - M^{(1)} - M^{(2)} - Y$ would be included in the indirect effect both for the analysis for $M^{(1)}$ and for the analysis for $M^{(2)}$. The approach described above circumvents this difficulty. However, we might then think that if the mediators do not affect one another—if, for example, the mediators are independent of one another conditional on A and C—then the sum of two mediated effects considered separately should equal the joint mediated effect when both mediators are considered together. In fact, even if the mediators are statistically independent and do not affect each other, this need not hold. The sum of two mediated effects considered separately may diverge from the joint mediated effect when there are interactions between the effects of the two mediators on the outcome. Note that such interaction can arise even if the mediators do not affect each other. We show in the Appendix that if there is no interaction in the effects of the two mediators at the individual counterfactual level and if the two mediators do not affect each other, then the sum of two mediated effects considered separately will equal the joint mediated effect when both are considered together. If the two diverge, then either the mediators must affect one another or there must be an interaction between the effects of the two mediators.

In some applications, mediators are considered one at a time and the proportion mediated is calculated for each of these. Sometimes, when doing this, the sum of the proportion mediated can exceed 100%. One possible explanation for this is that there are other pathways (that operate through other mediators) that affect the outcomes in the opposite direction from those under consideration. The sum of the proportion mediated may exceed 100% if there are other mediators with a "negative" proportion mediated. If this is thought not to be the case—that is, if all pathways are thought to operate on the outcome in the same direction—then the true proportion mediated for all mediators (known and unknown) considered jointly must be 100%. If the sum of the proportion mediated when each measured mediator is considered separately exceeds 100%, then the sum and the joint proportion mediated would be different and thus it must be the case either that the mediators affect one another or that there are interactions between the effects of the mediators on the outcome. The approach described above could accommodate these complications by considering all mediators jointly while the approach of assessing mediators one at a time cannot. In summary, if the sum of the proportion mediated exceeds 100%, then one of the following must be true: (i) There are other mediators with a negative proportion mediated, (ii) the mediators affect one another, or (iii) there are interactions between the mediators.

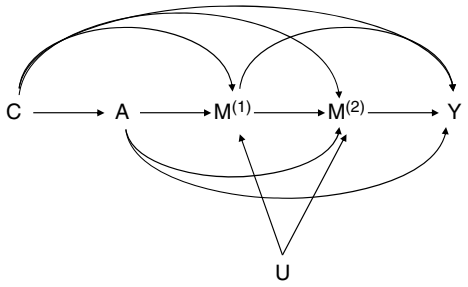

Figure 5.3 Two mediators in which one affects the other and they share an unmeasured common cause U.

Comparing the sum of the mediated effects to a joint mediated effect (or examining if the sum of the proportion mediated exceeds 100% if all mediators are thought to operate in the same direction) would thus constitute one strategy whereby an investigator could assess whether the approach of examining one mediator at a time might fail. An alternative approach might consist of examining the independence of the mediators more directly. For example, in the case of two mediators, $M^{(1)}$ and $M^{(2)}$, a regression of $M^{(2)}$ on $M^{(1)}, A, C$ should have $M^{(1)}$ independent of $M^{(2)}$ in the regression. Statistical dependence between the two conditional on A and C would indicate that the approach of examining mediators one at a time cannot be used. Statistical dependence between $M^{(1)}$ and $M^{(2)}$ conditional on A and C cannot distinguish between Figures 5.1 and 5.2 (or Figure 5.3, in which one mediator affects the other and they share an unmeasured common cause), but in either case, the approach of examining the mediators one at a time fails because assumptions required for such an approach are then violated.

5.2. A WEIGHTING APPROACH TO MULTIPLE MEDIATORS

5.2.1. Weighting for Multiple Mediators

Because of the aforementioned concerns about model incompatibility in models that involve mediator–mediator interactions and because the addition of mediators increases the need for modeling, we also present a simple alternative approach based on inverse probability weighting (VanderWeele and Vansteelandt, 2013). This alternative weighting approach does not require models for the mediators. Instead a model for the exposure is used and this then essentially overcomes the issue of model incompatibility. This weighting approach can be used for essentially any type of outcome, including non-rare binary outcomes, and regardless of whether there are mediator–mediator interactions. However, as with other weighting approaches, its performance is best when the exposure is binary or discrete with only a few levels. This approach essentially forms a straightforward generalization of the results in Albert (2012) to the case of a vector of mediators and is closely related to the imputation approach of Vansteelandt et al. (2012a).

The weighting approach estimates the marginal natural direct effect and the marginal natural indirect effect—that is, averaged over the covariates. Doing this requires the estimation of three weighted averages that we will call Q_1, Q_2, and Q_3. These weighted averages under assumptions (A5.1)–(A5.4) will correspond to the counterfactuals $\mathbb{E}[Y_{a^*\mathbf{M}_{a^*}}]$, $\mathbb{E}[Y_{a\mathbf{M}_a}]$, and $\mathbb{E}[Y_{a\mathbf{M}_{a^*}}]$, respectively. For Q_1 we take a weighted average of the subjects with $A = a^*$ where each subject i is given a weight

$$\frac{P(A = a^*)}{P(A = a^*|C = c_i)}$$

where c_i denotes the actual covariate value for subject i. For a binary exposure with $a = 1$ and $a^* = 0$, the probabilities $P(A = 0|C = c_i)$ could be fit, for example, using logistic regression to obtain the predicted probabilities for $P(A = 0|C = c_i)$ for each subject i. Likewise for the second weighted average Q_2 we proceed by taking a weighted average of the subjects with $A = a$ where each subject i is given a weight

$$\frac{P(A = a)}{P(A = a|C = c_i)}$$

where again c_i denotes the actual covariate value for subject i. And again for a binary exposure with $a = 1$ and $a^* = 0$, the probabilities $P(A = 1|C = c_i)$ could be fit, for example, using logistic regression to obtain the predicted probabilities for $P(A = 1|C = c_i)$ for each subject i.

Finally, for the weighted average Q_3, for each subject i with $A_i = a^*$, one uses an outcome model $\mathbb{E}(Y|A = a, \mathbf{M} = \mathbf{m}_i, C = c_i)$ (which can include exposure–mediator or mediator–mediator interactions) to obtain a predicted estimate of the outcome if the individual had had exposure $A_i = a$ rather than $A_i = a^*$, but using the individual's own values of the mediator, $\mathbf{M} = \mathbf{m}_i$, and covariates, $C = c_i$. Once these predicted values are calculated, one can obtain an estimate for the counterfactual $\mathbb{E}[Y_{a\mathbf{M}_{a^*}}]$ by taking a weighted average of these predicted values for subjects with $A_i = a^*$ where each subject i is again given the weight

$$\frac{P(A = a^*)}{P(A = a^*|C = c_i)}$$

Once the various weighted averages are obtained, we can estimate the natural direct effect by taking the difference $Q_3 - Q_1$, and we can estimate the natural indirect effect by taking the difference $Q_2 - Q_3$. Alternatively, on a ratio scale we could obtain risk ratios for the natural direct effect by taking the ratio Q_3/Q_1, and we called obtain risk ratios for the natural indirect effect by taking the ratio Q_2/Q_3—and likewise for effects on an odds ratio scale. We recommend bootstrapping for confidence intervals. An alternative weighting approach for multiple mediators, similar to that described in Section 4.5, was proposed by Lange et al. (2014) but requires that none of the mediators affect one another, which is a strong restriction. The weighting approach presented in this section (VanderWeele and Vansteelandt, 2013) does not impose this restriction.

5.2.2. SAS Code for the Weighting Approach for Multiple Mediators

We describe how the proposed weighting approach given above can be implemented in SAS statistical software (SAS Institute, Inc., Cary, North Carolina). Below we let `c`, `a`, `m` and `y` correspond to the observed confounders C, exposure A, mediator M, and outcome Y and assume, for the illustration, that A and Y are dichotomous.

```
proc logistic data = mydata;
model a = c;
score data = mydata out = preda;
run;
data preda;
set preda;
pa1 = P_1;
run;
data mydata0;
set mydata;
a = 0; output;
run;
data mydata1;
set mydata;
a = 1; output;
run;
proc logistic data = mydata;
model y = a m c;
score data = mydata0 out = predy0;
score data = mydata1 out = predy1;
run;
data predy0;
set predy0;
py0 = P_1;
run;
data predy1;
set predy1;
py1 = P_1;
run;
data mydataw;
merge preda predy0 predy1 mydata;
run;
data mydataw;
set mydataw;
w = a/pa1+(1-a)/(1-pa1);
run;
```

The mean $\mathbb{E}[Y_{1\mathbf{M}_0}]$ (except for standard errors) can now be estimated using:

```
proc reg data = mydataw;
where a = 0;
model py1 = ;
weight w;
run;
```

and the mean $\mathbb{E}[Y_{0\mathbf{M}_1}]$ (except for standard errors) using:

```
proc reg data = mydataw;
where a = 1;
model py0 = ;
weight w;
run;
```

5.2.3. Illustration of the Weighting Approach for Multiple Mediators

To illustrate the weighting approach, we will analyze 2003 US birth certificate data and will consider whether the exposure, A, of adequate or inadequate prenatal care ($n = 2,629,247$; those with intermediate or superadequate care are excluded from the analysis for the purposes of this illustration) on preterm birth (Y) is mediated by maternal smoking and/or drinking ($M^{(1)}$) or pre-eclampsia ($M^{(2)}$). Adequacy of prenatal care categories are determined from data on the month prenatal care was initiated, on the number of visits, and on gestational age, according to the American College of Gynecologists recommendation as encoded in a modification of the APNCU index (Kotelchuck, 1994; VanderWeele et al., 2009). In this analysis we will take age category (below 20 years, between 20 and 35 years, or above 35 years), ethnicity (black, Hispanic, native American, white), education, and marital status as baseline confounders (C). Our analysis is certainly a simplification of a more complex reality because prenatal care and maternal smoking are both ultimately time-varying and pre-eclampsia and preterm birth could be conceived of as processes whereas we will treat them as dichotomous.

Inverse probability weights were constructed on the basis of logistic regression models for adequate care. In view of the large sample size and the resulting computational complexity, standard errors and confidence intervals were constructed using the subsampling bootstrap (Politis and Romano, 1994). This is similar to the bootstrap, but involved repeating the analysis for 1000 subsamples of size $n = 13,146$ (0.5% of the total sample size); on the basis of the empirical standard deviation of the 1000 estimates, the standard error of the estimates that were obtained from the analysis of the full data set can be inferred (accounting for correlation resulting from the fact that some data points may be shared between subsamples).

We first consider maternal smoking and/or drinking as mediators. The analyses indicate the direct effect of adequate care, through pathways other maternal smoking and/or drinking ($M^{(1)}$), is a 5.6% (95% CI 5.5% to 5.7%) reduction in the risk of preterm birth and that the mediated effect via maternal smoking and/or drinking is a 0.09% (95% CI 0.08% to 0.10%) reduction in the risk of preterm birth. When pre-eclampsia is considered as an additional mediator and the weighting approach is applied ($M^{(2)}$), we find essentially the same direct and indirect effects of 5.6% (95% CI 5.5% to 5.7%) and 0.09% (95% CI 0.08% to 0.10%). The effect of adequate prenatal care on preterm birth by pathways through pre-eclampsia, but not through maternal smoking and drinking, thus seems minimal.

5.3. CONTROLLED DIRECT EFFECTS AND EXPOSURE-INDUCED CONFOUNDING

In the last two sections we have assumed that what is of interest is the effect of an exposure on some outcome through several mediators jointly. In the next three sections we will turn to a somewhat different set of questions that may arise when interest lies in one particular mediator but when there are other prior mediators affected by the exposure which in turn affect both the mediator of interest and the outcome. This was a violation of our fourth assumption about confounding that there was no mediator–outcome confounder affected by the exposure (assumption A2.4). In this section we will discuss the estimation of controlled direct effects in the presence of such exposure-induced mediator–outcome confounding. In the following section we will discuss alternative approaches to natural direct and indirect effects that can be used in the presence of exposure-induced mediator–outcome confounders even though natural direct and indirect effects themselves are not identified.

5.3.1. Difficulties with Regression Methods in the Presence of Exposure-Induced Confounding

Suppose we are in a setting like that of Figure 5.4 and we are interested in the effect of A on Y as mediated by or independent of M. In this figure, L here is a mediator–outcome confounder affected by the exposure. Note that this is essentially the same figure as Figure 5.1 except that we are now focusing our attention on a single mediator M rather than L and M together. In this setting, L is a mediator–outcome confounder affected by the exposure, and as discussed in Chapter 2, natural direct and indirect effects are not identified in this setting. Suppose then that instead we are interested in estimating the controlled direct effect of A on Y through pathways that do not involve M. We thus want to estimate the effects of two pathways A–Y and A–L–Y. These are the pathways not through M. Controlled direct effects are identified even in the presence of an exposure-induced mediator–outcome confounder, but in this setting the regression methods of Chapter 2 will not work. Regression models will not work here because L is both a mediator–outcome confounder and on the pathway from the exposure to the outcome. To intuitively see why regression control for L will not work, suppose first that we included L as a

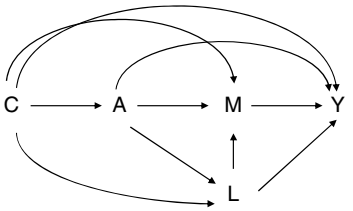

Figure 5.4 An example of a mediator-outcome confounder L that is affected by the exposure A.

covariate in our outcome regression. In this case we would potentially be blocking one of the direct effect pathways not through M that we were interested in, namely A–L–Y, by controlling for L. This would suggest that we should perhaps not adjust for L in the regression. But if we do not adjust for L in the regression, then our estimates of the effect of M on Y will be confounded (since L is a confounder of the M–Y relationship which we have not controlled for) and thus our direct and indirect effect estimates will be biased. It thus seems that whether we control for L or not in the regression, we will get bias if what we are interested in is the direct effect of A on Y not through M (but potentially through L). We get biased results if we adjust for L; we get biased results if we do not adjust for L. Regression methods cannot be used to estimate controlled direct effects in this setting.

Instead, in order to estimate controlled direct effects, we need a different class of methods. In this section we will discuss the use of two different methods to estimate controlled direct effects in this setting: marginal structural models (Robins et al., 2000; van der Laan and Petersen, 2008; VanderWeele, 2009) and structural mean models (Robins, 1999a; Vansteelandt, 2009a). These methods were developed originally to handle time-varying exposures outside the context of mediation (Robins, 1999b; Robins et al., 2000), but they are applicable to contexts of mediation as well. With an exposure-induced mediator-outcome confounder we modify the second of our two confounding assumptions. We will assume, as before, that [assumption (A2.1)] the effect of A on Y is unconfounded conditional on C, but now we will also assume [assumption (A5.5)] that the effect of M on Y is unconfounded conditional on (C,A,L). Note that our first assumption [i.e., (A2.1)] is the same assumption about exposure–outcome confounding that we made in Chapter 2. However, now instead of assuming that controlling for the covariates C suffices to control for mediator–outcome confounding, we are assuming that control for mediator–outcome confounding also requires that control be made for L, a variable that itself may be affected by the exposure. These two assumptions would be satisfied in Figure 5.4. They would also be satisfied even if there were an unmeasured confounder of L and Y as in Figure 5.5 since (C,A,L) still suffice to control for confounding for the effect of M on Y in this diagram. In any case, if these two assumptions do hold, then we can proceed with the methods described below.

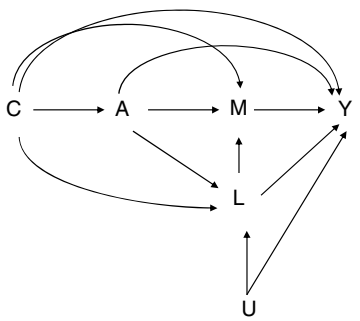

Figure 5.5 An example of a mediator-outcome confounder L that is affected by the exposure A, with an unmeasured common cause U of L and outcome Y.

5.3.2. Marginal Structural Models for Controlled Direct Effects in the Presence of Exposure-Induced Confounding

Under the assumptions [assumption (A2.1)] that the measured covariates C suffice to control for exposure–outcome confounding and [assumption (A5.5)] that C, A, and L together suffice to control for mediator–outcome confounding, we can estimate controlled direct effects using a method referred to as marginal structural models (Robins, 1999b; Robins et al., 2000). Such models are models for the average counterfactual outcomes themselves. The models are not fit directly by regression but instead by a weighting technique generally referred to as "inverse probability weighting." The basic analytic approach is to fit a regression model of Y on A and M but instead of controlling for the confounders C and L by regression, these are controlled for by weighting. Two sets of weights are needed, one for the exposure and one for the mediator. For each individual i in the sample, the exposure weight is calculated by

$$w_i^A = \frac{P(A = a_i)}{P(A = a_i | C = c_i)}$$

where a_i and c_i are the actual values of the exposure and the covariates for individual i. In the denominator of the weight we have the probability of receiving the treatment the individual in fact received conditional on the covariates C taking the value c_i; that is, this is the predicted probability of an individual actually having the treatment that they had, given they had their covariate values. For a binary exposure, this probability could be estimated, for example, by using a logistic regression model (e.g., see code below). The denominator of the weight is similar to what is sometimes called a propensity score (Rosenbaum and Rubin, 1983b), the probability of the exposure conditional on the covariates. For the exposed subjects, we use the predicted propensity score in the denominator. For the unexposed subjects, we use one minus the predicted propensity score. In other words, for each subject we use the predicted probability, conditional on a subject's covariates, of their having the exposure that they in fact had. Because this probability is in the denominator, the weighting approach being described here is sometimes referred to as inverse probability weighting. The numerator of the weight for the exposure is just the overall proportion of those with the exposure in the population. The ratio of these two is our weight for the exposure. To use the marginal structural model approach, we also need a second weight, a weight for the mediator. For each individual i in the sample, the mediator weight is calculated by

$$w_i^M = \frac{P(M = m_i | A = a_i)}{P(M = m_i | A = a_i, C = c_i, L = l_i)}$$

where m_i and l_i are likewise the actual values of the mediator M and the variable L for individual i. The denominator of this weight w_i^M is the probability of having the value of the mediator that the individual in fact had conditional on $A = a_i$, $C = c_i$, and $L = l_i$. For a binary mediator, this predicted probability could likewise be obtained by a logistic regression. The numerator probability is simply the probability of having the value of the mediator that the individual in fact had conditional

on having the exposure they in fact had. Once again for a binary mediator, this could be obtained by logistic regression and is somewhat analogous to a propensity score for the mediator.

Once these weights are estimated, we can obtain an overall weight for the individual by taking a product of these weights:

$$w_i = w_i^A \times w_i^M$$

We can then employ the marginal structural model approach to controlled direct effects (van der Laan and Petersen, 2008; VanderWeele, 2009a) by fitting a regression model of Y on A and M, potentially allowing for exposure–mediator interaction

$$\mathbb{E}[Y|a,m] = \gamma_0 + \gamma_1 a + \gamma_2 m + \gamma_3 am$$

but where each individual i is weighted by the overall weight we calculate for that individual $w_i = w_i^A \times w_i^M$. The controlled direct effect can then be obtained based on the resulting estimates from this final weighted regression. The controlled direct effect of A on Y with the mediator fixed to level $M = m$ for a change in exposure from a^* to a is given by

$$CDE(m) = \gamma_1(a - a^*) + \gamma_3 m(a - a^*) \tag{5.2}$$

This allows us to estimate the controlled direct effect even in the presence of an exposure-induced mediator–outcome confounder. If the exposure is binary so that $a = 1$ and $a^* = 0$, then this just reduces to

$$CDE(m) = \gamma_1 + \gamma_3 m$$

Several comments merit additional attention. First, as noted above, control for the covariates is done by weighting rather than regression. The covariates are not included in the regression model that we fit for Y. We regress Y only on A and M, but we weight this regression. Also, different sets of covariates are included in the denominator of the weight for the exposure versus the mediator. For the denominator of the exposure we estimate the probability of the exposure only conditional on the baseline covariates C. For the denominator of the mediator weight we estimate the probability of the mediator conditional on the baseline covariates C, the exposure A, and the exposure-induced confounder L. Thus L is included in the weight for the mediator but not in the weight for the exposure. It is this that essentially allows us to get around the problem with regression methods described in the previous subsection (that we get bias whether or not we control for L). By controlling for L only in the mediator weight and not the exposure weight, we get around this problem. This approach also works even if there are several exposure-induced mediator outcome confounders. We simply need to include them all in estimating our denominator probabilities for the mediator weights.

Second, the approach described above will allow us to estimate controlled direct effects. However, if we also want to obtain standard errors, then in the weighted regression we must calculate what are sometimes called "robust" or "sandwich"

standard errors. This then takes into account in the estimation of standard errors the weighting approach that has been used. We give code for this below.

Third, if the exposure and mediator are binary, then the predicted probabilities in the denominator and numerator for the weights for A and M could be fit by logistic regression (again code is given below). If one or both of the exposure or mediator are categorical or ordinal, then these numerator and denominator probabilities could be estimated by using a multinomial or ordinal logistic regression. If the exposure or mediator is continuous, the probabilities can be replaced by values from a probability density function; see Robins et al. (2000) for further description. However, the performance of this marginal structural model technique is usually not particularly good if the exposure and mediator are continuous. We describe an alternative technique in Section 5.3.5 which can work well with continuous exposure and mediator and can likewise estimate controlled direct effects in the presence of exposure-induced mediator–outcome confounding.

Fourth, if the outcome Y were binary, we could potentially still proceed with the approach described above to obtain controlled direct effects on the risk difference scale. However, if we wanted to obtained controlled direct effects on the odds ratio scale instead, we could replace the linear regression model for Y conditional on A and M with a logistic regression model of Y on A and M:

$$\text{logit}\{P(Y=1|a,m)\} = \gamma_0 + \gamma_1 a + \gamma_2 m + \gamma_3 am$$

where we once again weight each individual by the weights $w_i = w_i^A \times w_i^M$ above. The controlled direct effect odds ratio with the mediator fixed to level $M = m$ for a change in exposure from a^* to a is given by $OR^{CDE}(m) = \exp\{\gamma_1(a - a^*) + \gamma_3 m(a - a^*)\}$.

Finally, the approach described here could also be employed even if there were no exposure-induced mediator–outcome confounding. In other words, if there were no variables L, this would then be an alternative approach to the regression-based approaches described in Chapter 2. However, in the absence of exposure-induced mediator–outcome confounding, the regression-based approach tends to be more efficient (i.e., have smaller standard errors) and so it is recommended unless exposure-induced mediator–outcome confounding is indeed present.

5.3.3. SAS Code for Marginal Structural Models for Controlled Direct Effects

In this section we give SAS code to implement this marginal structural model approach to direct effects. This code could fairly easily be adapted to other software packages. Suppose the exposure A and mediator M are binary and suppose that the name of the dataset was "mydata," the name of the exposure variable "a," the mediator variable "m," the outcome "y," and exposure-induced mediator–outcome confounder "l" and that we have five baseline covariates "c1," "c2," "c3," "c4", and "c5." We could then first fit the models for the weights using the following code:

```
proc logistic data=mydata descending;
   model a= ;
```

```
   output out=mydata predicted=pn1;
run;

proc logistic data=mydata descending;
   model a= c1 c2 c3 c4 c5;
   output out=mydata predicted=pd1;
run;

proc logistic data=mydata descending;
   model m = a;
   output out=mydata predicted=pn2;
run;

proc logistic data=mydata descending;
   model m = c1 c2 c3 c4 c5 a l;
   output out=mydata predicted=pd2;
run;
```

The first two logistic regressions are for the exposure weights—the numerator and the denominator, respectively. The second two logistic regressions are for the mediator weights—the numerator and denominator, respectively. In each case the "predicted" command will predict the probability conditional on the individuals' actual covariate values, that the exposure (for the first two regressions) or the mediator (for the second two regressions) takes the value 1.

To obtain the weights, we need to take these predicted probabilities for the exposure and mediator being 1 and turn them into the probabilities of the individual having the values of the exposure and mediator that they in fact had. In other words, for the exposure numerator and denominator probabilities, if the individual did actually have the exposure, then we can just take the predicted probabilities themselves. If the individual did not have the exposure, then we want 1 minus these predicted probabilities (i.e., the predicted probability that the individual did not have the exposure). Likewise for the mediator numerator and denominator probabilities, if the individual actually had a value of the mediator $M = 1$, then we can just take the predicted probabilities themselves. If the individual had a value of the mediator $M = 0$, then we want 1 minus these predicted probabilities (i.e., the predicted probability that the individual had mediator level $M = 0$). The code below thus calculates the appropriate predicted probabilities and takes the ratio of the numerator and denominator probabilities for the exposure to form the exposure weights (w1) and takes the ratio of the numerator and denominator probabilities for the mediator to form the mediator weights (w2) and then takes the product of these weights to form the overall weights (ww). It also forms a new variable am_int that is the exposure–mediator interaction which will be used in the final marginal structural model.

```
data mydata;
   set mydata;
   if a=1 then w1=pn1/pd1; else w1=(1-pn1)/(1-pd1);
   if m=1 then w2=pn2/pd2; else w2=(1-pn2)/(1-pd2);
   ww=w1*w2;
   am_int = a*m;
run;
```

Finally, once we have calculated the weights, we can simply regress the outcome "y" on the exposure "a," the mediator "m," and the exposure–mediator interaction variable "am_int" where we weight each individual by the weights we have calculated. The final line of the procedure requests that "robust" or "sandwich" standard errors be calculated, which is necessary to obtain valid standard errors to take into account the weighting.

```
proc genmod data=mydata;
   class caseid;
   model y = a m am_int / error=normal link=id;
   weight ww;
   repeated subject = caseid/ type = unstr;
run;
```

This will then give us the parameters of the marginal structural model that we can use to calculate the controlled direct effect. If the outcome were binary and we wanted to obtain a controlled direct effect odds ratio, we could simply replace "error=normal link=id" with "error=binomial link=logistic."

The code above applies to binary exposures and mediators. If the exposure or mediator were categorical or ordinal, we would have to replace the logistic regressions for the predicted probabilities by multinomial or ordinal logistic regressions and obtain for each individual the predicted probabilities of their having the exposure or mediator level that was in fact present. If the exposure A or the mediator M is continuous, then we have to replace the probabilities for the weights from logistic regression with probability density functions obtained from linear regression (cf. Robins et al., 2000), but again this marginal structural model for controlled direct effects approach will in general be somewhat less stable for continuous exposures and mediators, and the approach we describe below in section 5.3.5 using structural mean models is then to be preferred.

5.3.4. An Example of Controlled Direct Effects using MSMs: Life Course Epidemiology

Analyses have repeatedly found that low socioeconomic status (SES) during childhood is associated with adverse health outcomes later in life. However, there remains debate as to whether this is because low SES during childhood affects adult SES, which in turn affects adult health (a "social trajectory" model), or whether childhood SES affects adult health through pathways other than through adult SES (a "latent effects/sensitive period" model), or both. Here we would then take childhood SES as the exposure A, adult SES as the mediator M, and some health outcome, say diabetes, as the outcome Y. One complicating factor in trying to address this question is that often there is a considerable temporal gap between the childhood and adult measures of SES. There may thus be factors L on the pathway from childhood SES to the adult SES measurement that also affect the outcome Y as in Figure 5.4. For example, risk behaviors prior to adult SES measurement might affect both adult SES M and the health outcome Y.

Nandi et al. (2012) used data from a national sample of 9760 persons from the Health and Retirement Study. Participants were between ages 50 and 62 during the baseline year 1992 and were followed up every two years subsequently. A composite measure of early-life SES was obtained via confirmatory factor analysis using parental education, father's occupation, region of birth, and childhood rural residence, all assessed retrospectively in 1992. A composite measure of adult SES in 1992 was obtained via confirmatory factor analysis using education, occupation, labor force status, household income, and household wealth. Any instance of self-reported diabetes, as assessed every two years through 2006, was used as an outcome in the analysis. Baseline confounders C for which adjustment was made included baseline age, sex, race, and self-rated childhood health. Adult risk factors L that potentially confounded the relation between adult SES and the diabetes outcome included high blood pressure, body mass index, and self-rated health and whether health problems limited the respondent's ability to work, current smoking, and alcohol consumption.

Each of the childhood and adult SES variables was categorized into quartiles for the purposes of the analysis and the estimation of weights. The marginal structural model approach described in Section 5.3.2 was used to estimate the controlled direct effect of early SES on diabetes not through adult SES, using weighing to appropriately adjust for adult risk factors L. Note that the controlled direct effect includes the pathway through adult risk factors, A–L–Y, so we cannot simply include L in a regression model. We must proceed by using the marginal structural model weighting approach instead. Using a marginal structural logistic regression model, the controlled direct effect odds ratio comparing the highest SES in childhood to increasingly lower SES in childhood were 1.06 (95% CI: 0.91, 1.23), 1.15 (95% CI: 0.97, 1.36), 1.23 (95% CI: 1.02, 1.48). There seems to be some evidence for a direct effect of low childhood SES on adult diabetes not through adult SES, using this approach.

In analyses that replicated what might be done with conventional regression, the corresponding effects using regression to control for the baseline confounders but without controlling for the adult risk factors L gave odds ratios of 1.02 (95% CI: 0.87, 1.19), 1.07 (95% CI: 0.90, 1.27), and 1.17 (95% CI: 0.98, 1.41). When regression was used but control was also made for the adult risk factors L, this gave odds ratios of 0.99 (95% CI: 0.83, 1.17), 0.99 (95% CI: 0.82, 1.19), and 1.08 (95% CI: 0.88, 1.32). In both of these alternative analyses, the direct effect odds estimates were smaller than when using the marginal structural model and all confidence intervals extended below zero. Note, however, that if the adult risk factors truly do constitute a mediator–outcome confounder affected by the exposure as in Figure 5.4, then both of the traditional regression analyses, with or without control for L, will be biased. The regression analysis not adjusting for L will be biased because L confounds the relationship between the adult SES mediator M and the diabetes outcome Y. The regression analysis adjusting for L will be biased because it will block part of the effect of the childhood SES exposure A on the diabetes outcome Y not through adult SES M; that is, it will block the path A–L–Y. A structural equation model could have been used but this would have required modeling the relationships between A and L, between L and M, between L and Y, and among

the variables in L itself; the marginal structural model only required specifying the relationship between L and M.

Of course, even the marginal structural model approach is a simplification of a more complex reality. There is likely feedback back and forth between the adult risk factors themselves and SES. Ideally, then multiple measures of the adult risk factors, and adult SES, over time would be collected and appropriately used in the analysis. Methods to do so will be the focus in Chapter 6. However, the analysis here arguably represents a step in the right direction, compared to either of the potential regression approaches, in addressing confounding for this research question in social epidemiology.

5.3.5. Structural Mean Models for Controlled Direct Effects in the Presence of Exposure-Induced Confounding

Here we describe another approach to estimate controlled direct effects in the presence of exposure-induced mediator–outcome confounding. When the exposure and/or mediator is continuous, this approach will often be more stable than the marginal structural model approach described in Sections 5.3.2 and 5.3.3. We assume again our assumptions [assumption (A2.1)] that the measured covariates C suffice to control for exposure–outcome confounding and [assumption (A5.5)] that C, A, and L together suffice to control for mediator–outcome confounding. The structural mean model will be a model for the controlled direct effect that we can fit by a two-stage regression approach (Vansteelandt, 2009; Joffe and Greene, 2009; cf. Robins, 1999a). In the first stage we fit a regression model of Y on A, M, C, and L, allowing for exposure–mediator interaction:

$$\mathbb{E}[Y|a,m,l,c] = \kappa_0 + \kappa_1 a + \kappa_2 m + \kappa_3 am + \kappa_4' c + \kappa_5' l$$

Once fit, we do two things with the coefficient estimates of this regression model. First, we save the estimate of regression coefficient for the exposure–mediator interaction, κ_3, because this will form part of our controlled direct effect estimate. Second, we take the coefficient estimates (call them $\widehat{\kappa}_2$ and $\widehat{\kappa}_3$) of κ_2 and κ_3 and for each individual i we calculate the following outcome residuals:

$$\widehat{Y} = Y - \widehat{\kappa}_2 m - \widehat{\kappa}_3 am$$

We then regress these residuals on the exposure A and covariates C:

$$\mathbb{E}[\widehat{Y}|a,c] = \gamma_0 + \gamma_1 a + \gamma_2' c$$

Once we have obtained a coefficient estimate $\widehat{\gamma}_1$ of γ_1 from our regression of $\widehat{Y}$ on A, our estimate of the controlled direct effect of A on Y with the mediator fixed to level $M = m$ for a change in exposure from a^* to a is given by

$$CDE(m) = \widehat{\gamma}_1(a - a^*) + \widehat{\kappa}_3 m(a - a^*) \tag{5.3}$$

Standard errors can be obtained by bootstrapping. See Vansteelandt (2009) for R code to implement this approach.

5.4. EFFECT DECOMPOSITION WITH EXPOSURE-INDUCED CONFOUNDING

The methods in the previous section allow for estimation of controlled direct effects even in the presence of exposure-induced mediator–outcome confounding. However, this does not allow for effect decomposition. We have already noted that if there is a mediator–outcome confounder that is affected by the exposure, then natural direct and indirect effects are not identified. Although these natural direct and indirect effects are not identified, certain analogues of these effects based on randomized interventions on the mediator will be identified even in the presence of such exposure-induced mediator outcome confounding.

5.4.1. Randomized Interventional Analogues of Natural Direct and Indirect Effects

With natural direct and indirect effects we would consider, for example, what would happen if, for an exposed person we fixed the mediator to the level it would have been for that person if they had been unexposed. We will instead now consider somewhat different effects wherein for an exposed person we will consider what would have happened if we fixed their mediator to a level that is drawn randomly from the subpopulation that is unexposed. Thus instead of using the individual's particular value for the mediator in the absence of exposure, we use the distribution of the mediator amongst all the unexposed. It turns out that this slight change is sufficient to identify these randomized interventional analogues of the natural direct and indirect effects even in the presence of exposure-induced mediator–outcome confounding.

Let $G_{a|c}$ denote a random draw from the distribution of the mediator amongst those with exposure status a conditional on $C = c$. Suppose a and a^* are two values of the exposure we wish to compare; for example, for a binary exposure we may have $a = 1$ for exposed and $a^* = 0$ for unexposed. Consider first the effect on the outcome of randomly assigning an individual who is given the exposure to a value of the mediator from the distribution of the mediator amongst those with exposure versus without exposure (conditional on covariates); this is an effect through the mediator. We will refer to this effect as NIE^R, the randomized interventional analogue of the natural indirect effect. In counterfactual notation, this would be $\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})$. Next consider the effect comparing exposure versus no exposure with the mediator in both cases randomly drawn from the distribution of the population when given no exposure (given covariates). We will refer to this effect as NDE^R, the randomized interventional analogue of the natural direct effect. In counterfactual notation, this is $\mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|c}})$. Finally, consider the effect comparing the expected outcome when having the exposure with the mediator randomly drawn from the distribution of the population when given the exposure (conditional on covariates) to the expected outcome when not having the exposure with the mediator randomly drawn from the distribution of the population when not exposed (conditional on covariates). We will refer to this effect

as TE^R, the randomized interventional analogue of the total effect. In counterfactual notation this is $\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{a^*G_{a^*|c}})$. We can decompose this total effect into the sum of the randomized interventional analogues of the direct and indirect effects, $TE^R = NIE^R + NDE^R$, so that the overall effect decomposes into the sum of the effect through the mediator and the direct effect. These are not the natural direct and indirect effects considered earlier but are instead analogues arising not from fixing the mediator for each individual to the level it would have been under a particular exposure, but rather from fixing it to a level that is randomly chosen from the distribution of the mediator amongst all of those with a particular exposure.

These various effects are similar to those described by Didelez et al. (2006). Didelez et al. (2006) restrict these effects to settings in which there is no exposure-induced mediator–outcome confounding. However, in fact, these effects are identified even in the presence of exposure-induced mediator–outcome confounding (VanderWeele et al., 2013c). Suppose there is one or more exposure induced mediator–outcome confounders L as in Figure 5.4. These randomized interventional analogues of the direct and indirect effects will be identified under the assumptions that conditional on C there is [assumption (A2.1)] no unmeasured exposure–outcome and [assumption (A2.3)] no unmeasured exposure–mediator confounding, along with the assumption [assumption (A5.5)] that that conditional on (A, C, L), there is no unmeasured confounding of the mediator–outcome relationship. These three assumptions would hold in the causal diagram in Figure 5.4. They would still hold even if the association between L and Y were also confounded by unmeasured factors as in Figure 5.5.

If there is no exposure-induced mediator–outcome confounder L so that assumptions (A2.1)–(A2.4) hold, then estimators for these randomized interventional analogues of the natural direct and indirect effects in fact then coincide with those for the natural direct and indirect effects themselves. However, if there is an exposure-induced mediator–outcome confounder L, then although the natural direct and indirect effects themselves are not identified, these randomized interventional analogues are and will permit us to carry out a particular type of effect decomposition even in this setting. By focusing on randomized interventions on the mediator, rather than fixing the mediator for each individual to what it would have been under a counterfactual scenario, we can identify these effects. In the next subsection we describe a weighting-based approach to estimating these effects.

5.4.2. A Weighting Approach to Estimate Randomized Interventional Analogues of Natural Direct and Indirect Effects

The weighting-based approach requires first estimating inverse probability weights, which can be obtained from regression models for the exposure A, mediator M and confounder L. We will focus on the case of a dichotomous exposure (coded 0 or 1), dichotomous mediator, and dichotomous exposure-induced confounder. The approach described below is in principle applicable to other settings (e.g., with these variable discrete with a few levels); but as already noted, most

weighting-based approaches do not work very well with continuous variables (other than a continuous outcome) because they are not very stable.

When the exposure, mediator, and exposure-induced mediator–outcome confounder are all binary, three logistic regressions can be used to construct the weights: (1) a logistic regression of the exposure A on covariates, (2) a logistic regression of the confounder L on the exposure and covariates and (3) a logistic regression of the mediator on the confounder, exposure, and covariates. The procedure we describe is for marginal effects (i.e., averaged over the covariates). Analogous estimators for conditional effects are described in the Appendix. Once the logistic regression models are fit and the predicted probabilities are calculated, weights can be formed by

$$\frac{\sum_l P(m|l,a^*,c)P(l|a^*,c)}{P(a|c)P(m|l,a,c)}$$

Under assumptions (A2.1), (A2.3), and (A5.5), weighting-based estimators for the randomized interventional analogues of the natural direct and indirect effects can then be obtained upon duplicating the dataset and adding an exposure variable A^* which is 0 for the first replication and 1 for the second. The randomized interventional analogue of the natural direct effect NDE^R is then obtained as the coefficient of A in a weighted regression of Y on A among individuals with $A^* = 0$; the natural indirect effect NIE^R is obtained as the coefficient of A^* in a weighted regression of Y on A^* among individuals with $A = 1$ in the duplicated dataset (VanderWeele et al., 2014c). Bootstrapping can be used for standard errors and confidence intervals.

5.4.3. SAS Implementation

We describe how the proposed weighting approaches to marginal direct and indirect effects given above can be implemented in SAS statistical software. Below we let `c`, `a`, `l`, `m`, and `y` correspond to the observed confounders C, exposure A, exposure-induced confounder L, mediator M, and outcome Y. We assume that a dataset has been created with these variables called "mydata."

```
data mydata0;
set mydata;
a = 0; output;
run;
data mydata1;
set mydata;
a = 1; output;
run;
proc logistic data = mydata;
model a = c;
score data = mydata out = preda;
run;
data preda;
set preda;
pa1 = P_1;
run;
```

```
proc logistic data = mydata;
model l = a c;
score data = mydata1 out = predl1;
score data = mydata0 out = predl0;
run;
data predl1;
set predl1;
pl1 = P_1;
run;
data predl0;
set predl0;
pl10 = P_1;
run;
data mydata00;
set mydata;
a = 0; l = 0; output;
run;
data mydata10;
set mydata;
a = 1; l = 0; output;
run;
data mydata01;
set mydata;
a = 0; l = 1; output;
run;
data mydata11;
set mydata;
a = 1; l = 1; output;
run;
proc logistic data = mydata;
model m = a l c;
score data = mydata1 out = predm1;
score data = mydata0 out = predm0;
score data = mydata00 out = predm00;
score data = mydata01 out = predm01;
score data = mydata10 out = predm10;
score data = mydata11 out = predm11;
run;
data predm1;
set predm1;
pm1 = P_1;
run;
data predm0;
set predm0;
pm10 = P_1;
run;
data predm00;
set predm00;
pm100 = P_1;
run;
data predm10;
set predm10;
pm110 = P_1;
run;
data predm01;
```

```
set predm01;
pm101 = P_1;
run;
data predm11;
set predm11;
pm111 = P_1;
run;
data mydataw;
merge preda predl1 predl0 predm1 predm0 predm00 predm01 predm10 predm11 mydata;
run;
data mydatanew;
set mydataw;
astar = 1-a;
if (a = 0) & (astar = 0) & (m = 1) then
  w3 = (1/(1-pa1))*(pm100*(1-pl10)+pm101*pl10)/pm10;
if (a = 0) & (astar = 0) & (m = 0) then
  w3 = (1/(1-pa1))*((1-pm100)*(1-pl10)+(1-pm101)*pl10)/(1-pm10);
if (a = 0) & (astar = 1) & (m = 1) then
  w3 = (1/(1-pa1))*(pm110*(1-pl1)+pm111*pl1)/pm10;
if (a = 0) & (astar = 1) & (m = 0) then
  w3 = (1/(1-pa1))*((1-pm110)*(1-pl1)+(1-pm111)*pl1)/(1-pm10);
if (a = 1) & (astar = 0) & (m = 1) then
  w3 = (1/pa1)*(pm100*(1-pl10)+pm101*pl10)/pm1;
if (a = 1) & (astar = 0) & (m = 0) then
  w3 = (1/pa1)*((1-pm100)*(1-pl10)+(1-pm101)*pl10)/(1-pm1);
if (a = 1) & (astar = 1) & (m = 1) then
  w3 = (1/pa1)*(pm110*(1-pl1)+pm111*pl1)/pm1;
if (a = 1) & (astar = 1) & (m = 0) then
  w3 = (1/pa1)*((1-pm110)*(1-pl1)+(1-pm111)*pl1)/(1-pm1);
run;
```

The results for the randomized interventional analogue of the natural direct effects can then be obtained from

```
proc surveylogistic data = mydatanew;
where astar = 0;
model y = a;
weight w;
run;
```

and those for the randomized interventional analogue of the natural indirect effect from

```
proc surveylogistic data = mydatanew;
where a = 1;
model y = astar;
weight w;
run;
```

where `astar` corresponds to A^* and where the use of `proc surveylogistic` ensures the use of robust standard errors.

5.4.4. An Example of Estimating Randomized Interventional Analogues of Natural Direct and Indirect Effects

To illustrate this approach, we will again analyze 2003 US birth certificate data but will consider a slightly differently question, namely whether the exposure, A, of adequate or inadequate prenatal care ($n = 2,629,247$; those with intermediate or superadequate care are excluded from the analysis for the purposes of this illustration) on preterm birth (Y) is mediated by pre-eclampsia (M) or other pathways, where maternal smoking (L) is taken as an exposure-induced mediator–outcome confounder. Maternal smoking may be affected by prenatal care and may in turn affect both pre-eclampsia and preterm birth. Adequacy of prenatal care categories are determined from data on the month that prenatal care was initiated, on the number of visits, and on gestational age, according to the American College of Gynecologists recommendation as encoded in a modification of the APNCU index (Kotelchuck, 1994; VanderWeele et al., 2009). In this analysis we adjusted for age category (below 20 years, between 20 and 35 years, or above 35 years), ethnicity (black, Hispanic, native American, white), education, and marital status as baseline confounders (C).

Inverse probability weights were constructed on the basis of logistic regression models for adequate care, smoking, and pre-eclampsia. In view of the large sample size and the resulting computational complexity, standard errors and confidence intervals were constructed using the subsampling bootstrap (Politis and Romano, 1994). This is similar to the bootstrap, but involved repeating the analysis for 1000 subsamples of size $n = 13,146$ (0.5% of the total sample size); on the basis of the empirical standard deviation of the 1000 estimates, the standard error of the estimates that were obtained from the analysis of the full dataset can be inferred.

When we apply the methods described above, we have that the marginal randomized interventional analogue of direct effect of adequate care, other than via pre-eclampsia, is to reduce the odds of preterm birth by 54.4% (95% CI 53.9% to 54.9%). The estimate of the marginal randomized interventional analogue of the natural indirect effect via pre-eclampsia was a negligible increase in the odds of preterm birth by 0.01% (95% CI −0.89% to 0.91%). Note that here, unlike the illustration in Section 5.2.3, we are reporting effects on the odds ratio scale, rather than the absolute risk scale. Conditional effects for this example, described in the Appendix, are reported in VanderWeele et al. (2014c).

5.5. PATH-SPECIFIC EFFECTS

When there are multiple mediators on the pathway from exposure to outcome, it may be of interest to separate out the effects through a variety of different pathways. In Sections 5.1–5.3 we considered how to assess the effect through and not through a particular set of mediators. In this section we will consider what further inferences might be possible when path-specific effects are of interest. As might already be clear, in general settings allowing for nonlinearities and interaction, some path-specific effects are not identified. The estimation of path-specific effects

is commonly done in the social sciences using a structural equation model. This effectively achieves the identification of these effects by assuming linearity and no interaction and by assuming away all possible confounding relationships (rather than just a few; cf. VanderWeele, 2012e). Perhaps in some settings these assumptions are not unreasonable or the techniques could be used for exploratory purposes. Here, however, we will not make these assumptions about linearity and the absence of interaction and consider what further inferences we may be able to obtain about path-specific effects. The reader is referred to Bollen (1989) and MacKinnon (2008) for descriptions of the estimation of path-specific effects using structural equation models.

5.5.1. Path-Specific Effects with Two Mediators

Consider again the setting with two mediators, call them L and M as in Figure 5.4. As noted above, natural direct and indirect effects with M considered alone as the mediator are not in general identified under the causal diagram of Figure 5.4. However, certain path-specific effects are identified. For example, although we cannot identify the effects mediated through pathways involving M (i.e., the combination of $A \rightarrow L \rightarrow M \rightarrow Y$ and $A \rightarrow M \rightarrow Y$) and the effects through pathways not involving M (i.e., the combination of $A \rightarrow Y$ and $A \rightarrow L \rightarrow Y$), Avin et al. (2005) showed that we can identify the effects (i) through pathways involving neither L nor M (i.e., $A \rightarrow Y$) (ii) through the additional pathways not involving L (i.e., $A \rightarrow M \rightarrow Y$), and (iii) through the pathways involving L (i.e., the combination of $A \rightarrow L \rightarrow M \rightarrow Y$ and $A \rightarrow L \rightarrow Y$). For simplicity, let us refer to these effects as $E_{A\rightarrow Y}$, $E_{A\rightarrow M\rightarrow Y}$, $E_{A\rightarrow LY}$. Formal counterfactual definitions of these effects are given in the Appendix. It turns out we can in fact decompose a total effect into the sum of these three effects: $E_{A\rightarrow Y} + E_{A\rightarrow M\rightarrow Y} + E_{A\rightarrow LY}$. Note that these expressions do not allow us to distinguish between effects through L and through M versus those that are through L but not through M. Thus we can identify some but not all of the path-specific effects that may be of interest. As noted in the Appendix, this identification holds irrespective of the models that are used or the presence of nonlinearities or interactions.

We can also use a fairly simple weighting approach to estimate these three path-specific effects. A weighting-based estimator can be obtained by merging three copies of the dataset and adding exposure variables A^* and A^{**}, where A^* equals the observed exposure for the first replication and $1 - A$ for the next two replications, and where A^{**} equals the observed exposure for the first two replications and $1 - A$ for the third replication. For each individual, a weight is now obtained by taking the product of the predicted probability (of the observed confounder value L) from the logistic regression for L had the exposure been A^* and the predicted probability (of the observed mediator value M) from the logistic regression for M had the exposure been A^{**}, divided by the product of the corresponding predicted probabilities from the two logistic regressions had the exposure been as observed:

$$\frac{P(l|a^*,c)P(m|l,a^{**},c)}{P(a|c)P(l|a,c)P(m|l,a,c)}$$

For $a = 1$ and $a^* = 0$, the natural direct effect $E_{A\rightarrow Y}$ is now obtained as the coefficient of A in a weighted regression of Y on A among individuals with $A^* = A^{**} = 0$; the natural indirect effect $E_{A\rightarrow LY}$ is obtained as the coefficient of A^* in a weighted regression of Y on A^* among individuals with $A = 1, A^{**} = 0$; finally, the natural indirect effect $E_{A\rightarrow M\rightarrow Y}$ is obtained as the coefficient of A^{**} in a weighted regression of Y on A^{**} among individuals with $A = A^* = 1$ (VanderWeele et al., 2014c). If robust standard errors are used, this will yield confidence intervals and estimates of the standard error that are conservative for the marginal effects (averaged over the covariates). With the weighting approach for conditional effects, described in the Appendix, standard errors and confidence intervals can be obtained via the bootstrap.

To implement this in SAS, one can first run the code in Section 5.4.2 and then further run the code

```
data mydatanew2;
set mydataw;
astar = a; astarstar = a; w2 = a/pa1+(1-a)/(1-pa1); output;
astar = 1-a; astarstar = a;
w2 = (a/pa1)*(l*pl10/pl1+(1-l)*(1-pl10)/(1-pl1)) +
((1-a)/(1-pa1))*(l*pl1/pl10+(1-l)*(1-pl1)/(1-pl10)); output;
astar = 1-a; astarstar = 1-a;
w2 = (a/pa1)*(l*(pl10/pl1)*(m*(pm101/pm111)+(1-m)*(1-pm101)/(1-pm111))
+(1-l)*((1-pl10)/(1-pl1))*(m*(pm100/pm110)+(1-m)*(1-pm100)/(1-pm110)))
+((1-a)/(1-pa1))*(l*(pl1/pl10)*(m*(pm111/pm101)+(1-m)*(1-pm111)/(1-pm101))
+(1-l)*((1-pl1)/(1-pl10))*(m*(pm110/pm100)+(1-m)*(1-pm110)/(1-pm100))); output;
run;
```

and then to obtain the estimates of the three path-specific effects, one can use the coefficients from the following three models:

```
proc surveylogistic data = mydatanew;
where (astar = 0) & (astarstar = 0);
model y = a;
weight w2;
run;
```

for $E_{A\rightarrow Y}$,

```
proc surveylogistic data = mydatanew;
where (a = 1) & (astarstar = 0);
model y = astar;
weight w2;
run;
```

for $E_{A\rightarrow M\rightarrow Y}$, and

```
proc surveylogistic data = mydatanew;
where (a = 1) & (astar = 1);
model y = astarstar;
weight w2;
run;
```

for $E_{A\rightarrow LY}$.

If we apply this approach to the example in Section 5.4.4, we obtain a direct effect $E_{A\rightarrow Y}$ estimate of a reduction in odds of preterm birth by 54.1% (95% CI 53.6% to 54.6%), the mediated effect via smoking $E_{A\rightarrow LY}$ amounts to a 0.10% (95% CI -0.79% to 1.01%) increase in the odds of preterm birth. That this effect is small is perhaps not surprising since the effect of adequate care by decreasing smoking mixes an inherent beneficial impact on preterm birth with a detrimental effect on preterm birth by increasing pre-eclampsia (since smoking sometimes prevents pre-eclampsia). The remaining mediated effect via preeclampsia, but not smoking, $E_{A\rightarrow M\rightarrow Y}$, is to increase the odds of preterm birth of 0.20% (95% -0.70% to 1.11%). Conditional effects for this example, described in the Appendix, are reported in VanderWeele et al. (2014c).

5.5.2. More General Identification Criteria and Estimation Approaches

In the previous subsection we described one approach, along with one set of assumptions, to identification and the estimation of path-specific effects, without making assumptions about linearity and the absence of interactions. These path-specific effects, like our natural direct and indirect effects, are defined formally in terms of certain nested counterfactuals. In fact, a whole theory is available for when such nested counterfactuals and path-specific effects are or are not nonparametrically identified (Avin et al., 2005; Shpitser and Pearl, 2008). However, the theory and assumptions required to determine whether a particular path-specific effect is identified can be quite technical. One simple condition, however, sometimes referred to as the "recanting witness criterion," is not quite as difficult to state. We can think of a path-specific effect in some sense as comparing (i) a baseline state in which we set the exposure to a particular level to (ii) some other state in which we fix the exposure to some other level but only for certain paths. We will refer to these paths with the exposure fixed to a different level as "active paths." For example in Figure 5.4 the natural indirect effect through M would essentially be activating the paths A–L–M–Y and A–M–Y. The recanting witness criterion states that a particular path-specific effect from exposure A to outcome Y is identified if and only if there is no variable W such that there is an active path from A to W and there is an active path from W to Y and also an inactive path from W to Y. Such a variable with an active path from A to W, an active path from W to Y, and an inactive path from W to Y is sometimes referred to as a "recanting witness" (Avin et al., 2005). The recanting witness criterion thus states that a path-specific effect is identifiable if and only if there is no recanting witness.

To see how this works, consider again the causal diagram in Figure 5.4. We noted above that the natural indirect effect would be activating the paths A–L–M–Y and A–M–Y. We have also noted that this effect is not identified on this diagram. We can see this using the "recanting witness criterion" by noting that L here is a recanting witness for the natural indirect effect on this diagram. To see this, note that there is an active path from A to L—for example, on the path A–L–M–Y; there is also an active path from L to Y, again on the path A–L–M–Y; but there is moreover an inactive path from L to Y, namely on the path, A–L–Y, which is not part of the

natural indirect effect and thus inactive. Thus we can see from the recanting witness criterion that the natural indirect effect is not identified from the data without making additional assumptions. We noted in the previous section, however, that certain path-specific effects were identified in this diagram—for example, A–Y and A–M–Y. The path-specific effect A–Y has no intermediate on it, and so there can be no recanting witness for it and it is thus identified. The path A–M–Y does have an intermediate, M, and there is an active path from A to M and also an active path from M to Y, but there is no inactive path for this path-specific effect, and so there is no recanting witness and thus this effect also is identified from the data. Finally, we noted that the combination of the paths A–L–M–Y and A–L–Y was also identified. Here we have two intermediates on these paths, L and M. For M there is no path from M to Y that is inactive, and so it is not a recanting witness. For L there is likewise no path from L to Y that is inactive, so it is also not a recanting witness. Thus the combination of these two paths is also identified. An approach similar to the identification of path-specific effects could also be used on other causal diagrams.

However, even if path-specific effects are identified, we still need practical approaches to estimation. Albert and Nelson (2011) describe a general estimation approach for path-specific effects, but one which makes strong independence assumptions about counterfactuals, far stronger than the assumptions we have been considering in this book and assumptions that do not follow from causal diagrams. They do this to get around the identification issues discussed above. They do also supply a sensitivity analysis approach to evaluate the sensitivity of estimates to some of their assumptions, but the parameters in the sensitivity analysis are difficult to interpret and specify in practice. Further methodological development concerning practical tools to assess path-specific effects is still needed.

5.6. SENSITIVITY ANALYSIS FOR EXPOSURE-INDUCED CONFOUNDING

In this section we will discuss sensitivity analysis that can be used for natural direct and indirect effects in the presence of exposure-induced mediator–outcome confounding. Unlike the sensitivity analysis techniques described in Chapter 3 that had a relatively easy interpretation as the effect of an unmeasured confounder on the outcome or as the prevalence of the unmeasured confounder in certain strata of the exposure and mediator, the sensitivity analysis parameters in this section will involve counterfactuals and thus essentially necessitate the use of counterfactual notation directly. Readers uncomfortable with this notation could either review again Section 2.16 or skip this section and move on to the end of the chapter discussion and then to Chapter 6.

5.6.1. Sensitivity Analysis for Exposure-Induced Confounding for a Continuous Outcome

The methods we describe will assume that control has been made for [assumption (A2.1)] exposure–outcome confounding and [assumption (A2.3)]

exposure–mediator confounding (or that the exposure itself has been randomized), but no further assumptions about confounding will be made beyond that. The sensitivity analysis approach will allow for mediator–outcome confounding variables including mediator–outcome confounding variables affected by the exposure. Either our covariate set C can be empty or we can consider analysis conditional on $C = c$. As a motivating example to which we will return to below to illustrate the approach, we will consider the extent to which the effects of an exposure A constituting a randomized care management intervention on a depression score outcome Y are mediated by an indicator of adherence to antidepressant medication M.

For each possible level of the mediator m, consider the following bias parameter:

$$\gamma_{mc} = \mathbb{E}[Y_{1m}|A = 1, M = m, c] - \mathbb{E}[Y_{1m}|A = 0, M = m, c] \tag{5.4}$$

In the context of the example, the bias parameter $\gamma_{1c} = \mathbb{E}[Y_{11}|A = 1, M = 1, c] - \mathbb{E}[Y_{11}|A = 0, M = 1, c]$ would be a contrast of depressive outcome scores for two subpopulations. The two subpopulations would be (a) those who had in fact received the care management intervention ($A = 1$) and adhered to the antidepressant ($M = 1$) and (b) those who had not received the care management intervention ($A = 0$) but had adhered to the antidepressant ($M = 1$). We would then consider what would have happened to depression scores for these two subpopulations had we intervened to give the care management program and ensure adherence antidepressant (i.e., we would consider Y_{11}); the contrast between the depression scores for these two subpopulations under this particular intervention is our sensitivity analysis parameter γ_{1c}. If we thought that the second subpopulation was overall healthier or more competent (e.g., because they adhered even though they did not have the care management intervention), then the depression scores might be higher in the first subpopulation and our sensitivity parameter γ_{1c} would in this case be positive.

The other bias parameter $\gamma_{0c} = \mathbb{E}[Y_{10}|A = 1, M = 0, c] - \mathbb{E}[Y_{10}|A = 0, M = 0, c]$ would also be a contrast of depressive outcome scores for two subpopulations. The two subpopulations in this case would be (a) those who had in fact received the care management intervention ($A = 1$) but had not adhered to the antidepressant ($M = 0$) and (b) those who had not received the care management intervention ($A = 0$) and had not adhered to the antidepressant ($M = 0$). We would consider what would have happened to depression scores for these two subpopulations had we intervened to give the care management program but had not allowed adherence to the antidepressant (i.e., we would consider Y_{10}); the contrast between the depression scores for these two subpopulations under this particular intervention is our sensitivity parameter γ_{0c}. Again, if we thought that the second subpopulation were healthier or more competent (e.g., because the first did not adhere even though they had the care management intervention), then the depression scores for the first subpopulation would be higher and our sensitivity parameter γ_{0c} would also be positive.

Suppose now that there is no unmeasured [assumption (A2.1)] exposure–outcome confounding or [assumption (A2.3)] exposure–mediator confounding (or, for example, that exposure is randomized) and that we proceed to attempt to

estimate natural direct and indirect effect using methods, such as those described in Chapter 2, which will be consistent for natural direct and indirect effects if no-unmeasured–confounding assumptions (A2.1)–(A2.4) hold but would not be if there were unmeasured mediator–outcome confounders or if there were exposure-induced mediator–outcome confounders. Let Q_{NIE} and Q_{NDE} denote the estimates obtained using these methods for the natural indirect effect and the natural direct effect, respectively. These quantities will be consistent for the natural indirect and direct effects, respectively, if assumptions (A2.1)–(A2.4) hold. Suppose now that assumptions (A2.1) and (A2.3)—no unmeasured exposure–outcome confounding and no unmeasured exposure–mediator confounding, respectively—hold but either (A2.2) or (A2.4) is violated; that is, either we have unmeasured mediator–outcome confounding or there is a mediator–outcome confounder affected by the exposure. If we had unmeasured mediator–outcome confounding variables that were not affected by exposure, we could use sensitivity analysis techniques in Chapter 2. However, if we have a mediator–outcome confounder that is affected by exposure, these techniques are not inapplicable.

Define the bias factor for natural indirect effect, B_c^{NIE}, as the difference between the observed estimate Q_{NIE} for the natural indirect effect and the true natural indirect effect; likewise define the bias factor for natural direct effect, B_c^{NDE}, as the difference between the observed estimate Q_{NDE} for the natural direct effect and the true natural direct effect. It can then be shown that if we have our sensitivity parameters $\gamma_{mc} = \mathbb{E}[Y_{1m}|A = 1, m, c] - \mathbb{E}[Y_{1m}|A = 0, m, c]$ and if we let $\Gamma_c = \sum_m \gamma_{mc} P(m|A = 0, c)$, then we have that (VanderWeele and Chiba, 2014)

$$\begin{aligned} B_c^{NIE} &= -\Gamma_c \\ B_c^{NDE} &= \Gamma_c \end{aligned} \tag{5.5}$$

If we estimate Q_{NIE} and Q_{NDE} using methods for natural indirect and direct effects from Chapter 2 but if the confounding assumptions (A2.1)–(A2.4) do not hold because of unmeasured mediator–outcome confounding or an exposure-induced mediator–outcome confounding variable, then our estimators will not be valid for the true natural indirect and direct effects. However, we can obtain corrected estimates of the natural indirect and direct effects by specifying the sensitivity analysis parameter γ_{mc} for each level of m (we will have two such parameters if M is binary as above) and then computing the bias factors from formula (5.5). To obtain corrected estimates for natural indirect and direct effects, we then subtract the bias factors, B_c^{NIE} and B_c^{NDE}, respectively, from our estimates of the natural indirect and direct effects. The corrected estimators will be consistent for the true natural indirect and direct effects if we have specified the bias parameters γ_{mc} correctly. Of course, we do not know what the true values of these bias parameters are, but we can vary them in a sensitivity analysis to assess the extent to which our conclusions about direct and indirect effects depend on the magnitude of these parameters as we did also in Chapter 3. We give an example of such a sensitivity analysis below.

The bias parameters in (5.5) depend on the probabilities $P(m|A = 0, c)$ which must be estimated from the data. This can make obtaining corrected confidence intervals for direct and indirect effect estimates more challenging. If γ_{mc} were

constant across strata of m, then the bias factors B_c^{NIE} and B_c^{NDE} would no longer depend on $P(m|A=0,c)$ and we could simply subtract B_c^{NIE} and B_c^{NDE} from both limits of the confidence intervals for Q_{NIE} and Q_{NDE} to obtain corrected confidence intervals for natural indirect and direct effects, respectively. Likewise, if the dataset is sufficiently large so that estimates of $P(m|A=0,c)$ are very precise (e.g., if the mediator were binary and the covariate set C empty) and the sample size is large, then approximate corrected confidence intervals for natural indirect and direct effects could be obtained by simply subtracting B_c^{NIE} and B_c^{NDE} from both limits of the confidence intervals for our observed estimates, Q_{NIE} and Q_{NDE}. In other contexts, however, in which we must estimate $P(m|A=0,c)$ from the data and our estimates of $P(m|A=0,c)$ are themselves subject to sampling variability, then to obtain corrected confidence intervals for natural indirect and direct effects we could proceed by bootstrapping wherein for each fixed value of the sensitivity analysis parameters γ_{mc} and with each bootstrapped sample we would obtain both observed estimates of Q_{NIE} and Q_{NDE} and estimates of $P(m|A=0,c)$ and subsequently B_c^{NIE} and B_c^{NDE} to derive a corrected estimate of the natural direct and indirect effects. Corrected confidence intervals could then be obtained by using a percentile method over the corrected estimates across the bootstrapped samples.

It also follows from the bias factor formulae in (5.5) that if the bias parameter γ_{mc} is non-negative for all values of m, then using the observed data and our observed estimates of the natural indirect and direct effects, Q_{NIE} and Q_{NDF}, we will underestimate the true natural indirect effect and overestimate the true natural direct effect. Conversely, if the bias parameter γ_{mc} is nonpositive for all values of m, then using the observed data and our observed estimates of the natural indirect and direct effects, Q_1 and Q_2, we will overestimate the true natural indirect effect and underestimate the true natural direct effect. We will use this result in the application below.

5.6.2. Example of Sensitivity Analysis for Exposure-Induced Confounding

Emsley et al. (2010) considered mediation in the Prevention of Suicide in Primary Care Elderly: Collaborative Trial (PROSPECT). They assessed whether the effect of randomized exposure (collaborative care management versus exposure as usual), A, on the score from the Hamilton Depression Scale, Y, was mediated by adherence to antidepressants, M. They assumed no interaction between the effect of A and M on Y and obtained an estimate of the direct effect of -2.66 (standard error = 0.93) and an estimate of the indirect effect of -0.49 (standard error = 0.43). If there was indeed no exposure–mediator interaction, and if assumptions (A2.1)–(A2.4) about confounding for the mediator M were satisfied, then their estimator of the direct effect would be consistent for the natural direct effect and their estimator of the indirect effect would be equal to the natural indirect effect.

In their analysis, however, it is likely that there are variables that confound the relationship between antidepressant adherence and depression scores. Medical co-morbidities might affect both depression scores and also whether a patient

is adherent to antidepressant since patient's medical co-morbidities may deter patients from taking antidepressant medications because of so many other medications necessitated by their medical condition. If these mediator–outcome confounding variables were not affected by exposure, we could potentially use the sensitivity analysis technique in Chapter 3. However, it may be the case that the medical co-morbidities are affected by the collaborative care management intervention. Moreover, there may be other mediator–outcome confounding variables that are affected by the exposure. For example, we might consider whether patients have a regular eating schedule during follow-up. If patients typically take antidepressant medications with meals, then having a regular eating schedule may affect adherence; regulation of meals and diet may also affect depression scores; and whether there is a regular eating schedule may itself be affected by whether collaborative care management is provided.

We may allow for such mediator–outcome confounding variables affected by exposure and still apply the sensitivity analysis approach in the previous subsection. As above, our sensitivity parameter $\gamma_{1c} = \mathbb{E}[Y_{11}|A = 1, M = 1, c] - \mathbb{E}[Y_{11}|A = 0, M = 1, c]$ compares what depression scores would have been with care management and adherence for two subpopulations, those who received the care management intervention $(A = 1)$ and adhered to the antidepressant $(M = 1)$, versus those had not received the care management intervention $(A = 0)$ but had adhered to the antidepressant $(M = 1)$. If we thought that the second subpopulation was overall healthier or more competent, we might specify a positive sensitivity analysis parameter, for example, $\gamma_{1c} = 1$ (roughly a quarter of a standard deviation for the depressive symptom scale). The other sensitivity parameter, $\gamma_{0c} = \mathbb{E}[Y_{10}|A = 1, M = 0, c] - \mathbb{E}[Y_{10}|A = 0, M = 0, c]$, compares what depression scores would have been with care management without adherence for two subpopulations, for those who received the care management intervention $(A = 1)$ but had not adhered to the antidepressant $(M = 0)$ versus those who had not received the care management intervention $(A = 0)$ and had not adhered to the antidepressant $(M = 0)$. If we thought that the second subpopulation was healthier or more competent, we might again specify a positive sensitivity analysis parameter, for example, $\gamma_{0c} = 0.5$.

The probability of the mediator in the control group in their data is 0.45 and we could then calculate $\Gamma_c = \sum_m \gamma_{mc} P(m|A = 0, c) = (0.55)(0.5) + (0.45)(1) = 0.725$. The corrected estimates for the direct and indirect effects would be $-2.66 - 0.725 = -3.39$ and $-0.49 - (-0.76) = 0.24$, respectively. The direct effect would still be quite substantial, but the indirect effect (the effect mediated by antidepressant adherence) would be detrimental (i.e., would increase depression), which does not seem likely. The sensitivity analysis parameters specified may be too extreme. We could of course specify different sensitivity analysis parameters as well. If we thought that the sensitivity analysis parameters were half of what was specified above, we would have $\Gamma_c = 0.36$ and corrected direct and indirect estimates of $-2.66 - 0.36 = -3.02$ and $-0.49 - (-0.36) = -0.13$, respectively. As noted above, if the sensitivity parameters are positive, then irrespective of their actual values we would have the true direct effect, which was in fact more negative (more protective) than the initial estimate of -2.66. Moreover, even if the

sensitivity parameters took the opposite sign, they would have to be fairly substantial in magnitude—for example, $\gamma_{1c} = \gamma_{0c} = 2.66$ (roughly half a standard deviation in depression scores)—to explain away the direct effect. The direct effect itself then seems fairly robust to potential unmeasured mediator–outcome confounding or to exposure-induced mediator–outcome confounding; however, the indirect, as we have seen, is not.

5.6.3. Sensitivity Analysis for Natural Direct and Indirect Effects on a Ratio Scale in the Presence of Exposure-Induced Confounding

We will now consider how a similar sensitivity analysis technique can be employed when natural direct and indirect effects on the ratio scale are of interest. Suppose again that control has been made for [assumption (A2.1)] exposure–outcome confounding and [assumption (A2.3)] exposure–mediator confounding (or that the exposure itself has been randomized) but that there may be unmeasured mediator–outcome confounders or the mediator–outcome confounders may be affected by the exposure. Suppose then that we proceed to attempt to estimate natural direct and indirect effects using methods, such as those described in Chapter 2 for natural direct and indirect effects on the ratio scale, which will be consistent for natural direct and indirect effects if no-unmeasured-confounding assumptions (A2.1)–(A2.4) hold, but will not be consistent if there are unmeasured mediator–outcome confounders or if there are mediator–outcome confounders that may be affected by the exposure. Let Q_{NIE} and Q_{NDE} denote the estimates obtained using these methods from Chapter 2 for the natural indirect effect and natural direct effect, respectively. Let OR^{NIE} and OR^{NDE} be the true natural indirect and direct effect odds ratios, respectively. Define the following bias factors:

$$B_c^i = \frac{1}{Q_{NIE}} - \frac{1}{OR^{NIE}}$$
$$B_c^d = Q_{NDE} - OR^{NDE}$$

If we specify the same sensitivity analysis parameters as before, namely $\gamma_{mc} = \mathbb{E}[Y_{1m}|A=1,m,c] - \mathbb{E}[Y_{1m}|A=0,m,c]$, and if we let $\Gamma_c = \sum_m \gamma_{mc} P(m|A=0,c)$, we then have that (VanderWeele and Chiba, 2014)

$$B_c^i = \frac{\Gamma_c}{\mathbb{E}[Y|A=1,c]}$$
$$B_c^d = \frac{\Gamma_c}{\mathbb{E}[Y|A=0,c]} \qquad (5.6)$$

And once we have calculated the bias parameters, we can then obtain corrected estimates of the natural indirect and direct effect odds ratios by

$$OR^{NIE} = \frac{Q_{NIE}}{1 - Q_{NIE} \times B_c^i}$$
$$OR^{NDE} = Q_{NDE} - B_c^d$$

If we estimate Q_{NIE} and Q_{NDE} using methods for natural indirect and direct effects on the ratio scale as in Chapter 2 but if the identification assumptions (A2.1)–(A2.4) do not hold because of unmeasured mediator–outcome confounding or an exposure-induced mediator–outcome confounding variable, then our estimators will not be consistent for the true natural indirect and direct effects. However, we can use (5.6) to obtain corrected natural indirect and direct effect estimates by specifying the sensitivity analysis parameters $\gamma_{mc} = \mathbb{E}[Y_{1m}|A = 1, m, c] - \mathbb{E}[Y_{1m}|A = 0, m, c]$ and then using this to obtain the bias factors B_c^i and B_c^d by using also empirical estimates for $\mathbb{E}[Y|A = 1, c]$ and $\mathbb{E}[Y|A = 0, c]$ and then using these bias factors to obtain corrected natural indirect and direct effect estimates on the ratio scale.

In general, to obtain corrected confidence intervals for natural indirect and direct effect risk ratios, we would have to use bootstrapping, wherein for each fixed value of the sensitivity analysis parameters γ_{mc} and with each bootstrapped sample, we would obtain both estimates of Q_{NIE} and Q_{NDE} and also estimates of $P(m|A = 0, c)$, $\mathbb{E}[Y|A = 1, c]$, and $\mathbb{E}[Y|A = 0, c]$ from the data and subsequently use these to calculate B_c^i and B_c^d and use the formulas in (5.6) to calculate a corrected estimate of the natural indirect and direct effect risk ratios. Corrected confidence intervals could then be obtained by using a percentile method over the corrected estimates across the bootstrapped samples. As was also the case on the difference scale, if the sensitivity analysis parameters are greater than or equal to 0 (i.e., $\gamma_{mc} \geq 0$ for all m), then $OR^{NIE} \geq Q_{NIE}$ (i.e., the true indirect effect odds ratio will be greater than or equal to our estimate) and $OR^{NDE} \leq Q_{NDE}$; that is, the true direct effect odds ratio will be less than or equal to our estimate. On the other hand, if the sensitivity analysis parameters are less than or equal to 0 (i.e., $\gamma_{mc} \leq 0$ for all m), then $OR^{NIE} \leq Q_{NIE}$ (i.e., the true indirect effect odds ratio will be less than or equal to our estimate) and $OR^{NDE} \geq Q_{NDE}$ (i.e., the true direct effect odds ratio will be greater than or equal to our estimate).

5.6.4. Other Sensitivity Analysis Techniques for Natural Direct and Indirect Effects for Exposure-Induced Confounding

Other techniques for sensitivity analysis for direct and indirect effects have also recently been developed that handle the presence of an exposure-induced mediator–outcome confounder. Imai and Yamamoto (2012) proposed a technique that requires specification of a linear structural equation model with random coefficients. The technique requires stronger modeling assumptions than the one presented here and also assumes that both the mediator and the outcome are continuous whereas the technique presented above is more broadly applicable. Their technique also assumes that data are available on the exposure-induced mediator outcome confounder L, whereas the one presented in the previous subsections does not. However, when the mediator and outcome are continuous, their technique can be a useful alternative to the one presented in this chapter.

Tchetgen Tchetgen and Shpitser (2012) proposed a technique that requires specifying as sensitivity analysis parameters the quantities $\mathbb{E}[Y_{1m}|A = a, M = m,$

$C = c] - \mathbb{E}[Y_{1m}|A = a, M \neq m, C = c]$ for each a and m. For a fixed level of c, their technique thus requires specifying a number of sensitivity analysis parameters equal to the product of the number of levels of exposure A and the number of level of the mediator M, whereas our technique only requires specifying a number of sensitivity analysis parameters equal to the number of levels of M. Moreover, for the technique of Tchetgen Tchetgen and Shpitser (2012), the parameters $\mathbb{E}[Y_{1m}|A = a, M = m, C = c] - \mathbb{E}[Y_{1m}|A = a, M \neq m, C = c]$ may be difficult to specify in practice if M is not binary because the parameters are not a simple contrast comparing two values of M, but rather comparing a single value $(M = m)$ to an entire set of values $(M \neq m)$. Note that when M is not binary, both their technique and the one presented here will require specifying a potentially large number of parameters, making it more difficult to use these techniques in practice.

Finally, Vansteelandt and VanderWeele (2012) also proposed a sensitivity analysis technique for an exposure-induced mediator–outcome confounder. Their technique, like that of Imai and Yamamoto (2012), requires that data are available on the exposure-induced mediator–outcome confounder L, whereas the technique presented in the sections above does not. The technique of Vansteelandt and VanderWeele (2012) also involves specifying a selection bias function that can be difficult to interpret in practice, but does have the advantage that it is essentially zero so long as there is no three-way interaction between A, L, and M. The interested reader is referred to these various papers for further details of these approaches.

Finally, it should be noted that in the last three sections we have focused on what is sometimes called nonparametric identification—that is, identifying effects making assumptions only about confounding and not about the functional form of the relationships between variables. Under stronger assumptions about functional forms, natural direct and indirect effects can be identified under weaker assumptions about confounding. Structural equation models constitute a very extreme case of making additional functional form assumptions [in which all variables are assumed to have linear relationships and generally no interactions are included (cf. Bollen, 1989; MacKinnon, 2008)]. But progress can also be made under rather weaker functional form assumptions. For example, Robins (2003) noted that if the exposure and the mediator did not interact in their effects on the outcome at the individual counterfactual level, then natural direct effects were equal to controlled direct effects and thus the natural direct and indirect effects could be identified whenever controlled direct effects were (and thus the methods in Section 5.3 would be applicable here as well). Tchetgen Tchetgen and VanderWeele (2014b) showed that if the effect of the exposure A on an exposure-induced mediator–outcome confounder L were monotonic and if L was binary then natural direct and indirect effects were identified even in the presence of such an exposure-induced mediator–outcome confounder. Likewise they also showed that when the exposure-induced mediator–outcome confounder L did not interact with the mediator M in their effects on the outcome, then natural direct and indirect effects could likewise be identified from the data even with an exposure-induced mediator–outcome confounder. Additional progress can also be made in other settings. For example, when the exposure A is randomized and the exposure-induced mediator–outcome confounder L is compliance status and when treatment A can only affect the mediator

M and outcome Y through treatment compliance L, then identification of certain natural direct and indirect effects is once again sometimes possible (cf. Yamamoto, 2014).

5.7. DISCUSSION

The methods in this chapter have focused on multiple mediators. The approaches here can roughly be divided into two categories. First we considered estimation approaches that could be employed when, rather than just one mediator, we were interested in a whole set of mediators simultaneously. These methods and the assumptions required naturally extended the methods we had considered in Chapter 2. A regression-based approach was described and also an alternative weighting approach that offered somewhat more flexibility was described. The second category of methods we have described in this chapter concerned the setting when multiple mediators might be present, but we were only interested in one of them. This created analytical challenges when the mediator we were principally interested in was not the first one on the pathway from the exposure to the outcome. If there was a variable on the pathway from the exposure to the mediator of interest and that variable also affected the outcome, then this was a setting we had described in previous chapters as an exposure-induced mediator–outcome confounder. The natural direct and indirect effects for the mediator of interest were then not identified in this setting. We described in this chapter, however, that controlled direct effects could still be identified in this setting and we moreover supplied a randomized interventional analogue of natural direct and indirect effects that were also identified and could be used for effect decomposition. We moreover discussed certain path-specific effects that were identified and could be estimated; and finally, we described a sensitivity analysis technique for natural direct and indirect effects. This second category of techniques where we are only interested in one mediator and there are other mediators that precede it thus required a more subtle approach, but we have described a number of tools that can be used even in this more challenging setting. These two sets of methods—those that focus on a whole set of mediators versus those that focus on just one while accounting for the presence of the others—are of course addressing different questions, and thought needs to be given to which effect is of interest and why before selecting the appropriate tool.

6

Mediation Analysis with Time-Varying Exposures and Mediators

In previous chapters we have considered one single fixed exposure and generally, until Chapter 5, one fixed mediator. In Chapter 5 we considered settings with multiple mediators, but each of these was considered to take on a single value. In this chapter we will consider mediation analysis and effect decomposition when both the exposure and the mediator may change over time. This will help address questions about direct and indirect effects and effect decomposition in longitudinal settings when data are available on the exposure and on the mediator over time. An alternative title for this chapter might have been "Mediation Analysis with Longitudinal Data," but, as is hopefully clear from the preceding chapters, rigorous mediation analysis generally requires longitudinal data even if there is only a single exposure and a single mediator. This is because we want to be sure that the exposure temporally precedes the mediator and that the mediator temporally precedes the outcome. We would thus expect, in an ideal scenario, even with a single exposure and a single mediator, that we would have data from three different time periods which, in most contexts, would then be referred to as longitudinal data. Therefore, what distinguishes the methods in this chapter from those in the previous chapters is not so much longitudinal data conceived of as data from multiple time periods, but rather a setting when the exposures and mediators themselves vary over time with multiple time measurements on each.

Some of the difficulty here, as in Chapter 5, is that natural direct and indirect effects are often not identified from the data in many settings involving time-varying exposures and mediators. As before, whenever there is a mediator–outcome confounder affected by the exposure, these natural direct and indirect effects are not identified irrespective of whether data are available on this exposure-induced confounder or not (Avin et al., 2005). This will be a common occurrence whenever the confounding variables also vary over time. Here we will present an

approach for estimating controlled direct effects when the exposure and mediator vary over time, even when there is time-varying confounding as well. We will also present an approach with time-varying exposures and mediators to estimate a randomized interventional analogue of natural direct and indirect effects, similar to the randomized interventional analogues we considered in Chapter 5 but now applicable to time-varying exposures and mediators as well. These effects will be identified even when there is exposure-induced mediator–outcome confounding as will generally be the case when the confounding variables vary over time. However, when there is no such exposure-induced mediator–outcome confounding, the same approach will be applicable to estimating the longitudinal equivalents of the natural direct and indirect effects themselves.

6.1. NOTATION AND DEFINITIONS

Suppose now that the exposure, mediators and possibly confounding variables vary over time. Let $(A(1),\ldots,A(T))$ denote values of the exposure as it varies over time at periods $1,\ldots,T$. Likewise let $(M(1),\ldots,M(T))$, and $(L(1),\ldots,L(T))$ denote the values of the mediator and time-varying confounders at periods $1,\ldots,T$. Let C denote baseline covariates that occur before the first exposure period. We assume a subsequent temporal ordering $A(t)$, $M(t)$, $L(t)$ as in Figure 6.1. We will revisit this question of temporal ordering again later on, however. For any variable W, let $\overline{W}(t) = (W(1),\ldots,W(t))$ denote the history of that variable up through time t and we let $\overline{W} = \overline{W}(T) = (W(1),\ldots,W(T))$ denote the entire history of the variable through the final period T. Let $\underline{W}(t) = (W(t),\ldots,W(T)$ denote the history of the variable from time t through the final period T.

Note that if the entire vector $A = (A(1),\ldots,A(T))$ is taken as the exposure and $M = (M(1),\ldots,M(T))$ is taken as the mediator, then the variable $L(1)$ is itself affected by the exposure (namely, by $A(1)$) and in turn confounds the mediator–outcome relationship between $M(2)$ and Y. So, as in other such settings when we have a mediator–outcome confounder affected by the exposure, natural direct and indirect effects will not be identified (Avin et al., 2005). However, in the next section we will consider the estimation of controlled direct effects in this setting and in the following section we will consider the estimation of randomized interventional analogues of natural direct and indirect effects which may still be identified.

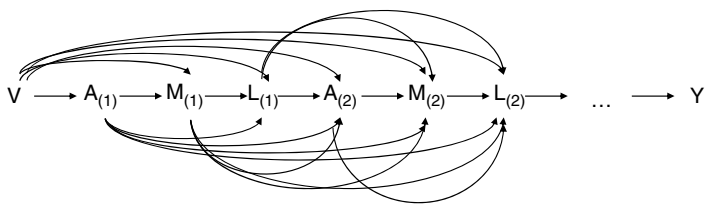

Figure 6.1 Time-varying exposures, mediators, and confounders with temporal ordering $A(t), M(t), L(t)$.

6.2. CONTROLLED DIRECT EFFECTS WITH TIME-VARYING EXPOSURES AND MEDIATORS

6.2.1. Marginal Structural Models for Controlled Direct Effects with Time-Varying Exposures and Mediators

To accommodate these settings with mediator–outcome confounders affected by prior exposure we will, as we did in Chapter 5, use a weighting estimation approach to marginal structural models. We will here, as we did there, address confounding not by covariate control in a regression but by weighting. Now, however, we will have multiple time periods and thus multiple sets of weights for the exposure and for the mediator. Doing so will allow us to estimate controlled direct effects.

As before, the controlled direct effect will compare two different levels of the exposure while in both scenarios fixing the mediator to some particular value. Now, however, the two different levels of the exposure that will be compared are not just exposure levels at a single point in time but entire exposure trajectories or histories. We will denote the two exposure histories we are comparing by $\overline{a}$ and $\overline{a}^*$. For example, if we had three measures of the exposure over time and the exposure at each time were binary, then $\overline{a}$ might correspond to being always exposed so we had $\overline{a} = (1,1,1)$ and $\overline{a}^*$ might correspond to being never exposed $\overline{a}^* = (0,0,0)$. But we could consider other comparisons as well. For example, we could compare being exposed during just the first two periods $\overline{a} = (1,1,0)$ to being exposed during just the last two periods $\overline{a}^* = (0,1,1)$, or any other comparison of exposure histories. When we consider controlled direct effects, we will be comparing these exposure histories while also fixing the mediator to some particular history or trajectory $\overline{m}$ for both of the exposure histories being compared. For example, we might be comparing our two exposure histories $\overline{a}$ and $\overline{a}^*$ while always fixing the mediator at every period to 0—for example, $\overline{m} = (0,0,0)$. Or we could choose some other mediator trajectory to fix the mediator when comparing the two exposure histories. For example, if the mediator were fixed to being present for just the last period, we would let $\overline{m} = (0,0,1)$. The controlled direct effect thus depends on what two exposure histories we are comparing, $\overline{a}$ and $\overline{a}^*$, and also what value to which we fix the mediator history $\overline{m}$. For two exposure histories , $\overline{a}$ and $\overline{a}^*$, the controlled direct effect may be different with each different level of $\overline{m}$, as would occur if there were interactions between the exposure and the mediator.

In counterfactual notation if we let $Y_{\overline{am}}$ be the counterfactual outcome if $\overline{A}$ were set to $\overline{a}$ and if $\overline{M}$ were set to $\overline{m}$, then the controlled direct effect is defined as $Y_{\overline{am}} - Y_{\overline{a}^*\overline{m}}$. As before, we will not in general be able to estimate this effect for an individual but we may, under certain assumptions, be able to estimate it on average for a population: $\mathbb{E}[Y_{\overline{am}} - Y_{\overline{a}^*\overline{m}}]$. To do so, we will again require certain no-unmeasured-confounding assumptions. However, now with longitudinal data we will require that these assumptions hold for each point in time. Specifically, we will require assumption (A6.1), which assumes that for each time t the effect of the exposure at time t, $A(t)$, on the outcome Y is unconfounded conditional on past treatment history $\overline{A}(t-1)$, past mediator history $\overline{M}(t-1)$, past confounder

history $\overline{L}(t-1)$, and the baseline confounders C. We also require assumption (A6.2), which assumes that for each time t the effect of the mediator at time t, $M(t)$, on the outcome Y is unconfounded conditional on past treatment history $\overline{A}(t)$, past mediator history $\overline{M}(t-1)$, past confounder history $\overline{L}(t-1)$, and the baseline confounders C. See the Appendix for formal counterfactual statements of these assumptions. These assumptions would be satisfied in the causal diagram in Figure 6.1. However, assumption (A6.1) would be violated if there were an unmeasured confounder U (not captured in the measured covariates $\overline{L}$) that affected exposure $A(t)$ at some time t and the outcome Y. Likewise, assumption (A6.2) would be violated if there were an unmeasured confounder U (not captured in the measured covariates $\overline{L}$) that affected the mediator $M(t)$ at some time t and the outcome Y. However, the assumptions (A6.1) and (A6.2) would not be violated if there were an unmeasured confounder U that affected the covariates $L(t)$ and Y; that is, the effect of the time-varying confounders do not themselves need to be unconfounded for the weighting approach described below to work. Also the assumptions would not be violated if there were an unmeasured variable U that affected the exposure $A(t)$ at any or all of the times t, but did not affect the outcome Y. And the assumptions would likewise not be violated if there were an unmeasured variable U that affected the mediator $M(t)$ at any or all of the times t, but did not affect the outcome.

Under assumptions (A6.1) and (A6.2), we can proceed to estimate the controlled direct effects using a weighting technique. As in Chapter 5, we will calculate weights and then we will fit a final model of the outcome on the exposure history and the mediator history, but the confounding variables will be controlled for by weighting rather than directly by regression adjustment. This will appropriately allow for control of the time-varying confounding variables (Robins et al., 2000). The final model that is fit, will, under assumptions (A6.1) and (A6.2), be a model for the counterfactual variables and will give us estimates of the controlled direct effects described above (Robins et al., 2000; van der Laan and Petersen, 2008; VanderWeele, 2009a). This model is referred to as a "marginal structural model."

The approach will be fairly similar to that described in Section 5.3.2 of the previous chapter, but we will now need weights for each point in time. As before, two sets of weights are needed, one for the exposure (now at each point in time) and one for the mediator (again now for each point in time). For each individual i in the sample, the exposure weight at time t is calculated by

$$w_i^A(t) = \frac{P\{A(t) = a_i(t) | \overline{a}_i(t-1), \overline{m}_i(t-1)\}}{P\{A(t) = a_i(t) | \overline{a}_i(t-1), \overline{m}_i(t-1), \overline{l}_i(t-1), c_i\}}$$

where $a_i(t)$, $m_i(t)$, $l_i(t)$, and c are the actual values of the exposure, the mediator, the time-varying covariates and the baseline covariates, respectively for individual i. In the denominator for the exposure weight at time t we have the probability of the individual receiving the exposure that was in fact received at time t, conditional on the individual's past exposure history, mediator history, time-varying covariate history, and baseline covariates. For a binary exposure, this probability could be estimated, for example, by using a logistic regression model (see code below). The numerator for the exposure weight at time t is the probability of the individual

receiving the exposure that was in fact received at time t conditional on just the individual's past exposure history and mediator history (not the covariates). For a binary exposure, this probability could also be fit by logistic regression. The ratio of these two probabilities is our weight for the exposure at time t. We calculate a different weight for each individual i at each time t. Note that the covariates only appear in the denominator, not the numerator; by weighting a final regression model using these weights, this will control for the covariates appropriately even in the presence of time-varying confounding variables (Robins et al., 2000). To use the weighting approach for marginal structural models, we also need a second set of weights, now for the mediator. For each individual i in the sample, the mediator weight at time t is calculated by

$$w_i^M(t) = \frac{P\{M(t) = m_i(t) | \overline{a}_i(t), \overline{m}_i(t-1)\}}{P\{M(t) = m_i(t) | \overline{a}_i(t), \overline{m}_i(t-1), \overline{l}_i(t-1), c_i\}}$$

where again $a_i(t)$, $m_i(t)$, $l_i(t)$, and c are the actual values of the exposure, the mediator, the time-varying covariates, and the baseline covariates, respectively, for individual i. The denominator is the probability of the individual having the mediator value that was in fact present at time t conditional on the individual's past exposure history, mediator history, time-varying covariate history, and baseline covariates. For a binary mediator, this probability could again be estimated using a logistic regression model. The numerator is the probability of the individual having the mediator value that was in fact present at time t conditional on just the individual's past exposure history and mediator history (not the covariates). For a binary mediator, this probability could also be fit by logistic regression. Again we calculate a different mediator weight for each individual i at each time t.

Once these weights are estimated, we can obtain an overall weight for the individual by taking a product of these weights across all times:

$$w_i = w_i^A(1) \times \cdots \times w_i^A(T) \times w_i^M(1) \times \cdots \times w_i^M(T)$$

We can then employ the marginal structural model approach to controlled direct effects (van der Laan and Petersen, 2008; VanderWeele, 2009a) by fitting a regression model of Y on $\overline{A}$ and $\overline{M}$. This model will, when weighted by the weights above w_i, and under assumptions (A6.1) and (A6.2) about confounding, be a model for the expected counterfactual outcomes $\mathbb{E}[Y_{\overline{am}}]$. Considerable flexibility is allowed in specifying the dependence of the counterfactual outcomes $\mathbb{E}[Y_{\overline{am}}]$ on $\overline{A}$ and $\overline{M}$. We could, for example, specify a linear link or a logistic link for this model. We could allow for a separate coefficient for each exposure period and each mediator period or we could specify the dependence to be only for the cumulative sum of all exposure periods $\text{cum}(\overline{A}) = \sum_{t=1}^{T} a(t)$ or all mediator periods $\text{cum}(\overline{M}) = \sum_{t=1}^{T} m(t)$ or some other function. We could allow for interactions and quadratic terms. Whatever form is decided upon for the model, a regression model is fit for the outcome Y only on the exposure and mediator variables (not the confounders) but the regression is weighted by the weights w_i to control for baseline and time-dependent confounding.

Suppose, for example, that there are only two exposure periods and two mediator periods and that the outcome Y is continuous. We might then specify the marginal structural model by having a different term for each exposure and mediator period:

$$\mathbb{E}[Y_{\overline{am}}] = \gamma_0 + \gamma_1 a(1) + \gamma_2 a(2) + \gamma_3 m(1) + \gamma_4 m(2)$$

To fit this marginal structural model, we would simply regress Y on the two exposure variables over time, $a(1)$ and $a(2)$, and the two mediator variables over time, $m(1)$ and $m(2)$, using the regression model

$$\mathbb{E}[Y|a,m] = \gamma_0 + \gamma_1 a(1) + \gamma_2 a(2) + \gamma_3 m(1) + \gamma_4 m(2)$$

but weighting each individual in this regression by the weights that were calculated, w_i. Under the assumption that our models are correctly specified and that the no-unmeasured-confounding assumptions (A6.1) and (A6.2) hold, the coefficients of in this weighted regression will estimate the parameters of the marginal structural model. The controlled direct effect can then be obtained based on the resulting estimates from this final weighted regression. For example, after fitting this marginal structural model, we would have that the controlled direct effect is given by

$$\begin{aligned}\mathbb{E}[Y_{\overline{am}}] - \mathbb{E}[Y_{\overline{a^*m}}] &= \gamma_0 + \gamma_1 a(1) + \gamma_2 a(2) + \gamma_3 m(1) + \gamma_4 m(2) \\ &\quad -[\gamma_0 + \gamma_1 a^*(1) + \gamma_2 a^*(2) + \gamma_3 m(1) + \gamma_4 m(2)] \\ &= \gamma_1[a(1) - a^*(1)] + \gamma_2[a(2) - a^*(2)]\end{aligned}$$

As noted above, we could also consider other specifications of the marginal structural model which depend only on the cumulative total of the exposure and the mediator, for example,

$$\mathbb{E}[Y_{\overline{am}}] = \gamma_0 + \gamma_1 \text{cum}(\overline{A}) + \gamma_2 \text{cum}(\overline{M})$$

or we could include interaction terms between exposures at different times or between exposures at particular times and mediators at particular times. We could also specify a logistic model instead if our outcome is binary. In each case, we again just fit a regular regression for our outcome Y on the exposure and mediator terms that have been chosen, but we weight the regression by the weights w_i that have been calculated. As in Chapter 5, to obtain standard errors, when we use the weighted regression we must calculate what are sometimes called "robust" or "sandwich" standard errors. This takes into account weighting approach that has been used. Most software packages allow for this and code for this is given below.

If the exposure and mediator are binary, then the predicted probabilities in the denominator and numerator for the weights for A and M could be fit by logistic regression as discussed above. If one or both of the exposure or mediator are categorical or ordinal, then the numerator and denominator probabilities could be estimated by using a multinomial or ordinal logistic regression. If the exposure or mediator is continuous, the probabilities can be replaced by values from a probability density function; see Robins et al. (2000) for further description. However,

the performance of this marginal structural model technique is usually not particularly good if the exposure and mediator are continuous. If there are many exposure and mediator periods, not just two, then when we form the weights, we might not regress the current exposure or mediator at time t on the entire exposure, mediator, and covariate history, but perhaps just on the previous period or the previous two periods. Sometimes, when fitting the final marginal structural model, the baseline covariates C (but not the time-varying covariates) are included in the final regression model. When this is done, then the baseline covariates also need to be included in the numerator of the exposure and mediator weights (and not just the denominator of the weights; though the time-varying covariates are still included only in the denominator). This can sometimes lead to more stable effect estimates. Further details on fitting marginal structural models are given in Robins et al. (2000) and Hernán and Robins (2015).

6.2.2. SAS Code for Marginal Structural Models for Controlled Direct Effects with Time-Varying Exposures and Mediators

In this section we give SAS code to implement this marginal structural model approach to controlled direct effects with time-varying exposures and mediators. We will give a code assuming two time periods for the exposure and two time periods for the mediator with both the exposure and the mediator binary. However, the code could fairly easily be extended to more than two time periods. The code could also fairly easily be adapted to other software packages. Suppose that the exposure $A(t)$ and mediator $M(t)$ at each time t are binary and suppose that the name of the dataset was "mydata", the name of the exposure variable at times 1 and 2 were "a1" and "a2" respectively, the mediator variables "m1" and "m2", the outcome "y," and time-varying covariates "l1" and "l2" and that we have three baseline covariates "c1," "c2," and "c3." We could then first fit the models for the weights using the following code:

```
proc logistic data=mydata descending;
   model a1= ;
   output out=mydata predicted=pna1;
run;

proc logistic data=mydata descending;
   model a1= c1 c2 c3;
   output out=mydata predicted=pda1;
run;

proc logistic data=mydata descending;
   model m1 = a1;
   output out=mydata predicted=pnm1;
run;

proc logistic data=mydata descending;
   model m1 = c1 c2 c3 a1;
   output out=mydata predicted=pdm1;
run;
```

```
proc logistic data=mydata descending;
   model a2= a1 m1;
   output out=mydata predicted=pna2;
run;

proc logistic data=mydata descending;
   model a2= c1 c2 c3 a1 m1 l1;
   output out=mydata predicted=pda2;
run;

proc logistic data=mydata descending;
   model m2 = a1 m1 a2;
   output out=mydata predicted=pnm2;
run;

proc logistic data=mydata descending;
   model m2 = c1 c2 c3 a1 m1 l1 a1;
   output out=mydata predicted=pdm2;
run;
```

The first two logistic regressions are for the exposure weight at time 1 (the numerator and denominator probabilities respectively); the second two logistic regressions are for the mediator weight at time 1; the third two logistic regressions are for the exposure weight at time 2; and the final two logistic regressions are for the mediator weight at time 2. In each case the "predicted" command will predict the probability, based on the individuals' actual past values of the variables, that the exposure or the mediator takes the value 1.

To obtain the weights we need to take these predicted probabilities for the exposure and mediator being 1 and turn them into the probability of the individual having the exposure that they in fact had. In other words, for the exposure numerator and denominator probabilities, if the individual did actually have the exposure then we can just take the predicted probabilities themselves. If the individual did not have the exposure then we want 1 minus these predicted probabilities (i.e. the predicted probability that the individual did not have the exposure). Likewise for the mediator numerator and denominator probabilities, if the individual actually had a value of the mediator $M = 1$ then we can just take the predicted probabilities themselves. If the individual had a value of the mediator $M = 0$ then we want 1 minus these predicted probabilities (i.e. the predicted probability that the individual had mediator level $M = 0$). The code below thus calculates the appropriate predicted probabilities and takes the ratio of the numerator and denominator probabilities for the exposure to form the exposure weights (wa1 for time 1 and wa2 for time 2) and takes the ratio of the numerator and denominator probabilities for the mediator to form the mediator weights (wm1 for time 1 and wm2 for time 2) and then takes the product of the exposure weights at times 1 and 2 and the mediator weights at times 1 and 2 to form the overall weights (ww).

```
data mydata;
   set mydata;
   if a1=1 then wa1=pna1/pda1; else wa1=(1-pna1)/(1-pda1);
   if a2=1 then wa2=pna2/pda2; else wa2=(1-pna2)/(1-pda2);
```

```
   if m1=1 then wm1=pnm1/pdm1; else wm1=(1-pnm1)/(1-pdm1);
   if m2=1 then wm2=pnm2/pdm2; else wm2=(1-pnm2)/(1-pdm2);
   ww=wa1*wa2*wm1*wm2;
run;
```

Finally, once we have calculated the weights, we can simply regress the outcome "y" on the exposure variables "a1" and "a2" and the mediator variables "m1" and "m2" where we weight each individual by the weights we have calculated. The final line of the procedure requests that "robust" or "sandwich" standard errors be calculated, which is necessary to obtain valid standard errors to take into account the weighting.

```
proc genmod data=mydata;
   class caseid;
   model y = a1 a2 m1 m2 / error=normal link=id;
   weight ww;
   repeated subject = caseid/ type = unstr;
run;
```

This will then give us the parameters of the marginal structural model that we can use to calculate the controlled direct effect. If the outcome were binary and we wanted to obtain a controlled direct effect odds ratio, we could simply replace "error=normal link=id" with "error=binomial link=logistic." The code above assumes just one time-varying variable L over times 1 and 2, but the code could also easily be adapted to include more than one time-varying confounder; for example, if we had a second time-varying confounder, call it V, with measures at times 1 and 2 of "v1" and "v2," we could instead put "l1 v1" in the regressions wherever "l1" occurs, and so on, so that both time-varying covariates are included.

The code above applies to binary exposures and mediators. If the exposure or mediator were categorical or ordinal, we would have to replace the logistic regressions for the predicted probabilities by multinomial or ordinal logistic regressions and obtain for each individual the predicted probabilities of their having the exposure or mediator that was in fact received. If the exposure A or the mediator M is continuous, then we have to replace the probabilities for the weights obtained by logistic regression with probability density functions obtained from linear regression, but again this marginal structural model for controlled direct effects approach will in general be somewhat less stable for continuous exposures and mediators. This approach above can also be used if multiple outcomes Y over time are also observed, but then the overall weights for the weighted regression model need to be different for each point in time and the final regression model that is fit to the data will be a repeated measures model, weighted by the time-varying weights. Intermediate values of the outcome can often then serve as a time-varying confounder and are adjusted for in the estimation of the weights. See VanderWeele (2009a) for using repeated measures marginal structural models for the estimation of controlled direct effects, and see Hernán et al. (2002) for a description of fitting repeated measures marginal structural models more generally.

Thus far we have assumed that the order of the variables was C, $A(1)$, $M(1)$, $L(1)$, $A(2)$, $M(2)$, $L(2)$, and so on, as in Figure 6.1. However, we could easily adapt this approach to other orderings. Suppose, for example, that the ordering was in fact

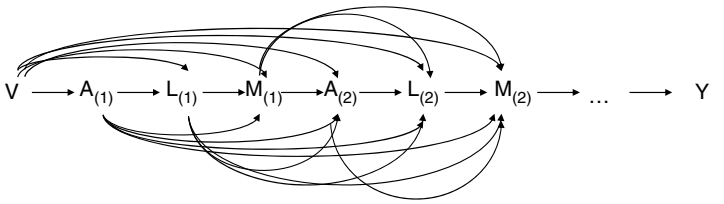

Figure 6.2 Time-varying exposures, mediators, and confounders with temporal ordering $A(t), L(t), M(t)$.

$C, A(1), L(1), M(1), A(2), L(2), M(2)$, and so on, so that $L(t)$ occurred between $A(t)$ and $M(t)$ as in Figure 6.2. We could use the same approach as that presented above, but in the denominator mediator regressions we would now also include $L(1)$ in the regression for $M(1)$ (i.e., in the fourth regression in the code above for the weights we would have "model m1 = c1 c2 c3 a1 l1" instead of simply "model m1 = c1 c2 c3 a1") and we would include $L(2)$ in the regression for $M(2)$ (i.e., in the eighth regression in the code above for the weights we would have "model m2 = c1 c2 c3 a1 m1 l1 a1 l2" instead of simply "model m2 = c1 c2 c3 a1 m1 l1 a1"). The basic principle is simply that to obtain the denominator weight for the exposure or mediator at each time t, we regress that variable on all of the past. If $L(t)$ occurs after $M(t)$, as in Figure 6.1, we would not include it in the regression for $M(t)$; we would only include the past covariates $\overline{L}(t-1)$. If $L(t)$ occurs before $M(t)$, as in Figure 6.2, we would include it in the regression for $M(t)$ as well. Using this same principle, we could likewise accommodate two different sets of time-varying confounding variables, one set what occurs between the exposure at time t and the mediator at time t and another set that occurs between the mediator at time t and the exposure at time $t+1$. Again, the principle is just that to obtain the weights we regress the exposure or mediator variable on all of the past.

6.2.3. Example of Controlled Direct Effects with Time-Varying Exposures and Mediators

Here we present an analysis of Mumford et al. (2011) to examine the controlled direct effects of dietary fiber intake on cholesterol, not through its effects on estrogen levels. Fiber intake (the exposure A) tends to decrease both estrogen levels (the intermediate M) and LDL cholesterol levels (the outcome Y) but lower estrogen levels tend to increase LDL cholesterol levels. There is thus potential interest in what the magnitude of the direct effects of fiber intake on cholesterol levels would be if estrogen levels were to remain fixed.

The data come from the BioCycle Study and consist of 259 healthy premenopausal women aged 18 to 44 from the western New York region who are followed up for two menstrual cycles with eight clinical visits per cycle timed, using fertility monitors, according to biologically relevant windows of the menstrual cycle. Measures of estrogen (estradiol), luteinizing hormone (LH), and follicle stimulating hormone (FSH) were measured in fasting serum samples collected at

each visit. The cholesterol outcome measures were total cholesterol, HDL cholesterol, LDL cholesterol, and triglycerides that were also obtained at each visit. Dietary measures, including fiber intake, energy intake, and vitamin E intake, were also obtained. Since there were no significant differences in dietary fiber intake across phases of the cycle, the average daily fiber intake per cycle was taken as the exposure variable. There was considerable evidence of a threshold effect, and fiber intake was thus dichotomized at more versus less than 22 grams per day. Baseline covariates included age, body mass index, and physical activity.

The exposure, fiber intake, itself then is time-varying with two measures (one for each of the two cycles), and the mediator, estrogen levels, which vary considerably across the cycle, was also time-varying with (16) measures (8 for each cycle). Luteinizing hormone (LH), follicle stimulating hormone (FSH), energy intake and vitamin E intake were taken as time-varying confounders that could be affected by past levels of fiber intake and estrogen and could also affect subsequent levels of estrogen and cholesterol. Intermediate values of cholesterol would likewise be affected by fiber intake and estrogen and affect subsequent levels.

The fiber intake exposure was binary, and therefore logistic regression models were used at each period to obtain the exposure weights at each period by regressing fiber intake on age, BMI, energy intake, vitamin E intake, physical activity, LH, and FSH levels, as well as past measurements of fiber and estrogen. The estrogen measures were continuous, and linear regression models assuming normally distributed weights were used to obtain predicted probability density weights as the weights for the mediator at each time (as indicated in Section 6.2.1). Estrogen level was thus likewise regressed at each period on age, BMI, energy intake, vitamin E intake, physical activity, LH, and FSH levels, as well as past measurements of fiber and estrogen to obtain these weights. In each of the models, the dependent variable was regressed on the most recent values of the other variables.

A final log-linear repeated measures marginal structural model was fit for each of the cholesterol outcomes (total, HDL, LDL, triglycerides) on fiber and estrogen. Evidence for exposure–mediator (i.e., fiber–estrogen) interaction was assessed for each of the outcomes. Only for triglycerides was there substantial evidence for such interaction. Controlled direct effects were calculated for different fixed levels of estrogen corresponding to 30, 45, or 110 pg/mL. Fixing estrogen levels to a specified level could, for example, be done by prescribing an oral contraceptive pill. Here we will present the analysis for LDL cholesterol and triglycerides. See Mumford et al. (2011) for a full description and interpretation of the results. For LDL cholesterol, the estimates of the controlled direct effect of fiber intake on LDL cholesterol with estrogen fixed to 30, 45, or 110 pg/mL, respectively, were, on a percentage change scale, -6.2% (95% CI: $-11.1\%, -1.5\%$), -6.2% (95% CI: $-11.1\%, -1.8\%$), and -6.4% (95% CI: $-11.0\%, -2.0\%$). In all cases there was a significant controlled direct effect, but this did not substantially differ across fixing estrogen to different levels. Again there was little evidence of exposure–mediator interaction here. Here, the total effect of fiber intake on LDL cholesterol was -5.6% (95% CI: $-9.7\%, -1.9\%$). Note that each of the controlled direct effects is somewhat larger than the total effect since the controlled direct effects correspond to blocking the effect of fiber intake on estrogen by fixing estrogen to a specified

level. Thus the effect of high fiber intake on lowering estrogen levels (and thereby also slightly increasing LDL cholesterol) is blocked and the controlled direct effects are therefore slightly higher than the total effect.

For triglycerides (where there was evidence of interaction), the estimates of the controlled direct effect of fiber intake on triglycerides with estrogen fixed to 30, 45, or 110 pg/mL respectively, were, on a percentage change scale, −7.8% (95% CI: −18.4%, −2.7%), −5.3% (95% CI: −15.1%, 4.4%), and −0.4% (95% CI: −9.2%, 9.9%). Here the controlled direct effect with estrogen fixed to 30 pg/mL is significant. When estrogen is fixed to 45 pg/mL, the effect is somewhat smaller, and when estrogen is fixed to 110 pg/mL, it is almost exactly 0. See Mumford et al. (2011) for further discussion of the results and details of the analysis.

6.3. NATURAL DIRECT AND INDIRECT EFFECTS AND THEIR RANDOMIZED INTERVENTIONAL ANALOGUES WITH TIME-VARYING EXPOSURES AND MEDIATORS

The methods in the previous section allow us to consider flexibility in estimating controlled direct effects even with time-varying exposures and mediators. However, the approach presented in the previous section cannot be used to decompose a total effect into a direct and a mediated effect. Throughout this book we have been using natural direct and indirect effects for this purpose. However, as we have already noted, we cannot, in general, directly use these concepts with longitudinal data and time-varying exposure and mediators because we will likely, whenever we have time-varying confounders as well, be in a setting with a mediator–outcome confounder affected by prior exposure, and natural direct and indirect effects are not identified in this setting. We will, however, instead be able to employ randomized interventional analogues to natural direct and indirect effects, as we did in Chapter 5, to help overcome this difficulty. We will give a conceptual overview of the relevant ideas, but practical methods and software remain to be developed.

We will let $\overline{G}_{\overline{a}|c}(t)$ denote a random draw from the distribution of the mediator $\overline{M}(t)$ that would have been observed in the population with baseline covariates $C = c$ if exposure status $\overline{A}$ had been fixed to $\overline{a}$. We let $\overline{a}$ and $\overline{a}^*$ be two distinct exposure histories that we are comparing. As a measure of the mediated effect, we could use $\{\mathbb{E}(Y_{\overline{a}\overline{G}_{\overline{a}|c}}|c) - \mathbb{E}(Y_{\overline{a}\overline{G}_{\overline{a}^*|c}}|c)\}$. This is the effect on the outcome of randomly assigning an individual with an exposure trajectory $\overline{a}$ to a trajectory of the mediator from the distribution of amongst those with exposure trajectory $\overline{a}$ versus exposure trajectory $\overline{a}^*$ (conditional on baseline covariates $C = c$); this is an effect through the mediator. For it to be nonzero, the change in the distribution of exposure trajectories from $\overline{a}^*$ to $\overline{a}$ would have to affect the mediator trajectory, and this change in the mediator trajectory would have to affect the outcome. As in Chapter 5, we will refer to this randomized interventional analogue of the natural indirect effect (now with time-varying exposures and mediators) as NIE^R. As a measure of the direct effect, we would use $\{\mathbb{E}(Y_{\overline{a}\overline{G}_{\overline{a}^*|c}}|c) - \mathbb{E}(Y_{\overline{a}^*\overline{G}_{\overline{a}^*|c}}|c)\}$. This effect compares the exposure trajectory $\overline{a}$ to $\overline{a}^*$ with the mediator trajectory in both cases randomly drawn from the distribution of mediator trajectory in the population

under exposure trajectory $\overline{a}^*$ (conditional on baseline covariates $C = c$). Again, as in Chapter 5, we will refer to this randomized interventional analogue of the natural direct effect (now with time-varying exposures and mediators) as NDE^R. Finally, $\mathbb{E}(Y_{\overline{a}\overline{G}_{\overline{a}|c}(t)}|c) - \mathbb{E}(Y_{\overline{a}^*\overline{G}_{\overline{a}^*|c}}|c)$ is the effect comparing the expected outcome when having the exposure trajectory $\overline{a}$ with the mediator randomly drawn from the distribution of the population under exposure trajectory $\overline{a}$ (conditional on baseline covariates $C = c$) to the expected outcome when having the exposure trajectory $\overline{a}^*$ with the mediator randomly drawn from the distribution of the population under exposure trajectory $\overline{a}^*$ (conditional on baseline covariates $C = c$). We will refer to this effect as TE^R, the randomized interventional analogue of the total effect. We can, as before, decompose this total effect into the sum of the randomized interventional analogues of the direct and indirect effects: $TE^R = NIE^R + NDE^R$.

Under assumptions over time about no unmeasured confounding, we can also identify these randomized interventional analogues of the natural direct and indirect effects from the data. To do so, we need assumptions (A6.1) and (A6.2) above; that is, we require that [assumption (A6.1)] for each time t the effect of the exposure at time t, $A(t)$, on the outcome Y is unconfounded conditional on past treatment history $\overline{A}(t-1)$, past mediator history $\overline{M}(t-1)$, past confounder history $\overline{L}(t-1)$, and the baseline confounders C and that [assumption (A6.2)] for each time t the effect of the mediator at time t, $M(t)$, on the outcome Y is unconfounded conditional on past treatment history $\overline{A}(t)$, past mediator history $\overline{M}(t-1)$, past confounder history $\overline{L}(t-1)$, and the baseline confounders C. Moreover, for these randomized interventional analogues of natural direct and indirect effects we require one further assumption about exposure–mediator confounding, namely that [assumption (A6.3)] for each time t the effect of the exposure at time t, $A(t)$, on the mediator $M(t)$ is unconfounded conditional on past exposure history $\overline{A}(t-1)$, past mediator history $\overline{M}(t-1)$, past confounder history $\overline{L}(t-1)$, and the baseline confounders C. Under these three assumptions, (A6.1)–(A6.3), the randomized interventional analogues of natural direct and indirect effects are identified. These assumptions would hold on the causal diagram in Figure 6.1. As before, the assumptions would not be violated if there were an unmeasured confounder U that affected the covariates $L(t)$ and Y (the effect of the time-varying confounders do not themselves need to be unconfounded). They would not be violated if there were an unmeasured variable U that affected the exposure $A(t)$ at any or all of the times t, but affected neither the mediator nor the outcome Y, or if there were a different unmeasured variable U that affected the mediator $M(t)$ at any or all of the times t, but did not affect the outcome. However, one setting in which controlled direct effects would be identified but the randomized interventional analogues of natural direct and indirect effects would not be identified [i.e., where assumptions (A6.1) and (A6.2) would hold but (A6.3) would not], would be if there were an unmeasured variable U that affected the exposure $A(t)$ at some (or all) time t and also affected the mediator at some subsequent time after t. Even if U did not affect the outcome, assumption (A6.3) would be violated because this would constitute exposure–mediator confounding. If such a U did not affect the outcome, assumptions (A6.1) and (A6.2) would still hold and so we could still identify controlled direct effects and use the estimation approach in the previous

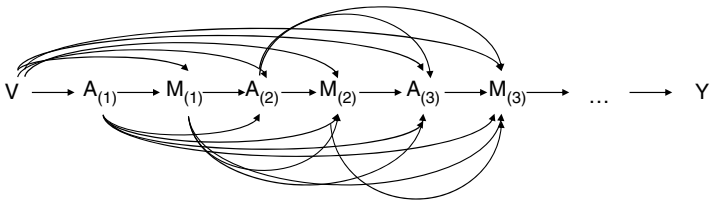

Figure 6.3 Time-varying exposures and mediators, with no time-varying confounders.

section but the randomized interventional effects would not be identified. As discussed in the Appendix, when there are no time-varying confounders as in Figure 6.3, then the natural direct and indirect effects themselves are identified, and under the causal diagram (Pearl, 2009) in Figure 6.3, the natural direct and indirect effects will be equal to the randomized interventional analogues. The Appendix gives formal identification formulae for these randomized interventional analogues of the natural direct and indirect effects when assumptions (A6.1)–(A6.3) hold as under Figures 6.1 and 6.2 and for natural direct and indirect effects under Figure 6.3, but practical methods and software to implement this remain to be developed.

The ideas and concepts in this section concerning randomized interventional analogues of natural direct and indirect effects can thus in principle be employed in the context of time-varying exposures and mediators, and further results and more formal details are given in the Appendix. Although the concepts carry over in a reasonably straightforward way from Chapter 2, the statistical modeling of such time-varying variables in the context of mediation and direct and indirect effects is challenging, and statistical methods and software to estimate these direct and indirect effects in the context of time-varying exposures and mediators require development. To illustrate the utility of these ideas, however, we will revisit a three-wave longitudinal model for mediation proposed by MacKinnon (2008) to show how the ideas in this section give insight into the interpretation of direct and indirect effects estimates in a longitudinal data setting.

6.4. COUNTERFACTUAL ANALYSIS OF MACKINNON'S THREE-WAVE MEDIATION MODEL

MacKinnon (2008) considered a three-wave mediation model with linear structural equations as depicted in Figure 6.4. We relabel indices somewhat to correspond to the notation of this chapter, and also add a set of baseline covariates C, but otherwise the model considered here is MacKinnon's model (MacKinnon, 2008, pp. 204–206, Autoregressive Model III). We let $A(0)$, $M(0)$, and $Y(0)$ denote baseline values of A, M, and Y that could be included in the baseline covariates C but are given here to make clearer the relation with MacKinnon (2008). Consider then the following regression models:

$$\mathbb{E}[M(1)|m(0),y(0),\overline{a}(1),c] = \beta_{10} + \beta_{11}a(0) + \beta_{12}a(1) + \beta_{13}m(0) + \beta_{14}y(0) + \beta_{15}'c$$

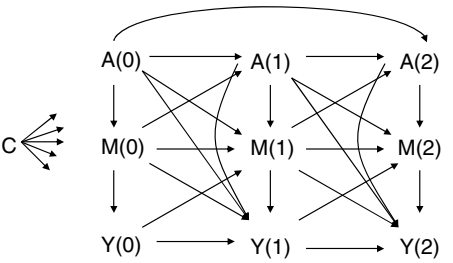

Figure 6.4 MacKinnon's (2008) three wave mediational model.

$$\begin{aligned}
\mathbb{E}[M(2)|\overline{m}(1),\overline{y}(1),\overline{a}(2),c] &= \beta_{20}+\beta_{21}a(1)+\beta_{22}a(2)\\
&\quad +\beta_{23}m(1)+\beta_{24}y(1)+\beta'_{25}c\\
\mathbb{E}[Y(1)|\overline{m}(1),y(0),\overline{a}(1),c] &= \theta_{10}+\theta_{11}a(0)+\theta_{12}a(1)+\theta_{13}m(0)\\
&\quad +\theta_{14}m(1)+\theta_{15}y(0)+\theta'_{16}c\\
\mathbb{E}[Y(2)|\overline{m}(2),\overline{y}(1),\overline{a}(2),c] &= \theta_{20}+\theta_{21}a(1)+\theta_{22}a(2)\\
&\quad +\theta_{23}m(1)+\theta_{24}m(2)+\theta_{25}y(1)+\theta'_{26}c
\end{aligned}$$

Note that in these models, the mediator and the outcome depend only on the two most recent past exposure values. The mediator model depends only on the most recent past mediator value and the most recent past outcome value. The outcome model depends on the two most recent mediator values and the most recent outcome value.

We show in the Appendix that if we take $Y(2)$ as the final outcome then under assumptions (A6.1)–(A6.3) with baseline covariates $(C,A(0),M(0),Y(0))$, with two intervention periods, $A(1)$ and $A(2)$, and $L(1)=Y(1)$, the randomized interventional analogues of the natural direct and indirect effects are given by

$$\begin{aligned}
\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|v}}|v)-\mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|v}}|v) &= (\theta_{21}+\theta_{12}\theta_{25})[a(1)-a^*(1)]\\
&\quad +\theta_{22}[a(2)-a^*(2)]\\
\{\mathbb{E}(Y_{\overline{a}G_{\overline{a}|v}}|v)-\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|v}}|v)\} &= \{\theta_{23}\beta_{12}+\theta_{25}\theta_{14}\beta_{12}+\beta_{21}\theta_{24}+\beta_{24}\theta_{12}\theta_{24}\}\\
&\quad \times[a(1)-a^*(1)]+\beta_{22}\theta_{24}[a(2)-a^*(2)]
\end{aligned}$$

The first expression is the randomized interventional analogue of the natural direct effect, and the second expression is the randomized interventional analogue of the natural indirect effect. A proof of this is given in the Appendix. The causal mediation framework allows for a clear causal interpretation of the estimates of the direct and indirect effects.

There is arguably a twofold advantage of using data like those in Figure 6.4, along with using a modeling approach like that described above, over simply applying the methods in Chapter 2, say, to one wave of data e.g. $A(1),M(1),Y(1)$. First, by having multiple waves of data, we can control for baseline levels of the exposure, mediator and outcome—that is, for $A(0),M(0),Y(0)$. This is potentially important because such baseline values of the exposure, mediator, and outcome may serve as

the most important confounders for the effects of subsequent values of exposure and mediator on the outcome. By including such baseline values of the exposure, mediator, and outcome, in our covariate set, our confounding assumptions required for a causal interpretation of our estimates are rendered much more plausible. This was discussed in Chapter 2 and we can see the point more fully here. Second, by using multiple waves of subsequent exposure and mediator and outcome data [i.e., by using $A(1), M(1), Y(1), A(2), M(2), Y(2)$ rather than just $A(1), M(1), Y(1)$ and by using the formulae above rather than those in Chapter 2], we may be able to more fully capture the dynamics of mediation over time. For example, we can pick up, in our indirect effect estimates, mediated effects of $A(1)$ through $M(1)$ to $Y(2)$ directly and also those from $A(1)$ through $M(1)$ to $Y(1)$ to $Y(2)$ or from $A(1)$ to $M(2)$ to $Y(2)$, and so on.

Here we have given a counterfactual analysis of one specific mediational model with three waves of data on the exposure, mediator, and outcome (MacKinnon, 2008). A similar approach could in principle be used for other complex longitudinal models to provide counterfactual-based interpretations of direct and indirect effect estimates. However, more methodological research is needed to develop general statistical techniques and methods that can be applied across a broad range of settings. The theory, as developed in the Appendix, provides the underlying principles, but practical details for statistical implementation across general settings still remain to be worked out.

6.5. DISCUSSION

In this chapter we have considered methods for time-varying exposures and mediators. The challenge here was similar to some of those encountered in the previous chapter: mediator–outcome confounder affected by the exposure. In most contexts with time-varying exposures and mediators as considered in this chapter, we will also have time-varying confounders. If this is the case, then we will likely have a mediator–outcome confounder that is affected by prior exposure, and natural direct and indirect effects will then not be identified from the data. As we have seen in this chapter, however, we can still make progress estimating controlled direct effects in this setting, and we can moreover also still estimate randomized interventional analogues of natural direct and indirect effects that can still be useful for effect decomposition. These randomized interventional analogues do reduce to the natural direct and indirect effects where there is no mediator–outcome confounder affected by exposure (e.g., when there are no time-varying confounders), but the randomized interventional analogues can be estimated in a broader range of settings even when natural direct and indirect effects are not identified with the data. The methods in this chapter thereby extend those in previous chapters to settings with longitudinal data and exposures and mediators that vary over time. Such rich longitudinal data can potentially increase power in the analysis of direct and mediated effects and help better ensure that questions of temporality in thinking about causal effects are clearer. Much work in this area remains to be done, however.

7

Selected Topics in Mediation Analysis

In this chapter we will consider a number of other, sometimes more technical and subtle, topics in mediation analysis. We will discuss alternative approaches to the estimation and identification of causal effects. We will also consider in more detail the very definitions of the effects that we have been discussing and the assumptions that have been making to estimate them. This chapter will not be focused on providing new methods but rather will present ideas that help with questions of interpretation. Other than estimation, such topics include difficulties with ill-defined mediators, controversies over the assumptions for direct and indirect effects, direct and indirect effects in health disparities research, more on incorporating interaction into mediation analyses, alternative identification strategies for direct and indirect effects, and power and sample size calculations for direct and indirect effects. Because some of the topics in this chapter concern more technical subtleties, we will sometimes have to employ counterfactual notation directly. However, many of the topics and sections in this chapter should still be fairly comprehensible and accessible to readers unfamiliar with or uncomfortable with this notation. The sections in this chapter for the most part consist of stand-alone additional topics on mediation analysis and need not be read consecutively. The reader can choose to read some sections and not others; they are for the most part independent of one another, and there is no need to read the sections in this chapter in order from beginning to end.

7.1. OTHER ESTIMATION APPROACHES

In Chapter 2 we discussed regression-based methods to estimate direct and indirect effects. These methods involved specifying a model for the outcome and a model for the mediator and then combining the results of these models to obtain direct and indirect effects. This could be done either by analytic calculations (VanderWeele and Vansteelandt, 2009, 2010; Valeri and VanderWeele, 2013) for which macros were provided or by using simulations (Imai et al., 2010a) for which macros were

also described. The simulation approach had the advantage of flexibility insofar as when different mediator or outcome models were specified, it was not necessary to derive new formulas for each case. The analytic approach had the advantage of much faster computation times, especially in obtaining standard errors, and especially in settings with a large number of observations.

In other chapters, we also discussed some alternative approaches to mediation analysis. In Chapter 4 for survival data we discussed a weighting approach which involved specifying a model for the exposure and a model for the mediator (Lange et al., 2012). This approach allowed us to estimate direct and indirect effects in Cox proportional hazards models with a common outcome, which was a case in which closed-form analytic expressions for direct and indirect effects were not possible to obtain using models for the mediator and the outcome. As noted in Chapter 4, the approach of using exposure and mediator weights could also be employed more generally to settings with binary, continuous, or count outcomes, not just with the Cox model (Lange et al., 2012; Hong, 2010). In Chapter 5, when considering mediation analysis with multiple mediators, we considered a weighting approach that involved specifying models for the exposure and the outcome. We considered this approach because, when the outcome and one of the mediators were binary, or when there were mediator–mediator interactions, it was difficult to obtain easily generalizable analytic expressions for the direct and indirect effects. By specifying models for the outcome and exposure, instead of outcome and mediator models, and using a weighting approach, we could overcome the difficulties in estimating direct and indirect effects with multiple mediators (VanderWeele and Vansteelandt, 2013). This approach of using exposure and outcome models, rather than mediator and outcome models, could likewise be employed in simple mediation settings with a single mediator involving a binary, continuous, or count outcome (Vansteelandt et al., 2012a; Albert, 2012).

More generally then we can proceed with estimating direct and indirect effects by specifying two of the three following models:

(i) a model for the outcome conditional on the exposure, mediator and covariates
(ii) a model for the mediator conditional on the exposure and covariates
(iii) a model for the exposure conditional on the covariates

The approach described in Chapter 2 used models for (i) the outcome and (ii) the mediator (VanderWeele and Vansteelandt, 2009, 2010; Valeri and VanderWeele, 2013; Imai et al., 2010a). The weighting approach described in Chapter 4 used models for (ii) the mediator and (iii) the exposure (Hong, 2010; Lange et al., 2012). The weighting approach described in Chapter 5 used models for (i) the outcome and (iii) the exposure (Vansteelandt et al., 2012a; Albert, 2012; VanderWeele and Vansteelandt, 2013).

However, in general, our emphasis in this book has been on using models for (i) the outcome and (ii) the mediator. This is because, when standard maximum likelihood estimators are used for these models (as in the approaches we have been

describing), this approach of using models for the outcome and the mediator are more efficient (i.e., have smaller standard errors and narrower confidence intervals) than the approaches that use either exposure or mediator weighting (Vansteelandt et al., 2012a; Tchetgen Tchetgen and Shpitser, 2012).

Nonetheless, there are cases in which using one of the alternative weighting approaches may be desirable even if we can also use the approach of outcome and mediator models. For example, if the exposure is randomized, then we do not need to model its distribution; we know it. Thus we could use the known exposure distribution and we would only have to specify one other model—for example, the outcome model (Albert, 2012; Vansteelandt et al., 2012a) or the mediator model. Estimation of direct and indirect effects would then only require correct specification of one statistical model, not two, and would therefore be making fewer modeling assumptions and would be more robust. We have also seen that in some cases the approach of using an outcome model and a mediator models leads to direct and indirect effect computations that are not easily analytically tractable and then one of the alternative weighting approaches can help address this concern. This was what motivated using the weighting approach in Chapters 4 and 5. Finally, Lange et al. (2012) and Vansteelandt et al. (2012a) note that the weighting approaches may be more attractive when trying to test whether the direct and indirect effects vary with the covariates (sometimes called "moderated mediation"; cf. Muller et al., 2005). This is because when using the regression-based approach with outcome and mediator models, it can be difficult to specify these models in a way that the direct effect does not depend on the covariates. As already noted, the weighting approaches work best when whatever variable is being modeled for the weights (the exposure or mediator) is binary or categorical. When the exposure is continuous, it is best not to use an approach that requires exposure weights. When the mediator is continuous, it is best not to use an approach which requires mediator weights.

Recently, other approaches for direct and indirect effects have been proposed (Tchetgen Tchetgen and Shpitser, 2012; Zheng and van der Laan, 2012) that involve specifying models for all of (i) the outcome, (ii) the mediator, and (iii) the exposure and these approaches will consistently estimate direct and indirect effects if any two of the three models are correctly specified. Such approaches are sometimes referred to as "multiply robust" or "triply robust" approaches (they are robust to the misspecification of any one of the three models, provided that the other two are correctly specified). An earlier proposal (van der Laan and Petersen, 2008) also considered a "doubly robust" estimator that required correct specification of the mediator model and either (i) the outcome model or (iii) the exposure model. Goetgeluk et al. (2008) developed doubly robust estimators for controlled direct effects. Unfortunately, at present many of these methods are not easy to implement using standard software and, although their theoretical properties are attractive in large sample sizes, their actual performance in smaller sample sizes is sometimes not particularly good (Vansteelandt et al., 2012a). However, adaptations and modifications of these methods may later develop into implementable and useful approaches. For example, an alternative doubly robust estimator was proposed by Vansteelandt et al. (2012a) which requires the additional specification of a "natu-

ral effect model" for the counterfactuals and then consistently estimates direct and indirect effects provided either (i) a model for the outcome is correctly specified or that both (ii) a model for the mediator and (iii) a model for the exposure are correctly specified. This approach can be implemented by making use of a replicated dataset and standard software with standard errors obtained by bootstrapping. The reader is referred to Vansteelandt et al. (2012a) for further details and code in R.

7.2. ILL-DEFINED MEDIATORS AND MULTIPLE VERSIONS OF THE MEDIATOR

The definitions of direct and indirect effects that we have been discussing in this book are defined in terms of potential or counterfactual outcomes, generally conceived of as the outcome that would have arisen under various hypothetical, possibly contrary-to-fact, interventions. Recall that with, for example, a binary exposure, the controlled direct effect assesses how much the outcome would change on average if the mediator were set to level m uniformly in the population but the treatment were changed from level $a^* = 0$ to level $a = 1$. The natural direct effect expresses how much the outcome would change if the exposure were set at level $a = 1$ versus level $a^* = 0$, but for each individual the mediator were kept at the level it would have taken in the absence of the exposure. The natural indirect effect expresses how much the outcome would change on average if the exposure were controlled at level $a = 1$, but the mediator were changed from the level it would take if $a^* = 0$ to the level it would take if $a = 1$.

In the context of mediation, then, we consider possible interventions on or settings of both the exposure and the mediator. It is sometimes stated that such potential or counterfactual outcomes are well-defined only to the extent that the investigator has clearly specified the intervention in view (Robins and Greenland, 2000; Hernán, 2005). Moreover, one's estimates of causal effects will reasonably correspond to the intervention envisioned only if the exposure and mediator that naturally occurred would give rise to the same outcomes as an intervention to set the exposure and mediator to this same level (Hernán, 2005). In the technical language of the causal inference literature, this is sometimes referred to as "consistency" assumptions (Robins, 1986; cf. VanderWeele, 2009b). Rubin (1986), in his formulation of the potential outcomes model, argued that two assumptions were important for having potential outcomes well-defined: One was that there be no multiple versions of treatment; the other was that there be "no interference" in the sense that the treatment of one individual does not affect the outcomes of other individuals in the study. He called these two assumptions the "Stable Unit Treatment Value Assumption" or "SUTVA." These assumptions were necessary for the usual potential outcomes notation and concepts to be well-defined.

These assumptions are problematic in settings in which there are multiple versions of the treatment or exposure; or, within the context of mediation, when there

are multiple ways to set the mediator to a particular value if these different hypothetical interventions to fix the mediator to the same value may have very different consequences for the outcome. For example, suppose the treatment were an educational intervention, the mediator a measure of classroom quality, and the outcome some measure of average test scores. In this context, there may be multiple ways to go about changing classroom quality (e.g., giving the teacher additional coaching, replacing the teacher, providing a more adequate room for the class, etc.). Even if each of these interventions on the mediator resulted in the same level of classroom quality under some measure, they may have very different consequences for the test scores of the class. Conceived of another way, we have "multiple versions of the mediator." This creates difficulties for the definition and interpretation of direct and indirect effects in the counterfactual framework. It is not clear what is necessarily meant by mediation when there are many ways the mediator may take on a particular value and when these have very different effects on the outcome. As another example, Caffo et al. (2008) studied the extent to which the effects of cumulative lead exposure in organolead manufacturers on cognitive function scores was mediated by brain volumes. Here, with brain volumes as the mediator, it is not clear how to intervene on this variable or what set of potential hypothetical interventions we might consider. In these settings, multiple versions of the mediator create difficulties for the counterfactual interpretation of direct and indirect effects.

An approach has, however, begun to develop within the causal inference literature to facilitate thinking about causal inference in settings with multiple versions of treatment (Hernán and VanderWeele, 2011; VanderWeele, 2012a; VanderWeele and Hernán, 2013). In this section we describe this approach in relation to questions of mediation when there are multiple versions of the mediator. More theoretical aspects and results are discussed in the Appendix, and the simpler setting with multiple versions of treatment outside the context of mediation is also discussed there. Here in this section we will focus on the issue of multiple versions of the mediator. We will consider two settings: one in which there may be multiple ways to intervene on the variable taken as the mediator in the analysis, and one in which the actual mediator has been coarsened (e.g., dichotomized) so that one value of the mediator measure corresponds to multiple values of the true mediator. Many of the concepts and issues of interpretation turn out to be quite analogous in these two settings and thus we will be able to consider them together. We will not consider the settings, addressed in some of the social science literature (Bollen, 1989; MacKinnon, 2008), in which multiple measures of an underlying latent mediator are available for each person in the study.

In what follows, we will use "version" to refer to the underlying true mediator or the set of interventions to fix the mediator value and "mediator measure" to refer to the variable actually used in the analysis. We will assume that the investigator does not have access to data on "version" but proceeds with the estimation of direct and indirect effects with the data on the mediator measure available (e.g., classroom quality or brain volumes, in the examples above), and we will consider the interpretation of the direct and indirect effect estimates that result. We will describe how these effects can be interpreted as the effects that would have arisen had the version of the mediator been randomly selected from the distribution of the versions

that arise naturally in certain subpopulations. In particular, what is estimated as an indirect effect does in fact correspond to a particular mediated effect wherein the version of the mediator is randomly selected from the exposed with particular values of the measured mediator. However, what is estimated as a direct effect is in fact equal to the sum of a direct effect plus a mediated effect corresponding to the effect of the exposure on the outcome mediated by versions of the mediator that is not captured by the mediator measure itself. This is the intuitive interpretation of the effects in the context of multiple versions of the mediator; the remainder of this section describes the precise technical interpretation; the non-technical reader may prefer to move to the next section.

7.2.1. Mediation Analysis with Multiple Versions of the Mediator

Suppose now that for every possible value of the mediator, $M = m$, there are potentially multiple interventions that allow M to be fixed at level m. We will let K denote the underlying version of the mediator. Several different values of K may correspond to the actual value of M (the variable M is effectively a coarsening of K). Suppose that K is the underlying version of the mediator, but we observe only M, our measurement of the mediator, and proceed with estimating direct and indirect effects using M. A causal diagram illustrating the relationships is presented in Figure 7.1. What then is the interpretation of the direct and indirect effects that we estimate using data only on M? The next several results provide an answer to this question.

Suppose now that the measured covariates C suffice for the analogues of our four no-unmeasured-confounding assumptions (A2.1)–(A2.4) to hold with respect to the effects of A and K on Y. That is, we will assume that [assumption (A7.1)] the effect the exposure A on the outcome Y is unconfounded conditional on C, [assumption (A7.2)] the effect the version K on the outcome Y is unconfounded conditional on C, [assumption (A7.3)] the effect the exposure A on the version K is unconfounded conditional on C, and [assumption (A7.4)] there is no effect of the exposure that itself confounds the version–outcome relationship. Formal statements of these in terms of counterfactual independence are given in the Appendix. We note that it may be difficult to satisfy these no-unmeasured-confounding assumptions in practice if we do not know what precisely is the version of the mediator that may be present in a population—a point we return to again below.

In trying to understand the effects, it will be helpful to have a bit more additional notation. Let Y_{ak} be the outcome that would have occurred if exposure had been set to level a and the version of the mediator to level k. Consider the subpopulation

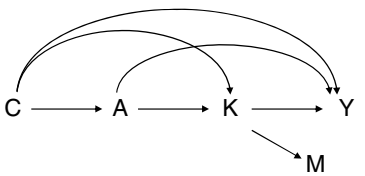

Figure 7.1 Mediation for the effect of A on Y with covariates C with multiple versions K of the mediator M.

with covariates $C = c$. Let G_a be a random draw from the distribution of the version of the mediator that involves first randomly selecting a value of M from amongst those with $A = a$ and $C = c$ and then randomly selecting a version of the mediator from among those with $M = m, A = 1$, and $C = c$. We will then let Y_{1G_a} be the value of the outcome that would arise if the exposure is set to 1 and the version of the mediator is randomly selected from the distribution G_a. Likewise we will let Y_{0G_a} be the value of the outcome that would arise if the exposure is set to 0 and the version of the mediator is randomly selected from the distribution G_a. It can be shown (VanderWeele, 2012a) that under our four no-unmeasured–confounding assumptions (A7.1)–(A7.4), the standard estimate of the natural indirect effect using data on only the mediator measurement M is equal to

$$\mathbb{E}[Y_{1G_1}|c] - \mathbb{E}[Y_{1G_0}|c]$$

In words, under the no-unmeasured-confounding assumptions (A7.1)–(A7.4), the quantity we think we are estimating as the natural indirect effect, using data only on our mediator measurement M, is in fact equal to a type of mediated effect. It is equal to the effect of fixing the exposure to $A = 1$ and then randomly drawing a version of the mediator from the distribution G_1 versus the distribution G_0, where G_1 first randomly selects a value of M from among those with $A = 1$ and $C = c$ and G_0 first randomly selects a value of M from among those with $A = 0$ and $C = c$ but then once the value of M has been selected, each then randomly selects a version of the mediator from among those with $M = m, A = 1$, and $C = c$. The expression is thus equal to a mediated effect, but with the versions of the mediator being drawn from the distribution of those with $M = m, A = 1$, and $C = c$.

The next result gives an analogue for the natural direct effect estimate but, as will be seen, the interpretation of the estimate now involves two parts. Under assumptions (A7.1)–(A7.4), the estimate of the natural direct effect using data on only the mediator measurement M is equal to

$$\big(\mathbb{E}[Y_{1G_0}|c] - \mathbb{E}[Y_{0G_0}|c]\big) + \big(\mathbb{E}[Y_{0G_0}|c] - \mathbb{E}[Y_{0H_0}|c]\big)$$

where Y_{0H_0} is the value of the outcome that would arise if the exposure is set to 0 and the version of the mediator is randomly selected from a distribution H_0 that first randomly selects a value of M from among those with $A = 0$ and $C = c$ and then randomly selects a version of the mediator from among those with $M = m, A = 0$, and $C = c$.

Contrary to the interpretation for the natural indirect effect estimate, for the natural direct effect estimate we have that if we proceed with the estimation of the natural direct effect using data only on the mediator measurement M (ignoring the different versions of the mediator), then what we think we are estimating as a natural direct effect is in fact the sum of two parts. The first part, $\big(\mathbb{E}[Y_{1G_0}|c] - \mathbb{E}[Y_{0G_0}|c]\big)$ in the expression above, is in fact a direct effect. It compares fixing exposure to $A = 1$ with fixing exposure to $A = 0$ when setting the version of the mediator to a value randomly selected from the distribution G_0. It is an effect of the exposure not through the mediator. The second part of the expression above, namely $\big(\mathbb{E}[Y_{0G_0}|c] - \mathbb{E}[Y_{0H_0}|c]\big)$, is, however, not a direct effect but a type of mediated effect. It involves comparing what the outcome would be if the exposure were fixed

to 0 and the version of the mediator were randomly set to some value from the distribution G_0 versus one from distribution H_0. Recall that G_0 involves first randomly selecting a value m of M from amongst those with $A = 0$ and $C = c$ and then randomly selecting a version of the mediator from among those with $M = m, A = 1$, and $C = c$; whereas H_0 involves first randomly selecting a value of M from among those with $A = 0$ and $C = c$ and then randomly selecting a version of the mediator from among those with $M = m, A = 0$, and $C = c$. The difference between these two distributions is in the second step wherein a version of the mediator is randomly selected from among those with $M = m, A = 1$, and $C = c$ versus those with $M = m, A = 0$, and $C = c$. Essentially then, what this second piece of the expression picks up is the effect of the exposure on outcome mediated through the portion of versions of the mediator that is not captured by mediator measurement M. The result is in some sense intuitive. If we ignore versions of the mediator in our mediation analysis, then the effect of the exposure on the outcome that is through versions of the mediator, K, but not through our mediator measure, M, will in fact be picked up by our direct effect estimate rather than our indirect effect estimate. This point should be taken into account in the estimation and interpretation of direct and indirect effects when multiple versions of the mediator exist but are not incorporated into the analysis.

A similar result also holds for the controlled direct effect. Under assumptions (A7.1) and (A7.2), the estimate of the controlled direct effect using data on only the mediator measurement M is equal to

$$\mathbb{E}[Y_{1F_1} - Y_{0F_1}|c] + \mathbb{E}[Y_{0F_1} - Y_{0F_0}|c]$$

where F_1 is a random draw of a version from the distribution $P(k|A = 1, m, c)$ and F_0 is a random draw of a version from the distribution $P(k|A = 0, m, c)$.

The first part, $\mathbb{E}[Y_{1F_1} - Y_{0F_1}|c]$ in the expression above, is a direct effect. It compares fixing exposure to $A = 1$ with fixing exposure to $A = 0$ when setting the version of the mediator to a value randomly selected from the distribution F_1—that is, of those with $A = 1, M = m$, and $C = c$. The second part of the expression, $\mathbb{E}[Y_{0F_1} - Y_{0F_0}|c]$, is a type of mediated effect. It involves comparing what the outcome would be if exposure were fixed to 0 and the version of the mediator were randomly set to some value from the distribution F_1 versus one from distribution F_0—that is, from among those with $A = 1, M = m$ and $C = c$ versus those with $A = 0, M = m$ and $C = c$.

7.2.2. Examples

We illustrate the ideas above with two examples, one when the mediator has been dichotomized and the other in which the true underlying versions of the mediator is unknown.

When questions of mediation are of interest, investigators may, out of convenience, dichotomize a potential mediator. Such dichotomization raises questions with regard to the interpretation of the resulting estimates of direct and indirect effects. Emsley et al. (2010), for example, considered mediation in the Prevention

of Suicide in Primary Care Elderly Collaborative Trial (PROSPECT). They assessed whether the effect of randomized treatment (collaborative care management versus treatment as usual), A, on the score from the Hamilton Depression Scale, Y, was mediated by adherence to antidepressants. In their analyses they dichotomized adherence. They assumed no interaction between the effect of A and M on Y and obtain an estimate of the direct effect of -2.66 (SE = 0.93) and an estimate of the indirect effect of -0.49. If there was indeed no exposure–mediator interaction and if no unmeasured confounding assumptions (A2.1)–(A2.4) for the mediator M were satisfied, and adherence were truly captured by the binary indicator M, then their estimate of the direct effect would be equal to both the controlled and natural direct effect and their estimate of the indirect effect would be equal to the natural indirect effect. However, dichotomization of the adherence is likely a coarsening of the actual underlying adherence and will likely not entirely capture the effects of adherence on the depression score. Conceived of another way, there are multiple versions of the adherence indicator M and multiple values of adherence that correspond to $M = 1$, for example.

Using the ideas above, we can consider the interpretation of the effect estimates of Emsley et al. (2010) when true adherence has been dichotomized for the purposes of the analysis. Suppose the no-unmeasured-confounding assumptions (A7.1)–(A7.4) for the effects of treatment and adherence on depression scores hold; note that assumptions (A7.1) and (A7.3) are guaranteed to hold by randomization of treatment. Under these no-confounding assumptions, the indirect effect could then be interpreted as the contrast of having the new treatment with the adherence indicator fixed to the value it would be with, versus without, the new treatment, in both cases then choosing randomly a particular value of actual adherence randomly from the distribution of those with that level of the adherence indicator, and who actually received the new treatment, conditional on covariates (i.e., $M = m, A = 1, C = c$). The direct effect estimate is equal to the sum of (i) the contrast of the new treatment versus treatment as usual, in both cases with the adherence indicator set to the level it would have been with treatment as usual, with the specific actual adherence value randomly drawn from the distribution of those who actually received the new treatment, conditional on covariates, and (ii) the effect of the new versus the standard treatment on the outcome mediated by adherence not captured by the adherence indicator itself.

We now consider a second example, one in which the underlying versions of the mediator is in fact unknown. Caffo et al. (2008) studied the extent to which the effect of cumulative lead dose, A, for organolead manufacturing workers on executive cognitive function test scores, Y, is mediated by white-matter brain volume, M, using data from 513 manufacturing workers. Brain volume was measured using magnetic resonance imaging which captures only brain volume differences and not more subtle neurobiologic changes to brain structure. Caffo et al. control for a number of covariates, C, including age, education, smoking, and alcohol consumption, and use a regression-based approach assuming no $A \times M$ product term (i.e., $\theta_3 = 0$) in regression models in Chapter 2. They obtain an estimate of the direct effect of a 3.79 point decline (95% confidence interval [CI] = -7.40 to -0.18) in executive functioning cognitive test scores per 1-μg/g increase in peak tibia lead

exposure, controlling for white matter in brain regions associated with lead; the indirect effect of lead exposure as mediated through white matter brain volume was a 1.21 ($P = 0.01$) point decline in executive functioning cognitive test scores per 1-μg/g increase in peak tibia lead exposure. If there was indeed no exposure–mediator interaction, and if assumptions (A2.1)–(A2.4) for the mediator M were satisfied, and if there were no multiple versions of the mediator, then their estimate of the direct effect would be equal to both the controlled and natural direct effect and their estimate of the indirect effect would be equal to the natural indirect effect.

In this setting, however, there is no clear unambiguous way to conceive of hypothetical interventions on the mediator, brain volume. It is not in fact clear at all what interventions would bring about changes to particular level. The versions of the mediator are effectively unknown. What then is the interpretation of the indirect effect estimate of Caffo et al. (2008)? Let us, for the time being, suppose the no-unmeasured-confounding assumptions (A7.1)–(A7.4) for the effects of the exposure and the versions of the mediator on the cognitive function outcome held. Consider a one-unit change in peak tibia lead exposure. By the discussion above, the indirect effect would then be interpreted as the contrast of cognitive functioning test scores between two scenarios. In both scenarios, lead exposure is set to the higher level, but then in one scenario brain volume is fixed to the value it would have been with higher lead exposure, and in the other scenario brain volume is fixed to its value with lower lead exposure, in both cases choosing a version of the mediator (to bring the specific brain volume about) randomly from the distribution of those who were actually at the higher level of lead exposure with that brain volume, conditional on covariates (i.e., $A = 1, M = m$, and $C = c$). The direct effect estimate is equal to the sum of (i) the contrast of the higher versus lower level of lead exposure, in both cases with brain volumes set to the level it would have been under the lower exposure using a version of the mediator (to bring the specific brain volume about) randomly from the distribution of those who were actually at the higher level of lead exposure with that brain volume conditional on covariates and (ii) the effect of the higher versus lower exposure on the outcome mediated by the various versions of the manner in which the actual brain volume levels came about that is not captured by the brain volume measurement itself. The effects estimated by Caffo et al. (2008) may indeed have an interpretation as causal effects, albeit somewhat subtle and complicated, even though there are multiple ways a particular level of white matter brain volume could come about and even though we are not able to comprehensively characterize what these various versions might be. These interpretations, however, do assume that our no-unmeasured-confounding assumptions hold with respect to the underlying unknown versions of brain volumes, and with the versions unknown, this assumption will be difficult to assess in practice. We come back to this point below.

7.2.3. Relevance and Limitations

The ability to appropriately interpret direct and indirect effect estimates in the presence of multiple versions of the mediator is arguably important for two reasons.

First, objections might be raised to the use of the counterfactual framework for mediation analysis in settings in which there are multiple versions of the mediator such that there is no single well-defined way to intervene on the mediator value of interest. Our discussion here shows that, although this context complicates analysis and interpretation, the use of the counterfactual approach to mediation analysis is still relevant in this setting. The results here help address violations of the no-multiple-versions-of-treatment assumption. Second, however, it is important to remember that what is estimated in this context as a direct effect in fact picks up not only a direct effect but also a mediated effect through the versions of the mediator that is not captured by the mediator measurement. In many contexts, this may lead to an overestimate of the direct effect and an underestimate of the indirect effect. Although this rule of thumb may be helpful in practice, it need not always hold. The effect mediated through the version of the mediator not captured by the mediator measurement will not necessarily be in the same direction as the natural indirect effect estimate through the mediator measurement. Indeed, when a mediator variable is dichotomized, without further assumptions, the natural indirect effect can be biased either upward or downward; and likewise with the natural direct effect (Ogburn and VanderWeele, 2012).

Arguably the most substantial limitation of the approach described here is the set of assumptions of no unmeasured confounding. This is an important limitation of any analysis that addresses questions of direct and indirect effects. However, in the context of multiple versions of the mediator, additional complications arise insofar as the assumptions of no unmeasured confounding are made with respect to the various versions of the mediator and not the mediator measurement itself. In many settings, an investigator may not be able to fully characterize all the possible versions of the mediator that may lead to a particular mediator measurement. In such cases, it will of course be very difficult to make assumptions about no unmeasured confounding with any certainty when the variable with respect to which these assumptions are being made (the versions of the mediator) is not entirely known. Interpreting direct and indirect effects in the presence of multiple versions of the mediator needs to be done cautiously. Importantly, however, abandoning the counterfactual framework is not really a "solution" to these problems of interpretation. Completely ignoring these issues of counterfactuals and the definition of interventions and proceeding simply with statistical techniques merely sweeps the issues under the carpet; it does not resolve them. The counterfactual framework provides an approach to help interpret direct and indirect effect estimates but also, perhaps just as importantly, it helps make clear when issues of interpretation are difficult and when we must proceed cautiously with interpretation.

7.3. CONTROVERSIES OVER ASSUMPTIONS AND ALTERNATIVE INTERPRETATIONS OF EFFECTS

Throughout our discussion of mediation we have employed our four assumptions about confounding, namely, that the measured covariates C suffice to control for [assumption (A2.1)] exposure–outcome confounding, [assumption (A2.2)]

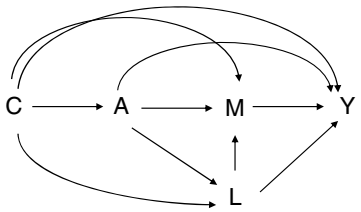

Figure 7.2 An example of a mediator-outcome confounder L that is affected by the exposure A.

mediator–outcome confounding, and [assumption (A2.3)] exposure–mediator confounding and finally that [assumption (A2.4)] there is also no mediator–outcome confounder that is itself affected by the exposure. This fourth assumption would thus be violated in Figure 7.2.

The technical assumptions are in fact somewhat more subtle than this. The statements we have made so far about the assumptions sufficing to identify the natural direct and indirect effects are correct, provided that the system underlying the variables is a causal diagram interpreted as a nonparametric structural equation model as in Pearl (2009; cf. Shpitser and VanderWeele, 2011). A causal diagram is a nonparametric structural equation model as in Pearl (2009) if (i) each variable is some arbitrary general function of the other variables with arrows into that variable and a random error term and (ii) if the random error terms are independent of one another. In general, for a causal diagram, if two random error terms were correlated, we would generally include an unmeasured variable U on the diagram with arrows into the two variables with correlated errors in order to make the errors, once the variable U has been added, independent. Essentially any time we simulate data, regardless of the distribution or functional form we specify, we are using an example of a nonparametric structural equation model. It is a very general class of models and under this class, if our four assumptions (A2.1)–(A2.4) hold on a causal diagram, then natural direct and indirect effects will be identified.

However, interpretations (Robins, 1986; Robins and Richardson, 2010; Richardson and Robins, 2013) can be given to causal diagrams other than that of Pearl's nonparametric structural equation models. These alternative interpretations make weaker assumptions about the causal diagram than Pearl's nonparametric structural equation models, and under these alternative interpretations, natural direct and indirect effects may not be identified even if assumptions (A2.1)–(A2.3) hold and there is no mediator–outcome confounder affected by exposure. In other words, under these weaker interpretations of causal diagrams, natural direct and indirect effects may not be identified even in a causal diagram like Figure 7.3. Robins and Richardson (2010) derive bounds for the natural direct and indirect when only assumptions (A2.1)–(A2.3) hold, without assuming Pearl's non-parametric structural equation model framework. The more general assumption that is needed [in addition to (A2.1)–(A2.3), stated directly as counterfactuals, rather than diagrams and arrows], is that the counterfactual Y_{am} is independent of M_{a*} conditional on the measured covariates C. This assumption is sometimes referred to as a "cross-world" independence assumption because it states that the counterfactual for Y,

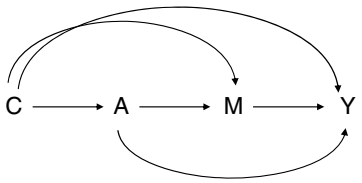

Figure 7.3 Mediation with exposure A, mediator M, and outcome Y with exposure-mediator-, mediator-outcome-, and exposure-mediator confounding controlled for by baseline covariates C.

when intervening to set A to a and M to m, is independent of the counterfactual for M when setting the exposure to a different level a^*; it is making an assumption about independence concerning what would happen in different "worlds"—that is, different settings of the exposure A. If we have a causal diagram like Figure 7.3 and we interpret it as a nonparametric structural equation model, then this assumption will hold. To see this, note that if ϵ_Y is the random error term for Y and ϵ_M is the random error term for M, then conditional on the covariates $C = c$, under Figure 7.3, the counterfactual Y_{am} will just be some function $f(a, m, c, \epsilon_Y)$ of the random error term ϵ_Y because a, m, c are all fixed and the counterfactual M_{a^*} will just be some function $h(a^*, c, \epsilon_M)$ of the random error term ϵ_M; however, under a nonparametric structural equation model, ϵ_Y and ϵ_M are assumed independent and any function of these variables will be independent and so the counterfactuals Y_{am} and M_{a^*} will be independent. It is this assumption, along with (A2.1)–(A2.3), that suffices to identify natural direct and indirect effects as shown in the Appendix. And again, this assumption will hold on a causal diagram interpreted as a nonparametric structural equation model if assumptions (A2.1)–(A2.3) hold and we have no mediator–outcome confounder affected by the exposure.

However, there is no way to check whether any observed data follows a nonparametric structural equation model. It is a very general class of models, but it makes untestable assumptions about independence of errors. We can include all measured and unmeasured common causes of any two variables on the diagram, but assumptions are still being made (Robins and Richardson, 2010). In some sense, this is not so very different from what is ordinarily done with causal inference from observational data, even when we are just interested in the total effect: We make assumptions about having controlled for confounding and we cannot empirically check these assumptions (Robins, 2003). However, in the case of total effects, we could, at least in theory, potentially design a randomized trial that would suffice to identify the effect of interest. In the case of natural direct and indirect effects, in general, no set of experimental interventions (even if we were able to intervene on both the exposure and the mediator) would suffice to identify these effects; we would always be making further assumptions. We can, as in Chapters 3 and 5, use sensitivity analysis to assess the sensitivity to violations of assumptions. Also certain clever experimental designs (Imai et al., 2013) can sometimes suffice to identify the sign of the natural direct and indirect effects, but the effects themselves are not identified from the data without making further assumptions.

However, even if we are not willing to make this cross-world independence assumption, or assume our causal diagrams corresponds to a set of nonparametric

structural equation models, the methods that we have been describing will still, under much weaker assumptions, have an interpretation as direct and mediated effects based on randomized interventions. We have considered such effects already in Chapters 5 and 6. In Chapters 5 and 6 we considered direct and mediated effect interpretations in trying to deal with issues of exposure-induced mediator–outcome confounders. However, even if there are no such variables, we could consider still interpretations of direct and mediated effects based on randomized interventions so that we do not need to make the assumption about cross-world independence or assume that a causal diagram is a nonparametric structural equation model.

Let a and a^* be two values of the exposure we wish to compare; for example, for binary exposure we may have $a = 1$ and $a^* = 0$. Let $G_{a^*|c}$ denote a random draw from the distribution of the mediator when setting the exposure to a^* amongst those with covariates $C = c$. Suppose then that assumptions (A2.1)–(A2.3) hold; that is, conditional on baseline covariates $C = c$ we have controlled for [assumption (A2.1)] exposure–outcome, [assumption (A2.2)] mediator–outcome, and [assumption (A2.3)] exposure–mediator confounding but we do not make the "cross-world" independence assumption. Then the methods described (e.g., in Chapter 2), for natural indirect effects, under these assumptions, will still estimate $NIE^R = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})$; that is, they will estimate the effect on the outcome of randomly assigning an individual who is given the exposure to a value of the mediator from the distribution of the mediator amongst those given exposure versus no exposure (conditional on covariates); this is an effect through the mediator; it is a randomized interventional analogue of the natural direct effect. Similarly, the methods for natural direct effects, will still estimate $NDE^R = \mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|c}})$; that is, they will estimate a direct effect comparing exposure versus no exposure with the mediator in both cases randomly drawn from the distribution of the population when given no exposure (conditional on covariates); this is a randomized interventional analogue of the natural direct effect. We might refer to these effects as interventional direct and indirect effects.

The effect $TE^R = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{a^*G_{a^*|c}})$ compares the expected outcome when having the exposure with the mediator randomly drawn from the distribution of the population when given the exposure (conditional on covariates) to the expected outcome when not having the exposure with the mediator randomly drawn from the distribution of the population when not exposed; it is a randomized interventional analogue of the total effect. With effects thus defined, we have the decomposition: $TE^R = NIE^R + NDE^R$ so that the overall effect decomposes into the sum of the effect through the mediator and the direct effect. These are not the natural direct and indirect effects considered earlier but are instead analogues arising from not fixing the mediator for each individual to the level it would have been under a particular exposure, but instead fixing it to a level that is randomly chosen from the distribution of the mediator amongst all of those with a particular exposure. These various effects are similar to those described by Didelez et al. (2006) and Geneletti (2007). These effects differ from the natural direct and indirect effects because instead of, for example, fixing the mediator for each individual to the level it would have been under exposure a^*, these randomized interventional

analogues fix the mediator to a random value of the mediator from the distribution of those with a^*. What the cross-world independence assumption gives us then is simply that these two effects, the natural direct and indirect effects, with their randomized interventional analogues, are equal. Importantly, however, even if we are not willing to make the cross-world independence assumption, our effect estimates can still be interpreted as particular types of direct and mediated effects. And this interpretation requires only assumptions (A2.1)–(A2.3), assumptions that we could guarantee to hold if we were able to randomize both the exposure and the mediator. One way to approach the interpretation of the estimators for direct and indirect effects then, if we can control for exposure–outcome, mediator–outcome, and exposure–mediator confounding, is that, without further assumptions, we can still interpret them as the interventional analogues of natural direct and indirect effects. If we are further willing to make the cross-world independence assumption, then we can also interpret these estimates as estimates of the natural direct and indirect effects themselves.

The question arises as to whether these interventional effects are really the effects of interest or whether natural direct and indirect effects are of interest. This will likely vary by context. In the following section, we describe an example from health disparities research where the interventional analogues, rather than the natural direct and indirect effects, are arguably the effects of interest. But in other settings, the natural direct and indirect effects (i.e., what would happen at the individual level if we fixed the mediator to a level it would have been under a counterfactual scenario) may be of interest. However, the best we can do, without making assumptions (that even in a doubly randomized trial cannot be guaranteed to hold), is to estimate the randomized interventional analogues, and again these will then equal the natural direct and indirect effects if we are willing to make the cross-world independence assumption or assume that the underlying causal diagram is a nonparametric structural equation model and assumptions (A2.1)–(A2.4) hold.

7.4. DIRECT AND INDIRECT EFFECTS IN HEALTH DISPARITIES RESEARCH

In health disparities research, when comparing outcomes across racial groups, often using a regression with race in the model, it is not infrequent to also control for socioeconomic status (SES) later in life. Doing so changes the interpretation of regression coefficients, because SES later in life is arguably on the pathway from race to some subsequent health outcome. Control for SES later in life is perhaps sometimes done so as to assess the extent to which health disparities across racial groups are in fact explained by differing SES levels later in life.

The application of methods for direct and indirect effects to the health disparities context is potentially problematic because the "effects of race" are not generally well-defined. In thinking about the "effects of race," we would generally not conceive of interventions to change race. Moreover, even if we did conceive of particular intervention, it is not always clear whether what we are after is the effects of skin color, parental skin color, genetic background, cultural background, and so on,

considered either singly or all together. A small literature has now begun to emerge discussing what might be meant by the "effects of race" from a counterfactual-based perspective (Kaufman, 2008; Greiner and Rubin, 2011; VanderWeele and Hernán, 2012; Sen and Wasow, 2013; VanderWeele and Robinson, 2014). Within the context of mediation, however, additional complications arise. When we think about, say, the "natural direct effect," we would be discussing what would happen under counterfactual scenarios in which we set the SES for a particular black individual, say, to what it would have been had they been white. Such counterfactuals statements generally do not strike us as particularly sensible. However, something along the lines of the randomized interventional analogues of the natural direct and indirect effects in the previous section are arguably somewhat more sensible. We might consider, for example, whether racial disparities in a particular health outcome would persist if the SES distribution for the black subpopulation were set to what it in fact was for the white population.

It can in fact be shown (VanderWeele and Robinson, 2014) that our methods for natural direct and indirect effects, when employed in the health disparities context, yield estimates with a coherent interpretation based on randomized interventions on the mediator. For simplicity, suppose we are comparing only two well-defined racial groups; for example, we let $A = 1$ denote black and $A = 0$ denote white. Suppose that we employ our methods for natural direct and indirect effects taking SES as the mediator and suppose that we have controlled for sufficient variables such that the associations between adult SES and the outcome actually reflect the effects of adult SES on the outcome. In this case the "direct effect" estimate that is obtained for race not through adult SES controlling for baseline covariates C measured at birth or conception (e.g., family and/or neighborhood SES at birth) could be interpreted as the health disparity that would remain for individuals, conditional on baseline covariates C, if within this population the adult SES distribution of the black population were set equal to that of the white population. We might refer to this as a "direct effect disparity measure" not through adult SES (i.e., how much of the disparity remains after accounting for adult SES). What is estimated as an indirect or mediated effect can be interpreted as how the health outcomes for the black population with baseline covariates C would change if the adult SES distribution of this black population were set equal to that of the black population versus that of the white population. We might refer to this as a "mediated disparity measure" through adult SES (i.e., how much of the disparity is due to difference in adult SES). The overall health disparity for those with baseline covariates C will be equal to the sum of these "direct" and "mediated" disparity measures. Importantly, we have this interpretation without having to define counterfactuals with respect to race; the only counterfactuals that are used concern potential interventions on SES. See the Appendix or VanderWeele and Robinson (2014) for further details and formalization.

As noted above, often SES later in life is added to in a regression model that has race as a covariate. This is effectively employing the difference method; thus when the difference method coincides with the natural direct and indirect effect methods of estimation, it will have the interpretation of the "direct" and "mediated" effects above as well. Thus, in particular, if the outcome is continuous and there is no statistical interaction between the exposure variable (race)

and the mediator variable (adult SES), then the coefficient for race in the model that includes adult SES (and other covariates) will correspond to a direct effect, and the difference in the coefficients for race in the models without versus with adult SES will correspond to the mediated effect. For a binary outcome with logistic regression, provided that the outcome is rare (or if a log-linear model is used with a common outcome), and if there is no statistical interaction between race and adult SES, then once again the coefficient for race in the model that includes adult SES (and other covariates) will correspond to the log of the direct effect, and the difference in the coefficients for race in the models without versus with adult SES will correspond to the log of the mediated effect (VanderWeele and Vansteelandt, 2010). On the odds ratios scale for logistic regression the overall disparity measure will decompose into a product (rather than the sum) of the "direct" and "mediated" disparity measures. However, the methods we have discussed in Chapter 2 can also be used to obtain direct and mediated effect estimates even when there is potential interaction between race and adult SES—for example, if the effects of adult SES differed by racial groups. And indeed there is some theory and empirical evidence for such interaction between race and SES for at least some health outcomes (cf. Sanchez-Vaznaugh et al., 2009; Chang and Lauderdale, 2005). The methods from Chapter 2 can accommodate such interaction and would still retain the interpretation of the direct and mediated disparity measures given above.

Of course, in practice, socioeconomic status (SES) itself is generally not precisely defined. Often a series of measures or indicators are combined into a single score. Sometimes a single indicator of SES, such as educational attainment, is used. In such cases, the direct and indirect effect disparity measures discussed above relate to how disparities would change if one particular dimension of SES, such as education, were equalized across racial groups. See also again Section 7.2 for discussion of the interpretation of direct and indirect effects when multiple indicators or measures are combined into a single score.

7.5. RUBIN'S SEEMINGLY PROBLEMATIC EXAMPLES

In a series of papers and seminars, Rubin (2004, 2005, 2010) has presented a number of examples that he suggests are problematic if one attempts to use concepts like "direct and indirect effects," and from this he argues that such concepts are unhelpful or meaningless. We will consider these examples here, one at a time.

Within the framework we have been considering, if we have controlled for exposure–outcome confounding and mediator–outcome confounding, then if, conditional on the covariates, a binary exposure is correlated with the outcome conditional on a binary mediator, then a controlled direct effect must be present (Pearl, 1995, 2009); in counterfactual notation at least one of $\mathbb{E}[Y_{1,m=1} - Y_{0,m=1}|c]$ or $\mathbb{E}[Y_{1,m=0} - Y_{0,m=0}|c]$ must be nonzero. Such correlation between the exposure and outcome conditional on the mediator is a sufficient condition for the presence of controlled direct effects. Rubin (2004) considers different direct effects of the form $\mathbb{E}[Y_1 - Y_0|M_0 = M_1 = m]$. An effect of the form $\mathbb{E}[Y_1 - Y_0|M_0 = M_1 = m]$ is the effect of the exposure on the outcome for the subpopulation for whom the mediator

Table 7-1. EXAMPLE FROM RUBIN (2004) IN WHICH THERE ARE PRINCIPAL STRATA DIRECT EFFECT BUT THE EXPOSURE AND OUTCOME ARE INDEPENDENT CONDITIONAL ON THE MEDIATOR

Principal Stratum	M_0	M_1	Y_0	Y_1
1	0	0	0	20
2	0	1	40	60
3	1	1	80	100

would take the same value m irrespective of whether the exposure were present or absent. For this subpopulation for whom the mediator would take the same value m irrespective of the exposure, we have that the exposure does not change the mediator and so any effect of the exposure on the outcome must be "direct." Effects of this form are sometimes referred to as "principal stratum direct effects," and we will consider such effects in more detail in the next chapter.

Rubin (2004) presents an example [see top half of Display 3 in Rubin (2004)] in which the population falls into three subsets, or "principal strata", based on the effect of the exposure on the mediator: In one stratum $M_0 = M_1 = 0$, in another stratum $M_0 = 0, M_1 = 1$, and in a third stratum $M_0 = 1, M_1 = 1$; in Rubin's example there is no one with $M_0 = 1, M_1 = 0$. In this example, Rubin (2004) considers the following potential outcomes presented in Table 7.1. In this example and all that follow Rubin assumes that the exposure A has been randomized.

Rubin assumes each of the three strata are equal size and considers what the observed average outcome would look like in this table conditional on exposure and mediator. He notes that with exposure randomized and each principal stratum of the same size we have that $\mathbb{E}[Y|A = 0, M = 0] = (0 + 40)/2 = 20$, $\mathbb{E}[Y|A = 0, M = 1] = 80$, $\mathbb{E}[Y|A = 1, M = 0] = 20$, and $\mathbb{E}[Y|A = 1, M = 1] = (60 + 100)/2 = 80$. He further notes that the principal stratum direct effect, $\mathbb{E}[Y_1 - Y_0|M_0 = M_1 = 0] = 20 - 0 = 20$, is nonzero, and the other principal stratum direct effect, $\mathbb{E}[Y_1 - Y_0|M_0 = M_1 = 1] = 100 - 80 = 20$, is also nonzero. We have nonzero principal strata direct effects. However, if we look at the correlation between the exposure and outcome condition on the mediator, we find that since $\mathbb{E}[Y|A = 0, M = 0] = 20 = \mathbb{E}[Y|A = 1, M = 0]$ and $\mathbb{E}[Y|A = 0, M = 1] = 80 = \mathbb{E}[Y|A = 1, M = 1]$, we in fact have that the exposure and outcome are independent conditional on the mediator. Rubin suggests that since the exposure and outcome are uncorrelated conditional on the mediator, it would seem that all of the effect is mediated and that there is no direct effect. However, when we look at the potential outcomes in Table 7.1 we see that there are principal strata direct effects. The suggestion then is that approaches employing direct and indirect effects are unhelpful or problematic. However, this is to confuse a *sufficient* condition with a *necessary* condition. If, conditional on covariates, control has been made for exposure–outcome confounding and mediator–outcome confounding, then correlation between exposure and outcome conditional on the mediator implies a controlled direct effect. But this is a *sufficient* condition, not a *necessary* one. There may be controlled direct effects even when the exposure and outcome

Table 7-2. Example from Rubin (2004) in Which There Are No Principal Strata Direct Effects but the Exposure and Outcome Are Correlated Conditional on the Mediator

Principal Stratum	M_0	M_1	Y_0	Y_1
1	0	0	0	0
2	0	1	40	60
3	1	1	80	80

are uncorrelated conditional on the mediator. In fact it can be shown that whenever a principal stratum direct effect is present, there must be individuals for whom the controlled direct effect is nonzero (VanderWeele, 2008; cf. Appendix). Using this result, we could in fact immediately conclude from the potential outcome in Table 7.1 that there must be a controlled direct effect for some individuals in the population. There is no contradiction in Rubin's example.

The second seemingly problematic example in Rubin (2004) is based on the potential outcomes given in Table 7.2.

In this example the principal stratum direct effects, $\mathbb{E}[Y_1 - Y_0|M_0 = M_1 = 0] = 0 - 0 = 0$ and $\mathbb{E}[Y_1 - Y_0|M_0 = M_1 = 1] = 80 - 80 = 0$, are both zero. If each of the three principal strata are of the same size, then the observed average outcomes conditional on exposure and mediator would be as follows: $\mathbb{E}[Y|A = 0, M = 0] = (0+40)/2 = 20$, $\mathbb{E}[Y|A = 0, M = 1] = 80$, $\mathbb{E}[Y|A = 1, M = 0] = 0$, and $\mathbb{E}[Y|A = 1, M = 1] = (60 + 80)/2 = 70$. Here we have $\mathbb{E}[Y|A = 0, M = 0] = 20 \neq 0 = \mathbb{E}[Y|A = 1, M = 0]$ and $\mathbb{E}[Y|A = 0, M = 1] = 80 \neq 70 = \mathbb{E}[Y|A = 1, M = 1]$. Here the exposure and the outcome are correlated conditional on the mediator, but we have no principal strata direct effects! Rubin (2004) argues that from the correlation of the exposure and outcome conditional on the mediator, it would appear that we had evidence of a direct effect, but in fact the principal stratum direct effects in this example are both zero. The suggestion once again is that approaches employing direct and indirect effects are unhelpful or problematic. But let us again return to, what implications actually do and do not hold. As we have noted several times now, if, conditional on covariates, control has been made for exposure–outcome confounding and mediator–outcome confounding, then correlation between exposure and outcome conditional on the mediator implies a controlled direct effect. Here, in this example, we must be in one of two scenarios: there is no exposure–outcome confounding because exposure has been randomized, there (1) may or (2) may not be mediator–outcome confounding. If there is no mediator–outcome confounding, then correlation between the exposure and the outcome conditional on the mediator implies the presence of a controlled direct effect. This may indeed be the case in this example. There can be controlled direct effect without there being principal stratum direct effects (VanderWeele, 2008). This can arise if one of the controlled direct effects, $Y_{1m} - Y_{0m}$, in principal stratum 2 is nonzero, or if the controlled direct effect $Y_{1,m=1} - Y_{0,m=1}$ in principal stratum 1 is nonzero, or if the controlled direct effect $Y_{1,m=0} - Y_{0,m=0}$ in principal stratum 3 is nonzero. All that "no principal stratum direct effect" tells us is that the controlled direct effect

$Y_{1,m=0} - Y_{0,m=0}$ in principal stratum 1 is zero, and the controlled direct effect $Y_{1,m=1} - Y_{0,m=1}$ in principal stratum 3 is zero. However, other controlled direct effects may be nonzero. Again, although a principal stratum direct effect implies a controlled direct effect, we may have controlled direct effects without there being principal strata direct effect. Thus one possibility in Table 7.2 is that there is no mediator–outcome confounding and that the observed data would indeed imply the presence of a controlled direct effect, even though there is no principal strata direct effects. The other possibility is that there is mediator–outcome confounding. In that case, as we have noted throughout the book, direct and indirect effects are not identified from the data; we would then not be able draw conclusions about a controlled direct effect one way or another from the data. But this too would not be a contradiction. There is nothing problematic about the second example in Rubin (2004) if we remember the issues of confounding control and consider what implications hold in which circumstances.

Importantly, in both the first and second examples, Rubin (2004) does not discuss mediator–outcome confounding, nor does he consider counterfactuals of the form Y_{am}. Without these counterfactuals we cannot say from the potential outcomes in Tables 7.1 and 7.2 what the controlled direct effects (or the natural direct and indirect effects) are. We also cannot ascertain whether or not mediator–outcome confounding is present. These two facts (no counterfactuals of the form Y_{am} and in general no way to assess mediator-outcome confounding) essentially form the grounds of all of Rubin's other seemingly problematic examples. Let us consider two further examples presented in Rubin (2005). Consider Table 7.3, replicated from Figure 2 of Rubin (2005), which presents both potential outcomes (columns 2–5) and observed data (columns 6–8).

In Rubin's example, treatment A, a fertilizer, is assumed to be randomized to different plots and data are subsequently collected on the number of plants in each plot, M, and the crop yield of the plots, Y. Different portions of the population have potential outcomes and observed data given in the table. The first two rows are the same types of plots (i.e., they have the same potential outcomes), but the plots in the second row are given the fertilizer and the plots in the first row are not. Likewise, the third and fourth rows are considered to be the same types of plots (they have the same potential outcomes), but the plots in the fourth row are given the fertilizer and the plots in the third row are not. From the table we see that the treatment increases the number of plants by 1 for all plots; that is, $M_1 - M_0 = 1$ for all plots. We also see that treatment has no effect on the crop yield outcome Y for any of the plots; that is, $Y_1 - Y_0 = 0$ for all plots. Rubin (2005) notes that if we were to compare the outcome across treatment groups while conditioning on M (e.g., $\mathbb{E}[Y|A = 1, M = 3] - \mathbb{E}[Y|A = 0, M = 3]$), we would obtain an apparent negative direct effect of $\mathbb{E}[Y|A = 1, M = 3] - \mathbb{E}[Y|A = 0, M = 3] = 10 - 12 = -2$. What appears to be a negative "direct effect" estimate seems to be at odds with the fact that there is no effect of the treatment on the outcome whatsoever; that is, $Y_1 - Y_0 = 0$ for all plots.

If we employ our concepts of mediation from Chapter 2, there are two possibilities in this example. One possibility is that -2 corresponds to an actual negative direct effect; the other possibility is that the naive estimate of -2 is a biased estimate

Table 7-3. Example from Rubin (2005) with What Seems to Be No Overall Effect of A on Y, a Positive Effect of A on M but a Seemingly Negative "direct Effect" of A on Y

Fraction of Population	M_0	M_1	Y_0	Y_1	A	M	Y
1/4	2	3	10	10	0	2	10
1/4	2	3	10	10	1	3	10
1/4	3	4	12	12	0	3	12
1/4	3	4	12	12	1	4	12

Table 7-4. Expansion of Potential Outcomes to Include $Y_{a=0,m=3}$ and $Y_{a=1,m=3}$, Consistent with the Data in Table 7.3, and Indicating a True Negative Controlled Direct Effect of A on Y of $\mathbb{E}[Y_{a=1,m=3} - Y_{a=0,m=3}] = -2$

Fraction of Population	M_0	M_1	Y_0	Y_1	A	M	Y	$Y_{a=0,m=3}$	$Y_{a=1,m=3}$
1/4	2	3	10	10	0	2	10	12	10
1/4	2	3	10	10	1	3	10	12	10
1/4	3	4	12	12	0	3	12	12	10
1/4	3	4	12	12	1	4	12	12	10

due to mediator–outcome confounding. We do not have the potential outcomes of the form Y_{am} in this table and so we cannot determine, from Table 7.3 alone, which of these two possibilities is the case. We will illustrate both of them, however, by supplementing the table with some of the potential outcomes of the form Y_{am}. Consider the extended data in Table 7.4 that is consistent with that in Table 7.3 but also gives potential outcomes for $Y_{a=0,m=3}$ and $Y_{a=1,m=3}$.

The potential outcomes in Table 7.4 are consistent with what is presented in Table 7.3 but we now have information on the potential outcomes for $Y_{a=0,m=3}$ and $Y_{a=1,m=3}$ as well. We can see from Table 7.4 now that the true controlled direct effect is $Y_{a=1,m=3} - Y_{a=0,m=3} = -2$ for all plots. If Table 7.4 corresponded to the true potential outcomes then our naive estimate from the observed data, $\mathbb{E}[Y|A = 1, M = 3] - \mathbb{E}[Y|A = 0, M = 3] = 10 - 12 = -2$, would turn out to be correct. At first this may seem puzzling: How can we have no overall effect but a negative direct effect? This is of course possible if the effect mediated by the number of plants M is positive. For example, we can see that in Table 7.4, for the final two rows of the table, the natural direct effect is $Y_{1M_0} - Y_{0M_0} = Y_{a=1,m=3} - Y_{a=0,m=3} = 10 - 12 = -2$ and the natural indirect effect is $Y_{1M_1} - Y_{1M_0} = (Y_{1M_1} - Y_{0M_0}) - (Y_{1M_0} + Y_{0M_0}) = (Y_1 - Y_0) - (Y_{1M_0} + Y_{0M_0}) = (12 - 12) - (10 - 12) = 2$. The natural direct and indirect effects are in opposite directions and of equal magnitude, and we thus get a total effect, $Y_1 - Y_0$, of 0. There is nothing contradictory here. Importantly, the reason that the naive estimate from the observed data, $\mathbb{E}[Y|A = 1, M = 3] - \mathbb{E}[Y|A = 0, M = 3] = 10 - 12 = -2$, gives the correct direct effect in Table 7.4 is that there is no mediator–outcome confounding present in this table. Recall that the assumption that there was no mediator–outcome confounding, stated formally, was that

Table 7-5. EXPANSION OF POTENTIAL OUTCOMES TO INCLUDE $Y_{a=0,m=3}$ AND $Y_{a=1,m=3}$, CONSISTENT WITH THE DATA IN TABLE 7.3, AND INDICATING NO TRUE CONTROLLED DIRECT EFFECT OF A ON Y, WITH THE APPARENT DIRECT EFFECT DUE TO MEDIATOR–OUTCOME CONFOUNDING

Fraction of Population	M_0	M_1	Y_0	Y_1	A	M	Y	$Y_{a=0,m=3}$	$Y_{a=1,m=3}$
1/4	2	3	10	10	0	2	10	10	10
1/4	2	3	10	10	1	3	10	10	10
1/4	3	4	12	12	0	3	12	12	12
1/4	3	4	12	12	1	4	12	12	12

Y_{am} is independent of M conditional on A. [assumption (A2.2)]. From Table 7.4, within strata of observed treatment A, the distribution of the potential outcome $Y_{a=0,m=3}$ is independent of observed M, and the distribution of the potential outcome $Y_{a=1,m=3}$ is also independent of observed M. We have no mediator–outcome confounding, and thus we get the correct direct effect estimate. Thus one possibility with regard to Table 7.3 is that the naive direct effect estimate $\mathbb{E}[Y|A=1,M=3]-\mathbb{E}[Y|A=0,M=3]=10-12=-2$ is in fact correct, even though the total effect is 0.

The other possibility is that the naive direct effect estimate $\mathbb{E}[Y|A=1,M=3]-\mathbb{E}[Y|A=0,M=3]=10-12=-2$ is biased due to mediator–outcome confounding. This is illustrated in Table 7.5.

As before, the potential outcomes in Table 7.5 are consistent with what is presented in Table 7.3, but we now have information on the potential outcomes for $Y_{a=0,m=3}$ and $Y_{a=1,m=3}$ as well. In Table 7.4 the true controlled direct effect is $Y_{a=1,m=3}-Y_{a=0,m=3}=0$ for all plots. The apparent direct effect estimate $\mathbb{E}[Y|A=1,M=3]-\mathbb{E}[Y|A=0,M=3]=10-12=-2$ is thus incorrect. In Table 7.5, the bias of the naive estimate is due to uncontrolled mediator–outcome confounding. Once again, the assumption that there is no mediator–outcome confounding, stated formally, is that (A2.2) Y_{am} is independent of M conditional on A. If we look in Table 7.5, we see that this is violated because, for example, conditional $A=1$ we have that $M=4$ implies $Y_{a=1,m=3}=12$ but $M=3$ implies $Y_{a=1,m=3}=10$; the independence does not hold. Thus a second possibility consistent with Table 7.3 is that the naive direct effect estimate $\mathbb{E}[Y|A=1,M=3]-\mathbb{E}[Y|A=0,M=3]=10-12=-2$ is biased due to mediator–outcome confounding.

One of these two possibilities must hold: Either the naive direct effect estimate is correct or there is unmeasured confounding. Of course, we cannot tell from Table 7.3 alone which of these possibilities is correct, nor can we be sure with the actual observed trial data whether mediator–outcome confounding is present. In many cases, mediator–outcome confounding will be present and the naive analysis for the direct effect will be biased. As we had emphasized repeatedly in Chapter 2, the assumption that mediator–outcome confounding has been controlled for is a strong assumption. It is the assumption that allows us to obtain valid estimates for controlled direct effects, but it is an assumption than we can never be sure holds. It was for this reason that in Chapters 3–5 of this book, we devoted considerable space to sensitivity so that we have tools to assess the extent to which unmeasured

mediator–outcome confounding might or might not undermine our conclusions in specific analyses.

Rubin (2005) presents another similar example (see Figure 3 in that paper) in which the total effect is positive and the naive direct effect estimate is negative. In that example, from the potential outcomes alone that are provided, it is possible to verify that mediator–outcome confounding is in fact present in the example.[1] But the same principles apply as before: Either the naive estimator of the direct effect is valid, or there is mediator–outcome confounding; in the first example of Rubin (2005), either possibility could be the case; in the second example of Rubin (2005), the latter explanation, mediator–outcome confounding, is correct. As we have emphasized before, mediator–outcome confounding is clearly a threat to all analyses of direct and indirect effects, and sensitivity analysis is essential. In Rubin's

1. If the reader is viewing Figure 3 from Rubin (2005), it is possible to see that there is mediator–outcome confounding in this example as follows. In Rubin's notation, treatment is W and the intermediate is C. The assumption that there is no mediator–outcome confounding, stated formally, is that Y_{wc} is independent of C conditional on W. If, in Figure 3 of Rubin (2005), we consider possible values for $Y_{w=0,c=3}$, we can see from the potential outcomes and the observed data that in the third and fourth rows, $Y_{w=0,c=3}$ must be 12 (since in these rows when $W = 0$ we have $C = 3$ and $Y = 12$ and thus $Y_{w=0,c=3} = Y_{WC} = Y = 12$) and likewise in the fifth and sixth rows $Y_{w=0,c=3}$ must be 14 (since in these rows when $W = 0$ we have $C = 3$ and $Y = 14$ and thus $Y_{w=0,c=3} = Y_{WC} = Y = 14$). The potential outcome $Y_{w=0,c=3}$ for the first and second rows is not fixed by the information presented in Figure 3 of Rubin (2005). However, for $Y_{w=0,c=3}$ to be independent of C conditional on W, it would be required that when $W = 0$, the mean of $Y_{w=0,c=3}$ would have to be the same irrespective of whether $C = 2$ or $C = 3$. When $W = 0, C = 3$ we have from Figure 3 of Rubin (2005) that the mean of $Y_{w=0,c=3}$ is $(14 + 12)/2 = 13$. This would thus require that $Y_{w=0,c=3}$ is 13 in the first two rows. However, for $Y_{w=0,c=3}$ to be independent of C conditional on W, we would also need that when $W = 1$, the mean of $Y_{w=0,c=3}$ would have to be the same irrespective of whether $C = 3$ or $C = 4$. When $W = 1, C = 4$, we have from Figure 3 of Rubin (2005) that the mean of $Y_{w=0,c=3}$ is 14. When $W = 1, C = 3$, we have from Figure 3 of Rubin (2005) that the mean of $Y_{w=0,c=3}$ is the average of 13 and the value of $Y_{w=0,c=3}$ is the first two rows; thus for the mean of $Y_{w=0,c=3}$ when $W = 1, C = 3$ to be equal to that when $W = 1, C = 4$, we would need that the value of $Y_{w=0,c=3}$ in the first two rows was 15. But this would contradict the requirement that the value of $Y_{w=0,c=3}$ in the first two rows is 13, which was required for $Y_{w=0,c=3}$ to have the same mean when $W = 0, C = 2$ as when $W = 0, C = 3$. Thus there is no set of potential outcomes for $Y_{w=0,c=3}$ that is consistent with Figure 3 of Rubin (2005) and for which Y_{wc} is independent of C conditional on W. There is thus mediator–outcome confounding in this example.

As a somewhat more subtle technical point, although we cannot have $Y_{w=0,c=3}$ independent of C conditional on $W = w$ for all w in Figure 3 of Rubin (2005) and thus this assumption for the identification of natural direct and indirect effects is not satisfied, if all we want are correct average controlled direct effects, then all that is needed is that $E[Y_{w=0,c=3}|W = 0, C = 3] = E[Y_{w=0,c=3}|W = 0]$ and $E[Y_{w=1,c=3}|W = 1, C = 3] = E[Y_{w=1,c=3}|W = 1]$. These equalities could in fact be satisfied with potential outcomes for $Y_{w=0,c=3}$ and $Y_{w=1,c=3}$ in Figure 3 of Rubin (2005) by having the potential outcomes for $Y_{w=0,c=3}$ in the six rows being $13, 13, 12, 12, 14, 14$ and the potential outcomes for $Y_{w=1,c=3}$ in the six rows being $11, 11, 13, 13, 12, 12$. Such potential outcomes would still be consistent with Figure 3 of Rubin (2005) and would also imply that the naive direct effect estimate of -1 in Rubin's example was in fact correct. However, it would still be the case that the distributions of the $P(Y_{w=1,c=3})$ and $P(Y_{w=0,c=3})$ would not be identified from the data because even with these potential outcomes for $Y_{w=0,c=3}$ and $Y_{w=1,c=3}$ given as above, we would only have absence of mediator–outcome confounding in expectation, not in distribution.

examples, there is nothing problematic about the concepts of direct and indirect effects that we have been employing throughout the book. The problem lies in trying to control for mediator–outcome confounding, which is, as we have discussed, no easy task.

Interestingly in these examples, the exposure that Rubin (2005) considers is fertilizer, the "mediator" is number of plants, and the outcome is crop yield. Here we may have difficulty imagining interventions on the mediator; we may be less willing to entertain counterfactual of the form Y_{am}. In this case, it may not seem unreasonable that Rubin does not have counterfactuals of the form Y_{am} in the potential outcome tables. Indeed another of Rubin's objections to the notions of direct and indirect effects (Rubin, 2013) is precisely that counterfactuals of the form Y_{am} are often not well-defined; there may be multiple ways to set the mediator variable to some level M, and these may have very different implications for the outcome, leaving a counterfactual of the form Y_{am} not well-defined. This is arguably a more relevant objection to the notions of direct and indirect effects that we have been considering in this book. Indeed, it is an objection that we have discussed at some length in Section 7.2 of this chapter. We saw that that even with multiple versions of the mediator, we need not necessarily abandon notions of controlled direct effects and natural direct and indirect effects, but we saw also that the interpretation of these effects and interpretation of the estimators generally used to identify them becomes considerably more subtle. We noted there that under assumptions about confounding control, the natural indirect effect does capture a particular type of mediated effect, but that the natural direct effect estimator, even under confounding control assumptions, captures both a direct effect and an effect mediated through the versions of the mediator not captured by the mediator measurement. Similar considerations were relevant to the controlled direct effect estimators. The interpretation of these effects were not entirely straightforward when multiple versions of the mediator were present, and we also noted that the confounding control assumptions were considerably more subtle and difficult to establish in the presence of multiple versions of the mediator. We need not then entirely give up methods that estimate direct and indirect effects when there are multiple versions of the mediator and the counterfactuals of the form Y_{am} are not well-defined but we do need to proceed more cautiously, and the objection that counterfactuals of the form Y_{am} are ill-defined (Rubin, 2013) is not an unreasonable one.

Indeed this issue and objection not only arises in the context of mediation, but also applies to assessing the overall effects of a single exposure for which multiple interventions on that exposure are conceivable. Here too we need not abandon discussion of causation or the estimation of causal effects, but we do need to take the presence of multiple versions of the exposure or treatment seriously and consider approaches and interpretations that take multiple versions of the exposure or treatment into account (Hernán and VanderWeele, 2011; VanderWeele and Hernán, 2013). Within the context of an intermediate variable M, an advantage of the principal stratum direct effect, $\mathbb{E}[Y_1 - Y_0 | M_0 = M_1 = m]$, is that it can be defined without reference to interventions on M or counterfactuals of the form Y_{am}. If interventions on the exposure are well-defined and unambiguous (or if the exposure is randomized) but interventions on the mediator are not well-defined, then the

principal stratum direct effect, $\mathbb{E}[Y_1 - Y_0 | M_0 = M_1 = m]$, poses fewer conceptual challenges than controlled direct effects or natural direct and indirect effects. But it is also important to understand that the principal stratum direct effects also address different questions of interest. We will consider principal stratum direct effects and their uses and limitations within the context of mediation in the following chapter.

7.6. A THREE-WAY DECOMPOSITION INTO DIRECT, INDIRECT, AND INTERACTIVE EFFECTS

As we have discussed on numerous occasions, the counterfactual-based approach to mediation analysis allows for effect decomposition even in the presence of exposure–mediator interaction. However, as was noted in Chapter 2, there are essentially two ways of going about this decomposition, essentially amounting to different ways of accounting for the interaction. In fact, arising from this, it turns out that it is possible to decompose a total effect into three parts: what might be called a pure indirect effect, a pure direct effect, and a mediated interaction. In this section we will consider this decomposition approach.

7.6.1. Different Decompositions

In Chapter 2, in discussing counterfactuals, we defined the natural direct effect as $Y_{1M_0} - Y_{0M_0}$ and defined the natural indirect effect as $Y_{1M_1} - Y_{1M_0}$, and we then had the decomposition $Y_1 - Y_0 = Y_{1M_1} - Y_{0M_0} = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$ where the decomposition is obtained by adding and subtracting the counterfactual Y_{1M_0}. The natural direct and indirect effects defined above are referred to by Robins and Greenland (1992) as "pure direct effects" and "total indirect effects," respectively. Robins and Greenland use the terminology "pure" and "total" because there are different ways of decomposing an overall effect into direct and indirect effect components. Above, we decomposed the overall or total effect as follows: $Y_1 - Y_0 = Y_{1M_1} - Y_{0M_0} = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$. For the natural direct effect, $Y_{1M_0} - Y_{0M_0}$, we compared outcomes under exposure versus no exposure, in both cases setting the mediator to what it would have been in the absence of exposure. We might instead compare exposure to no exposure, now in both cases setting the mediator to what it would have been in the presence of exposure. This would be the counterfactual contrast $Y_{1M_1} - Y_{0M_1}$. Likewise in the decomposition above, for the natural indirect effect, $Y_{1M_1} - Y_{1M_0}$, we compared outcomes when exposure is set to present and the mediator is set to the level it would have been with versus without exposure. We might instead compare outcomes when exposure is set to absent and the mediator is set to the level it would have been with versus without exposure. This would be the counterfactual contrast $Y_{0M_1} - Y_{0M_0}$. Robins and Greenland refer to $Y_{1M_1} - Y_{0M_1}$ as the "total direct effect" and $Y_{0M_1} - Y_{0M_0}$ as the "pure indirect effect," in contrast to the "pure direct effect" and "total indirect effect" considered above. We also then have an alternative effect decomposition of an overall effect: $Y_1 - Y_0 = (Y_{1M_1} - Y_{0M_1}) + (Y_{0M_1} - Y_{0M_0})$. We can thus decompose an overall effect, $Y_1 - Y_0$, either into a total indirect effect and a pure direct effect,

$(Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$, or into a total direct effect and a pure indirect effect, $(Y_{1M_1} - Y_{0M_1}) + (Y_{0M_1} - Y_{0M_0})$.

The "pure" and "total" terminology used by Robins and Greenland essentially arises from different ways of accounting for an interaction that is also mediated. As will be seen below, this mediated interaction involves the exposure changing the mediator and is thus a type of mediated effect; but it also involves an interaction, and thus an effect of the exposure not through the mediator is therefore in some sense direct also. It could thus be attributed either to the direct effect or to the indirect effect. When we decompose an overall or total effect into a pure direct effect and a total indirect effect, the indirect effect "picks up" the mediated interaction; the "pure" in "pure direct effect" effectively indicates that the direct effect does not pick up the interaction. When we decompose an overall effect into a total direct effect and a pure indirect effect, the direct effect picks up the interaction; the "pure" in "pure indirect effect" effectively indicates that the indirect effect does not pick up the interaction [cf. Hafeman (2008) for similar points using the sufficient cause framework]. We thus have two different decompositions depending on how we account for the mediated interaction. Traditionally the decomposition has been into the pure direct effect and the total indirect effect. This was arguably in part because of historical reasons because this decomposition was the one initially suggested by Pearl (2001); however, under certain "monotonicity" assumptions, the total indirect effect, in contrast to the pure indirect effect, would also give more evidence for the actual operation, rather than just the presence, of mediating mechanisms (Suzuki et al., 2011; VanderWeele, 2011c). In some cases, one decomposition may be preferred to another on substantive grounds because they constitute different effects, and different decompositions are answering different questions (Hafeman and Schwartz, 2009; Pearl, 2012). In other cases, deciding between the two may be less clear. The two decompositions remain and there is some level of arbitrariness or ambiguity in choosing between them. Again, this ambiguity of the choice between the two essentially arises from different ways of accounting for the mediated interaction, and in fact this ambiguity can be eliminated by a three-way decomposition of a total effect into three components: (i) a pure direct effect, (ii) pure indirect effect, and (iii) a mediated interaction.

7.6.2. A Three-Way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects

For simplicity we will consider the setting of a binary exposure and binary mediator. A more general decomposition for categorical or continuous exposure and mediator is given in the Appendix. For binary exposure A, binary mediator M, and outcome Y, we show in the Appendix that we have the following decomposition:

$$\begin{aligned} Y_1 - Y_0 = {} & (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) \\ & + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) \end{aligned} \tag{7.1}$$

The first term in this decomposition is the pure direct effect considered in the previous subsection. The second term in this decomposition is the pure indirect

effect considered in the previous subsection. The third term in this decomposition, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, is the product of an additive interaction between the exposure and the mediator on the outcome, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$, and the effect of the exposure on the mediator, $(M_1 - M_0)$. This interactive effect will be nonzero if and only if the exposure has some effect on the mediator, $(M_1 - M_0) \neq 0$, and if the additive interaction contrast, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$, is nonzero. We might thus refer to this interactive effect as a "mediated interaction." The contrast $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$ is a counterfactual measure of additive interaction. It is considered in more detail in Part II of the book when we turn to interaction analyses. It can be rewritten as $(Y_{11} - Y_{00}) - \{(Y_{10} - Y_{00}) + (Y_{01} - Y_{00})\}$. It will be nonzero for an individual if the effect on the outcome of setting both the exposure and the mediator to present differs from the sum of the effects of setting just the exposure to present and of setting just the mediator to present. In the Appendix, it is shown that this mediated interactive effect is equal to the difference between the total indirect effect and the pure indirect effect, $(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})$; the mediated interactive effect is also equal to the difference between the total direct effect and the pure direct effect, $(Y_{1M_1} - Y_{0M_1}) - (Y_{1M_0} - Y_{0M_0})$. The three-way decomposition above, along with the mediated interactive effect, essentially resolves the ambiguity above concerning the choice between decomposition into a pure direct and total indirect effect, or a total direct and pure indirect effect. The ambiguity was created by different ways of accounting for interaction. Instead of specifically assigning such mediated interaction to either the direct effect or the indirect effect, we can simply account for it separately.

As with natural direct and indirect effects more generally, we cannot identify the components of the decomposition individually, but we can identify them on average for a population under certain assumptions. In fact, these assumptions turn out to be no stronger than what was required for natural direct and indirect effects. We can identify the averages $\mathbb{E}[Y_{1M_0} - Y_{0M_0}|c]$, $\mathbb{E}[Y_{0M_1} - Y_{0M_0}|c]$ and $\mathbb{E}[(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)|c]$ simply under assumptions (A2.1)–(A2.4)—that is, that the covariates C suffice to control for [assumption (A2.1)] exposure–outcome, [assumption (A2.2)] mediator–outcome, and [assumption (A2.3)] exposure–mediator confounding and that [assumption (A2.4)] none of the mediator–outcome confounders are affected by the exposure. We describe a regression-based approach for the estimation of these effects below.

It was noted above that when using a two-way decomposition of a total effect into a direct and an indirect effect, there was ambiguity in how this was done and in the manner in which interaction was accounted for. The total effect could be decomposed into the sum of a total indirect effect and a pure direct effect or into a pure indirect effect and a total direct effect. The three-way decomposition arguably lends support to the approach of using the total indirect effect and the pure direct effect. This is because the total indirect effect is itself composed of the pure indirect effect and a mediated interaction. If the indirect effect that we use in a two-way decomposition of a total effect into direct and indirect effects is to capture the entirety of the effect that is in some sense mediated, then it arguably ought to include the mediated interaction as well. Fortunately, it is the decomposition of a total effect into

a total indirect effect and a pure direct effect that has most often been employed in practice and in software and there are moreover other theoretical arguments for sometimes preferring this particular decomposition (Suzuki et al., 2011; VanderWeele, 2011c). However, once again, with the three-way decomposition, one need not decide between alternative two-way decompositions and alternative approaches to accounting for interaction. The mediated interactive effect can be left as its own component in the decomposition.

7.6.3. A Three-Way Decomposition on the Ratio Scale

Thus far we have been considering the definition of these direct, indirect, and mediated interactive effects on a difference scale. Often, however, risk ratios or odds ratios are used for convenience, or ease of interpretation, or to account for study design and we considered the use of ratio scales in Chapter 2. On the risk ratio scale we can define the conditional total effect risk ratio by $RR_c^{TE} = \mathbb{E}[Y_1|c]/\mathbb{E}[Y_0|c]$, the pure direct effect risk ratio by $RR_c^{DE} = \mathbb{E}[Y_{1M_0}|c]/\mathbb{E}[Y_{0M_0}|c]$, and the pure indirect effect risk ratio by $RR_c^{IE} = \mathbb{E}[Y_{0M_1}|c]/\mathbb{E}[Y_{0M_0}|c]$. On this ratio scale we have the following decomposition for the excess relative risks:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + \left(\frac{\mathbb{E}[Y_{1M_1}|c]}{\mathbb{E}[Y_{0M_0}|c]} - \frac{\mathbb{E}[Y_{1M_0}|c]}{\mathbb{E}[Y_{0M_0}|c]} - \frac{\mathbb{E}[Y_{0M_1}|c]}{\mathbb{E}[Y_{0M_0}|c]} + 1\right) \quad (7.2)$$

On the left-hand side of this equation, the term $(RR_c^{TE} - 1)$ is the excess relative risk for the total effect. On the right-hand side of the equation, we have a three-way decomposition. The first term in this decomposition is the excess relative risk for the pure direct effect, the second term is the excess relative risk for the pure indirect effect, and the final term could be interpreted as a measure of mediated excess relative risk due to interaction. We will refer to this quantity as $RERI_{mediated}$. When using a ratio scale, investigators will sometimes use a quantity called the "relative excess risk due to interaction" or the "interaction contrast ratio," which we will consider in more detail in Part II of this book on interaction analyses. The causal relative excess risk due to interaction if M were binary would be defined as

$$RERI_{causal} = \frac{\mathbb{E}[Y_{11}|c]}{\mathbb{E}[Y_{00}|c]} - \frac{\mathbb{E}[Y_{10}|c]}{\mathbb{E}[Y_{00}|c]} - \frac{\mathbb{E}[Y_{01}|c]}{\mathbb{E}[Y_{00}|c]} + 1 \quad (7.3)$$

It assesses whether there is additive interaction but does so using ratios. The mediated relative excess risk due to interaction in (7.2) is analogous to the regular causal relative excess risk due to interaction in (7.3) but with replacing $m = 1$ and $m = 0$ in (7.3) with M_1 and M_0, respectively, in (7.2). Analogous to the decomposition for the total effect defined on a difference scale, we can decompose the excess relative risk for a total effect into the sum of the excess relative risk for the pure direct effect, the excess relative risk for the pure indirect effect, and the mediated relative excess

risk due to interaction:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated} \quad (7.4)$$

These quantities are likewise all identified under our four no-unmeasured-confounding assumptions (A2.1)–(A2.4). Similar decompositions would hold also for an odds ratio scale and for a hazards ratio scale, as was considered in Chapter 4 and as described in the Appendix.

7.6.4. Three-Way Decomposition into Direct, Indirect, and Interactive Effects with Regression

Here we describe how the three-way decomposition can be implemented using regressions that accommodate exposure–mediator interaction. Suppose that assumptions (A2.1)–(A2.4) hold, that Y and M are continuous, and that the following regression models for Y and M are correctly specified:

$$\mathbb{E}[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

In Chapter 2 we gave expressions for natural direct and indirect effects from these two regressions. However, as discussed above, we can further decompose such effects into a pure direct effect, a pure indirect effect, and a mediated interactive effect. For exposure levels a and a^* the pure direct effect and pure indirect effect are given by

$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$$
$$\mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] = (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)$$

and the mediated interactive effect is given by

$$\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}} - Y_{a^*M_a} + Y_{a^*M_{a^*}}|c] = \theta_3\beta_1(a - a^*)(a - a^*)$$

The sum of the pure indirect effect and the mediated interactive effect is equal to $(\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$, which is the total indirect effect given in Chapter 2. If the exposure were binary the pure direct, pure indirect and mediated interactive effects would respectively simply be $\{\theta_1 + \theta_3(\beta_0 + \beta_2'\mathbb{E}[C])\}$, $\theta_2\beta_1$, and $\theta_3\beta_1$. Standard errors for estimators of these quantities could be obtained using the delta method or by using bootstrapping.

Suppose now instead that Y were binary and M continuous and normally distributed, that assumptions (A2.1)–(A2.4) held, that the outcome was rare, and that the following regressions were correctly specified:

$$\text{logit}(P(Y = 1|a,m,c)) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

Then the pure direct effect risk ratio and the pure indirect effect risk ratio would be given by

$$RR_c^{DE} = \exp\left[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right]$$
$$RR_c^{IE} = \exp\left[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)\right]$$

where σ^2 is the variance of the error term in the linear regression model for M and the mediated interaction $RERI_{mediated}$ is given by

$$\begin{aligned}
&\exp\left[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*)\right.\\
&\quad \left. + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right]\\
&- \exp\left[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right]\\
&- \exp\left[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)\right] + 1
\end{aligned}$$

Again standard errors for these expressions could be obtained using the delta method or by using bootstrapping.

7.6.5. Illustrations of the Three-Way Decomposition

In Chapter 2 we considered the extent to which the effect of chromosome 15q25.1 rs8034191 C alleles on lung cancer risk was mediated by cigarettes smoked per day. rs8034191 C alleles had been found to be associated with both smoking and lung cancer, but there had been debate as to whether the effects on lung cancer were direct or mediated by cigarettes per day. Because the outcome, lung cancer, is rare, odds ratios approximate risk ratios. Using lung case–control data and logistic regression models, controlling for sex, age, education, restricting to Caucasians, and allowing for gene-by-smoking interaction, it was found that comparing 2 to 0 C alleles gave a pure direct effect odds ratio of 1.72 (95% CI: 1.34, 2.21), a total indirect effect odds ratio of 1.028 (95% CI: 0.99, 1.07), and a total effect odds ratio of $1.72 \times 1.028 = 1.77$ (95% CI: 1.38, 2.26), with proportion mediated $RR_c^{DE}(RR_c^{TIE} - 1)/(RR_c^{TE} - 1) = 1.72(1.028 - 1)/(1.77 - 1) = 6.3\%$. Most of the effect was found to be not through increasing cigarettes per day i.e. direct. If we now use the three-way decomposition for risk ratios:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}$$

we find $RR_c^{DE} = 1.72$, $RR_c^{IE} = 1.014$, $RERI_{mediated} = 0.036$. Thus for the excess relative risk we have $(1.77 - 1) = 0.77$, for the total effect, $(1.72 - 1)/0.77 = 93.7\%$ is attributable to the pure direct effect, $(1.014 - 1)/0.77 = 1.7\%$ is attributable to the pure indirect effect, and $0.036/0.77 = 4.6\%$ is attributable to the mediated interaction; once again the overall proportion mediated is $1.7\% + 4.6\% = 6.3\%$. Of the mediated effect, which is itself small proportion, most of this mediated effect is due to the mediated interaction, rather than being attributable a pure indirect effect. In this case, a priori, there is little reason to prefer either the pure indirect effect or the total indirect effect to assess the substantive question of interest; in this case, reporting the three-way decomposition may make most sense.

In Chapter 3 we examined the extent to which the effect of placental abruption on perinatal mortality was mediated by preterm birth using NCHS birth certificate files from 1995–2002. Allowing for potential interaction between abruption and preterm birth, and controlling for various sociodemographic variables, it was reported that the pure direct effect risk ratio was 10.18 (95% CI: 9.80, 10.58), the total indirect effect risk ratio was 1.35 (95% CI: 1.33, 1.38), and the total effect risk ratio was $10.18 \times 1.35 = 13.76$ (95% CI: 13.45, 14.08), with proportion mediated: $RR_c^{DE}(RR_c^{TIE} - 1)/(RR_c^{TE} - 1) = 10.18(1.35 - 1)/(13.76 - 1) = 28.1\%$. If we now use the three-way decomposition for ratios:

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}$$

we find $RR_c^{DE} = 10.18$, $RR_c^{IE} = 2.47$, and $RERI_{mediated} = 2.11$. Thus for the excess relative risk we have $(13.76 - 1) = 12.76$ for the total effect, $(10.18 - 1)/12.76 = 71.9\%$ is attributable to the pure direct effect, $(2.47 - 1)/12.76 = 11.5\%$ is attributable to the pure indirect effect, and $2.11/12.76 = 16.6\%$ is attributable to the mediated interaction; once again the overall proportion mediated is $11.5\% + 16.6\% = 28.1\%$. From this analysis we see that although a substantial portion of the effect of abruption on infant mortality is mediated by increasing the likelihood of preterm birth, in the majority of these cases of mediation, it is the interaction between abruption and preterm birth that brings about infant mortality.

As with the analyses of these examples in Chapters 2 and 3, the three-way decomposition here requires that assumptions (A2.1)–(A2.4) hold. Sensitivity analysis techniques described in Chapter 3 could be applied to the pure direct and pure indirect effects; however, new sensitivity analysis techniques would need to be developed for the mediated interaction, and these are not yet available.

The examples do raise the question of which of these decompositions is to be preferred—the two-way or the three-way—and of what it is that is ultimately of interest when we carry out effect decomposition. The two-way decomposition is simpler, but the three-way decomposition has the potential to give additional insight. It allows us to assess how much of the total indirect effect is due to a mediated interaction versus a pure indirect effect. It makes clearer the role of interaction in mediation analysis. A researcher interested in mediation and effect decomposition should perhaps consider first whether the pure indirect effect or the total indirect effect more closely corresponds to what might be the mediated effect of interest. As has been pointed out previously in the literature, these two effects have different substantive interpretations and one or the other may be useful in different contexts (Hafeman and Schwartz, 2009). In other cases, however, such as perhaps in the genetics example above, there may be no clear reason to choose between the two; in such instances, no choice need be made; an investigator can decompose the total effect into three components. In other cases, a researcher may want to know what portion of a mediated effect requires also the joint operation of the exposure and the mediator. In these cases, the three-way decomposition can give further insight into this question, as was perhaps the case in the perinatal example above. In summary, then, if the researcher is clearly interested in either the pure indirect effect or total indirect effect on substantive grounds, then there is perhaps

no reason to pursue the three-way decomposition. If the substantive setting is such that the choice between the two seems arbitrary or if further insight is desired into what portion of a mediated effect requires also the joint operation of the exposure and the mediator, then the three-way decomposition might be pursued. In Chapter 14 we will return once again to the relationship between mediation and interaction and consider an even more refined decomposition.

7.7. ALTERNATIVE IDENTIFICATION STRATEGIES USING CONFOUNDING CONTROL

In Chapter 2 and in many of the subsequent chapters, we had articulated the identification assumptions for natural direct and indirect effects as [assumption (A2.1)] no unmeasured confounding of the exposure–outcome relationship conditional on measured covariates C, [assumption (A2.2)] no unmeasured confounding of the mediator–outcome relationship conditional on measured covariates C, [assumption (A2.3)] no unmeasured confounding of the exposure–mediator relationship conditional on measured covariates C, and [assumption (A2.4)] no mediator–outcome confounder that is itself affected by the exposure. We noted that assumptions (A2.1) and (A2.3) would hold if the exposure were randomized. Assumptions (A2.1) and (A2.3) are likewise the assumptions that would generally be made in an observational study of the overall effect of the exposure on an outcome. We have also considered in Chapter 3 sensitivity analysis techniques for violations of some of these assumption, focusing principally on assumption (A2.2) about mediator–outcome confounding, because this is an assumption that would not hold even if the exposure were randomized. In Chapter 5 we also discussed (a) approaches to effect decomposition that could be used under violations of assumption (A2.4)—that is, in settings in which there is a mediator–outcome confounder affected by the exposure—and (b) sensitivity analysis for this setting.

A number of other sets of assumptions for the identification of natural direct and indirect effects have been given (Pearl, 2001, 2014; Taylor et al., 2005; Petersen et al., 2006; Imai et al., 2010c; Hafeman and VanderWeele, 2011). These can roughly be divided into two categories. First there are sets of assumptions that suffice to identify natural direct and indirect effects using the regression-based or simulation-based approach that arise from Pearl's "mediation formula" (Pearl, 2012) that was described in Section 2.16. These assumptions can differ in technical details or precise articulation from those that we have described as assumptions (A2.1)–(A2.4), but these differences do not amount to much in practice. For example, Imai et al. (2012c) articulate the identification assumptions as essentially (i) there is no unmeasured confounding of the exposure–outcome and mediator–outcome relationships jointly, conditional on baseline measured covariates C, and (ii) there is no unmeasured confounding of the mediator–outcome relationship conditional on baseline measured covariates C. At first sight it may appear that Imai et al. (2010c) do not require our assumption (A2.4), namely, the absence of mediator–outcome confounders affected by exposure. But upon closer examination, this assumption is in fact contained in (i) and (ii) above. Imai et al. (2010c) require that (ii) there is no unmeasured confounding of the mediator–outcome relationship conditional

on baseline measured covariates C. The key here is that the measured covariates C that control for mediator–outcome confounding must be baseline covariates. Thus if there were a mediator-outcome confounder affected by the exposure, the baseline covariates themselves would not suffice to control for mediator–outcome confounding; if the mediator–outcome confounder affected by exposure were to be included in the set of covariates C, then C would no longer be a set of baseline covariates because there would be a covariate that occurred after (and was affected by) exposure. Thus assumption (i) of Imai et al. (2010c) in fact correspond to assumptions (A2.1) and (A2.3), and assumptions (ii) of Imai et al. (2010c) correspond to assumptions (A2.2) and (A2.4). The same assumptions must effectively be made.

It can in fact be shown that for natural direct and indirect effects to be identified using the mediation formula (e.g., the regression based approaches in Chapter 2) on an arbitrary causal diagram as interpreted by Pearl (2001, 2009), assumptions (A2.1)–(A2.4) are both necessary and sufficient for nonparametric identification (Shpitser and VanderWeele, 2011). Recall that Pearl's mediation formula, described in Section 2.16, said that under assumptions (A2.1)–(A2.4), natural indirect effects were identified by the empirical expression:

$$\sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}$$

This is the formula that was used in the Appendix to derive our analytic expression for natural indirect effects using the regression models; it is also the formula that underlies the simulation-based approach described in Chapter 2. What the result of Shpitser and VanderWeele (2011) says is that if we are to use this formula in general settings, then we essentially must make assumptions (A2.1)–(A2.4). With other sets of identification assumptions for Pearl's mediation formula to hold (Petersen et al., 2006; Imai et al., 2010c; Hafeman and VanderWeele, 2011), one can come up with mathematical examples in which one set of identification assumptions hold and another does not, or settings in which there is no interaction or where certain variables are binary and monotonicity relationships hold in which one set of identification assumptions hold and another does not. However, whenever one articulates these technical identification assumptions in terms of a causal diagram (e.g., in terms of the intuitive statements about confounding), all of these assumptions amount to the same thing. They only differ in technicalities and in certain very special cases. Essentially to use Pearl's mediation formula or the regression-based and simulation-based approaches in Chapter 2, assumptions (A2.1)–(A2.4) are needed.

However, this statement about identification applies only to Pearl's mediation formula, and this brings us to another class of identification assumptions: those that achieve identification without using the mediation formula (Pearl, 2001, 2014). There are certain causal diagrams, interpreted as nonparametric structural equation models as in Pearl (2001, 2009), in which natural direct and indirect are identified, but they are not identified using the mediation formula. This can arise if conditioning on different sets of covariates is necessary to make the exposure–outcome, mediator–outcome, and exposure–mediator relationships

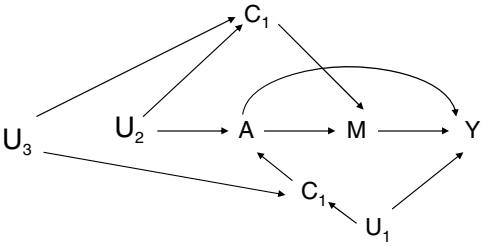

Figure 7.4 Diagram in which natural direct and indirect effects are identified from observed data (A, M, Y, C_1, C_2) with (U_1, U_2, U_3) missing.

unconfounded. For example, Pearl shows that in the rather complicated diagram in Figure 7.4, natural direct and indirect effects can be nonparametrically identified from the data irrespective of the distribution of the observed variables (A, M, Y, C_1, C_2) even with data missing on (U_1, U_2, U_3). However, the mediation formula cannot be used for identification; a different formula, given in Pearl (2014), is needed. In other instances it is also sometimes possible to use so-called "front-door" paths to identify certain direct and indirect effects (Kuroki and Miyakawa, 1999; Shpitser and Pearl, 2008; VanderWeele, 2011a). In fact, a complete theory for when natural direct and indirect effects are or are not identified from data is now available (Shpitser and Pearl, 2008). However, statistical methods have not yet been developed that provide estimates of natural direct and indirect effects in most of these settings, and it is not clear how often such settings arise in practice (Imai et al., 2014).

7.8. IDENTIFICATION USING BASELINE COVARIATES THAT INTERACT WITH EXPOSURE

One final identification strategy for direct and indirect effects that is sometimes applicable even when there is mediator–outcome confounding is worth discussing. Suppose we are interested in the controlled direct effect of the exposure on the outcome not through the mediator, $\mathbb{E}[Y_{am} - Y_{a^*m}|c]$, and we are willing to assume that this effect is constant across m; that is, there is no exposure–mediator interaction, so that $\mathbb{E}[Y_{am} - Y_{a^*m}|c] = \beta(a - a^*)$ for some fixed constant β. Suppose that the exposure–outcome and exposure–mediator relationships are unconfounded but that there may be mediator–outcome confounding as in Figure 7.5. Suppose further that we believe that there is one or more covariates X such that the effect of A on M on the additive scale varies across strata of X, but that the direct effect on the additive scale of A on Y not through M does not vary across strata of X. Suppose also that A and M do not interact on the additive scale in their effects on Y and that X and M do not interact on the additive scale in their effects on Y. In summary, we assume that there is an interaction between X and A in their effect on M (and this interaction must be present for this identification approach to work), but there is no interaction in their effects on Y between X and A, or between X and M, or between A and M.

Under these assumptions, Ten Have et al. (2007) show that the controlled direct effect, $\mathbb{E}[Y_{am} - Y_{a^*m}|c] = \beta(a - a^*)$, is identified and methods to estimate this

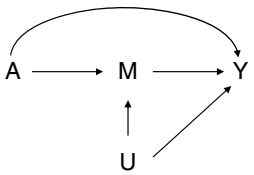

Figure 7.5 Randomized exposure A but with an unmeasured common cause U of mediator M and outcome Y.

effect under these assumptions are described by Ten Have et al. (2007), Emsley et al. (2010), Ten Have and Joffe (2012), and Emsley and Dunn (2012). The principle behind the identification is that the $X \times A$ interaction effectively serves as an "instrument" for the effect of the mediator on the outcome (i.e., the $X \times A$ interaction affects the mediator but only affects the outcome through the mediator) and allows for the identification of the effect of the mediator on the outcome and thus also (under the no interaction assumptions) for the controlled direct effect. We will discuss instrumental variable methods (outside the context of mediation) in somewhat more detail in the next chapter, but this alternative approach using baseline covariates that interact with the exposure in affecting the mediator is essentially just using such interactions as an instrument for the effect of the mediator on the outcome.

Emsley and Dunn (2012) describe a relatively straightforward approach to estimate this controlled direct effect under the assumptions above when this effect is identified. For example, when the outcome is continuous and the treatment is randomized so no further covariates C are needed to control for exposure–outcome or exposure–mediator confounding, then with outcome "y," exposure "a," mediator "m," and baseline covariate "x" that interacts with the exposure in its effects on the mediator in Stata, one can use ivregress 2sls y ax (m = a*x). See also Emsley and Dunn (2012) for implementation when the outcome is binary. If it is assumed not only that the average controlled direct effect, $\mathbb{E}[Y_{am} - Y_{a^*m}|c]$, does not vary with m but also that there is no interaction between A and M at the individual counterfactual level, then the controlled direct effect estimated using this approach will equal the natural direct effect (Robins, 2003) and one can then use the difference between the total effect and the controlled direct effect as an estimate of the natural indirect effect.

It should be noted that the assumption that the covariate X does not interact with the exposure A in its effects on Y (except through M) is crucial when using this approach. If it is violated, this approach can be severely biased. The assumption cannot be verified because although one could empirically test in Figure 7.5 whether the effect of A on Y varies with X, the assumption that is required here is in fact that the direct effects on the additive scale of A and X on Y, not through M, do not interact, and this is not empirically verifiable. Indeed if there is an interaction between X and A in their effects on M, then we might even expect there to be an interaction between X and A in their overall effects on Y since there is an interaction between X and A in their effects on M and M itself affects Y. It may thus be difficult to establish on substantive grounds that this assumption that the effect

on the additive scale of A on Y, not through M, does not vary across strata of X. The use of different sites in a trial is sometimes proposed as such a baseline covariate, but it has to be established that the effect of the treatment on the mediator varies by site but that the direct effect of treatment on the outcome does not vary by site.

Note also that the approach described above also assumes no interaction between A and M, or between X and M in their effects on Y. These assumptions likewise may not always be reasonable. Finally, we really do need to have an interaction between the effects of X and A on M for this method to work. If this interaction is weak, the approach can perform poorly, and biases that arise from violations of any of the other assumptions can also be amplified when this interaction is weak. If there are strong substantive reasons to believe there is interaction between a baseline covariate and the exposure in their effects on the mediator, but not in their effects on the outcome, then this alternative identification approach is certainly worth considering. Otherwise the approaches we have been considering in Chapters 2–6 in conjunction with sensitivity analysis is probably a more reliable approach to assessing direct and indirect effects.

7.9. POWER AND SAMPLE SIZE CALCULATIONS FOR MEDIATION ANALYSIS

One issue we have not discussed which is relevant to mediation, especially in designing a study, is that of power and sample size calculations for direct and indirect effects. Unfortunately, the current literature on this topic is somewhat limited and further development is still needed. Fritz and MacKinnon (2007) present some basic power and sample size requirements using simulations corresponding to small-, medium-, and large-sized effects for the exposure on the mediator and the mediator on the outcome. However, these do not allow an investigator to precisely calculate power or sample size when specifying exact effect sizes other than the scenarios they give. Kenny and Judd (2014) show that in many settings, power to detect indirect effects is greater than that for total effects, and power to detect direct effects is less than that for total effects. Vittinghoff et al. (2009) present a variety of power- and sample-size formulae when one is willing to assume that the exposure has an effect on the mediator (so that all that needs to be tested is whether the mediator has an effect on the outcome), but this approach presupposes part of the hypothesis that is to be tested in mediation (that the exposure affects the mediator). Freedman et al. (1992) and Freedman and Schatzkin (1992) present power and sample size formulae for the proportion mediated measure discussed in Section 2.13 when using the difference method. However, as discussed above, this can be a volatile measure, especially when the total effect is small, and if what is of interest is simply testing the mediated effect, then one would be better off testing whether the indirect effect itself is different than zero, rather than testing the proportion mediated. Moreover, none of the existing approaches accommodate exposure–mediator interaction. Additional research needs to be done on power- and sample-size calculations for natural direct and indirect effects. Until such formulae are available, the tables in Fritz and MacKinnon (2007) can provide some indication of what sort of sample sizes might be required.

7.10. DISCUSSION

In this chapter we have considered a number of additional topics in mediation analysis. We have considered alternative estimation approaches to assess direct and indirect effects, but have also noted that in many settings the regression-based approach presented in Chapters 2, 4, and 5 is to be preferred unless certain considerations make estimation difficult, or analytic calculations intractable, or models potentially incompatible. We have also considered different interpretations of our estimators for natural direct and indirect effects when these effects are potentially not identified or not well-defined. We have considered different interpretations of our direct and indirect effect estimators when there are multiple versions of the mediator (i.e., multiple ways in which the mediator might be changed) or when we are not willing to make certain technical assumptions which allow us to identify counterfactual quantities, or when we are applying methods for direct and indirect effects in the context of health disparities research in which it does not make sense, say, to talk about what would have happened had someone been of a different race. In each of these settings the direct and indirect effect estimates can still be interpreted, not as the counterfactual-based definitions of natural direct and indirect effects that we have been considering in previous sections, but rather as the effects of certain randomized interventions. In this chapter we have also discussed the ambiguity of accounting for interaction in mediation analysis and how multiple effect decompositions are possible. One way to resolve this ambiguity is to account for the mediated interaction separately so that one obtains a three-way decomposition of a total effect into a pure direct effect, a pure indirect effect, and a mediated interaction. We have also discussed alternative identification strategies to estimate direct and indirect effects. Many of the alternative strategies are not very different from, or even fall essentially within the same set of assumptions as, the no-unmeasured-confounding assumptions we considered in the previous chapter. However, one approach, which tries to leverage covariates that interact with the exposure in their effects on the mediator but do not interact with the exposure in their direct effects on the outcome, was quite distinct and, at least in some settings, could be used if unmeasured confounding is thought to be intractable and sensitivity analysis is inconclusive. We have discussed the limited work on power and sample size calculations for direct and indirect effects, but clearly this is an area in which further research and development is required.

Some of the topics in this chapter touch on the frontiers of current research on mediation. In the last several years there has been rapid expansion of the methodological literature on causal inference approaches to mediation, and this will likely continue in years to come. This chapter has provided a snapshot of some of these recent developments. In the next chapter we will consider other concepts and methods that relate to intermediate variables between an exposure and an outcome but do not fall within the context of mediation—that is, of trying to assess the extent to which the effect of an exposure on an outcome operates through an intermediate.

8

Other Topics Related to Intermediates

In this chapter we will consider a number of other topics related to intermediate variables that may occur sometime between a particular exposure and some outcome. These other topics and approaches do not constitute "mediation analysis" per se, conceived of as assessing the effects that operate through a particular intermediate variable versus through other pathways, but they are topics that are related, sometimes employ similar methods, and are occasionally conflated with questions of mediation. Each of these topics could merit a book in and of itself, so the focus of this chapter will simply be on pointing out the conceptual similarities and differences between these various other topics on the one hand and mediation analysis on the other. These topics include principal stratification, surrogate outcomes, instrumental variables, and Mendelian randomization.

8.1. PRINCIPAL STRATIFICATION

As discussed in Chapter 7, the concepts of direct and indirect effects that we have employed thus far have been defined in terms of hypothetical interventions on the exposure and the intermediate variable. We noted in Chapter 7 that in some cases, the ability to intervene on an intermediate may seem rather implausible. We discussed in that chapter the interpretation of direct and indirect effect estimates when there were potentially multiple ways (or even unknown ways) to bring about particular values of the intermediate.

An alternative approach to reason about intermediate variables is what is sometimes referred to as "principal stratification" (Frangakis and Rubin, 2002). The approach to the analysis of intermediates based on principal stratification only considers hypothetical interventions on the exposure variable and then assesses the effects of the exposure on outcome across strata defined by the effect of the exposure on the intermediate.

Let A denote some binary treatment or exposure variable and Y some outcome, and let S denote a variable that occurs between A and Y. We will use S rather than M as our intermediate variable in this section and the next section because,

as will be seen below, the approaches that are considered do not really correspond to mediation itself. We let S_a denote the potential outcome or counterfactual outcome for each individual that we would have observed had A, possibly contrary to fact, been a. A principal stratum is simply a subgroup of individuals homogenous in their joint potential outcomes (S_0, S_1). If S is also binary, then we have four principal strata: $(S_0 = 0, S_1 = 0)$ sometimes called "never-takers"; $(S_0 = 0, S_1 = 1)$ sometimes called "compliers"; $(S_0 = 1, S_1 = 0)$ sometimes called "defiers"; and $(S_0 = 1, S_1 = 1)$ sometimes called "always takers."

As in previous chapters, let Y_a denote the potential outcome for each individual that we would have observed had A, possibly contrary to fact, been a. The overall causal effect for the population is then given by $\mathbb{E}[Y_1 - Y_0]$. However, we could also consider the causal effect of A on Y conditional on the principal stratum, that is, $\mathbb{E}[Y_1 - Y_0 | S_0 = s_0, S_1 = s_1]$. This is what Frangakis and Rubin (2002) call a "principal causal effect." Conditioning on the principal stratum has certain advantages over conditioning on the observed posttreatment variable S. In general, if we condition on a post-treatment variable S, we will induce bias in the analysis, either (a) by potentially blocking some of the effect of the exposure on the outcome if we are interested in the total effect or (b) because of unmeasured common causes of the intermediate and the outcome if we are interested in the direct effect. However, the principal stratum, (S_0, S_1), that an individual belongs to is essentially viewed as a pretreatment characteristic of an individual and thus we can, in principle, stratify on it, as we could any other pretreatment variable. The difficulty is that we do not know who is in which principal stratum. As is discussed below, this creates problems for identifying quantities like $\mathbb{E}[Y_1 - Y_0 | S_0 = s_0, S_1 = s_1]$ from observed data and so, like our notions of direct and indirect effects, fairly strong assumptions are needed to identify these quantities of interest.

These notions of principal stratification have been employed in a number of different contexts, and we will briefly describe some of these; we will then turn to the question of how this approach answers different questions from the mediation analysis approaches we have been describing. The notion of principal stratification is most closely associated with a paper of Frangakis and Rubin (2002). Although the idea of principal stratification had clear antecedents (Robins, 1986; Angrist et al., 1996), Frangakis and Rubin (2002) proposed that this approach to thinking about causal effects be used to address a broad class of related problems concerning noncompliance, censoring-by-death, and surrogate outcomes, topics to which we now turn.

8.1.1. Principal Stratification and Censoring by Death (and the Analysis of Post-infection Outcomes)

One area of application of the idea of principal stratification that has received considerable attention in the literature is the analysis of outcomes that have been censored or "truncated" by death. Consider a randomized trial comparing two drugs (the variable A being an indicator of which of the two drugs was assigned)

and suppose we were interested in comparing quality of life outcomes (Y) at 6-months follow-up under these two drugs. If, however, some individuals die before the 6-month follow-up, their quality of life is not simply missing, it is undefined. We could attempt to compare outcomes amongst those who actually survived $(S = 1)$: We could examine the contrast $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$. The trouble with this is that survival is a post-treatment variable, and it may be affected by treatment; conditioning on it would essentially break randomization and could induce bias. Perhaps drug 1 was more likely to kill patients who are very sick at baseline than drug 2. A comparison of the quality of life outcomes between the two drugs amongst survivors would essentially be an unfair comparison because the sick patients are included in the average quality-of-life scores for drug 2 but they are not for drug 1 (because under drug 1, they die). An alternative comparison that would make sense in this setting is to compare the quality-of-life outcomes for the group that would have survived irrespective of which drug they were given. In the notation given above this is, $\mathbb{E}[Y_1 - Y_0|S_0 = 1, S_1 = 1]$. This is a principal stratum causal effect, sometimes referred to as the survivor average causal effect (SACE). In this context in which outcomes are effectively censored or truncated due to death, this is really the only comparison that is fair. The principal stratification approach is thus of considerable importance in addressing these questions as well, and a number of papers have provided methods to try to assess this survivor average causal effect when outcomes are truncated due to death (Robins, 1986; Zhang and Rubin, 2003; Rubin, 2006; Frangakis et al., 2007; Imai, 2008; Egleston et al., 2009; Chiba and VanderWeele, 2011; Tchetgen Tchetgen et al., 2012). A very closely related (essentially isomorphic) problem concerns the analysis of the effect of some treatment or vaccine on a post-infection outcome (e.g., HIV viral load) which is only defined for persons who are infected. In this context, one would want to know the effect of treatment on the post-infection outcome within the principal stratum who would develop the infection irrespective of whether they were given treatment. Many of the important methodological contributions to analyzing these principal strata effects have been developed within this infectious disease context (Gilbert et al., 2003; Hudgens et al., 2003; Hudgens and Halloran, 2006; Shepherd et al., 2006, 2007; Jemiai et al., 2007). With the analysis of problems concerning censoring by death or post-infection outcomes, the principal stratification framework has once again given considerable insight.

As noted above, one of the central challenges of this approach is that the principal stratum effect of interest is not in general identified because we do not in general know which individuals are in which principal stratum. As a result, most of the methodological approaches to the analysis of principal strata effects use either (i) bounds for the principal strata effects or (ii) sensitivity analysis techniques or (iii) take a Bayesian approach. With a sensitivity analysis approach, one does not obtain a single point estimate but rather different estimates for each possible value of the sensitivity analysis parameters. With a Bayesian approach, because of lack of point identification, the length of posterior intervals will not shrink to 0 as the sample size increases to infinity and the posterior will depend on the prior even as the sample size tends to infinity (Richardson et al., 2011). Here we will present one particularly simple sensitivity analysis approach (Chiba and VanderWeele, 2011).

Assessing the survivor average causal effect is made considerably easier by an assumption sometimes referred to as "monotonicity." Stated formally, the monotonicity assumption [assumption (A8.1)] is $S_0 \leq S_1$. The assumption states that, for all individuals, survival under the treatment is always at least as good as survival under the control condition. In other words, survival under control implies survival under treatment, and death under treatment implies death under control. If treatment cannot render death more likely than the control condition for any individual, this will be a reasonable assumption. Note that the assumption would also hold if treatment had no effect on survival. The sensitivity analysis approach expresses the principal stratum effect as the difference between the crude comparison of the outcomes across arms amongst survivors and a sensitivity analysis parameter. Specifically, if treatment A is randomized and the monotonicity assumption holds, then the survivor average causal effect, $\mathbb{E}[Y_1 - Y_0 | S_0 = 1, S_1 = 1]$, is equal to

$$\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1] - \alpha$$

where $\alpha = \mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1]$. The result states that, to obtain the survivor average causal effect, one can use the crude difference in outcomes Y between the treated and control subjects amongst those who survived, $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$, and then subtract the sensitivity analysis parameter α. The sensitivity analysis parameter α is set by the investigator according to what is thought plausible. The sensitivity analysis parameter can be varied over a range of plausible values to examine how conclusions vary under different values for the parameter. To obtain confidence intervals for the survivor average causal effect for a fixed value of the parameter α, one can simply subtract α from the upper and lower confidence limits for $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$.

The parameter itself is the average difference in the outcome that would have been observed under treatment comparing two different populations: The first population is the population that would have survived under treatment $(A = 1, S = 1)$; the second population is the population that would have survived without treatment $(A = 0, S = 1)$. Because the second population consists of individuals who survived even without treatment, it will likely overall be a healthier population than the population who would have survived with treatment. The interpretation of α then is simply the difference in expected outcomes under treatment for these two populations. This simple approach, however, only works under the monotonicity assumption (A8.1) that survival under the treatment is always at least as good as survival under the control condition. Assessing the survivor average causal effect is more complicated when this monotonicity assumption does not hold.

The fact that the population who would have survived without treatment is likely healthier overall than the population who would have survived with treatment also gives rise to a sufficient condition under which the crude comparison of the outcome is conservative for the survivor average causal effect. This will require one further assumption [i.e., (A8.2)]: $\mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1] \leq 0$. This second assumption requires that the sensitivity analysis parameter $\alpha = \mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1]$ be less than or equal to 0. If the outcome Y were quality of life and if it were indeed the case that the population who

would have survived without treatment $(A = 0, S = 1)$ is healthier overall than the population who would have survived with treatment $(A = 1, S = 1)$, and if owing to the fact that this former group was healthier, it also would have had higher quality of life outcomes under treatment, then assumption (A8.2) would be satisfied. Under the monotonicity assumption (A8.1) and assumption (A8.2), the crude comparison of the outcomes Y between the treated and control subjects amongst those who survived will give a lower bound for the survivor average causal effect.

If quality of life were the outcome of interest, then, under assumptions (A8.1) and (A8.2), we could use the data to estimate $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$. And, we would know that this crude estimate was conservative for the survivor average causal effect—that is conservative for the extent to which the treatment increased quality of life amongst the subpopulation that would have survived irrespective of whether treatment was given or not.

Note that if assumption (A8.2) is reversed so that $\mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1] \geq 0$, then the conclusion would be modified such that the survivor average causal effect is less than or equal to the crude estimate $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$. If the outcome is quality of life, then the "greater than or equal to" result stated above will be the result of interest. However, we give an illustration where the reverse of (A8.2) is likely plausible and in which the "less than or equal to" result is applicable.

Example. The ARDSnet clinical trial (Acute Respiratory Distress Syndrome Network, 2000) compared two methods of ventilation for patients with acute lung injury and acute respiratory distress syndrome. The two methods compared were low-volume ventilation $(A = 1)$ and traditional volume ventilation $(A = 0)$. The study found a significant decrease in 180-day mortality $(p = 0.003)$ comparing the low-volume group, 146/473 (31%), and the traditional volume group, 173/429 (40%). The study also assessed a variety of outcomes that were only defined for those who had survived up through 180-day mortality. One of these outcomes was days to return home. It was found that those in the low-volume group required fewer days: -7.15 (95% CI: $-13.73, -0.56$), to return home (mean 33.55) as compared with the traditional volume group (mean 40.70). These means were calculated amongst those who survived 180 days. Because survival in the low-volume group was higher, some of the individuals who survived in the low-volume group may have died had they been in the traditional-volume group. The crude comparisons of means may thus not be an adequate measure of the extent to which the low-volume ventilation method decreases days to return home. We might thus instead be interested in the effect of low-volume ventilation versus traditional-volume amongst the subset that would have survived under either ventilation method—that is, the survivor average causal effect (the effect within the principal stratum in which $S_1 = S_0 = 1$).

The low-volume ventilation method significantly reduced mortality, and the monotonicity assumption (A8.1) may be reasonable; we will, however, return to this point below. Under the monotonicity assumption, we can obtain different estimates of the survivor average causal effect under different specifications of the sensitivity analysis parameter: $\alpha = \mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1]$. The

sensitivity analysis parameter compares the days to return home under the low-volume ventilation method between two populations: the population that would have survived under low-volume ventilation and the population that would have survived under traditional ventilation. Because traditional ventilation is more likely to result in mortality, those who would have survived under traditional ventilation are likely a healthier population and one that is more likely to return home sooner if given low-volume ventilation. If we thought the difference between these populations were small, we might specify a difference to return home of $\alpha = 1$ day, which, under monotonicity, would give an estimate of the principal strata effect of -8.15 (95% CI: $-14.73, -1.56$); if we thought the difference in the populations were somewhat larger, we might specify a difference of $\alpha = 4$ days, which, again provided that the monotonicity assumption held, would give an estimate of the principal strata effect of -11.15 (95% CI: $-17.73, -4.56$). Irrespective of the actual value of α, if we thought that the population who would survive under traditional ventilation was indeed healthier and would return home sooner with low-volume ventilation than the population who would survive under low-volume ventilation, then we would have that $\alpha \geq 0$—that is, the reverse of assumption (A8.2). We thus would have the survivor average causal effect was in fact less than or equal to the crude estimate, $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$. We would then conclude that -7.15 days (95% CI: $-13.73, -0.56$) was an upper bound on—that is, conservative for—the principal strata direct effect.

We have focused on the setting of a clinical trial in which the primary treatment of interest is randomized. The approach above, however, also holds in an observational study if the effect of the treatment on the outcome is unconfounded conditional on some set of covariates C. If the effect of the treatment on the outcome is unconfounded conditional on C, then assumption (A8.2) also needs to be modified to be conditional on C. The results above also assumed that the monotonicity assumption (A8.1) holds—that is, that the treatment would, for no individual, cause death. Other approaches have been developed to assess the survivor average causal effect when this monotonicity assumption does not hold (Hayden et al., 2005; Chiba and VanderWeele, 2011), but these are more difficult to implement and require the specification of multiple sensitivity analysis parameters.

8.1.2. Principal Stratification and Noncompliance

The principal stratification framework has also been of use in addressing issues of noncompliance (Angrist et al., 1996; Imbens and Rubin, 1997; Balke and Pearl, 1997; Cheng and Small, 2006; Cuzick et al., 2007; Little et al., 2009). Consider a randomized trial with noncompliance in which the group assigned the placebo had no access to treatment (i.e., no "defiers"—no one for whom $S_0 = 1, S_1 = 0$). Let A be the randomized treatment, S compliance status (assume that compliance is all or nothing), and Y the outcome. The overall affect of treatment assignment on the outcome, $\mathbb{E}[Y_1 - Y_0]$, is simply given by a comparison of the average outcomes in the treatment versus the control groups, $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$.

The effect is sometimes referred to as the intent-to-treat estimate. We might, however, instead be interested in the effect of the treatment taken (not simply assigning treatment). Sometimes the intent to treat estimate is divided by the proportion of compliers, $\frac{\mathbb{E}[Y|A=1]-\mathbb{E}[Y|A=0]}{\mathbb{E}[S|A=1]-\mathbb{E}[S|A=0]}$, as an alternative estimate of the treatment effect. This is sometimes referred to as the instrumental variable (IV) of the treatment effect.

Suppose now that we were willing to assume that there was no effect of treatment assignment A on the outcome Y except through the actual treatment taken (compliance status) S. This assumption is sometimes referred to as the "exclusion restriction." Suppose further we were willing to assume that there were no defiers—that is, no one who would take the treatment if assigned to the control group, but who would not take the treatment if assigned to the treatment group (no one with $S_0 = 1, S_1 = 1$). Angrist et al. (1996) showed that under these assumptions, the instrumental variable estimator above, $\frac{\mathbb{E}[Y|A=1]-\mathbb{E}[Y|A=0]}{\mathbb{E}[S|A=1]-\mathbb{E}[S|A=0]}$, would in fact be a consistent estimator for the treatment effect among the compliers, $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$. Thus this estimator gives the effect of treatment itself but only for the compliers—that is, for those who would take the treatment if assigned to the treatment group and would not if assigned to the control group. In other words, this treatment effect estimate pertains only to the group for which treatment assignment actually changes treatment taken. In a sense, this is the best we might be able to hope for because for the "never-taker" principal stratum ($S_0 = 0, S_1 = 0$) we don't have any data on what would have happened to them had they actually taken treatment, and for the "always-taker" principal stratum ($S_0 = 1, S_1 = 1$) we don't have data on what would have happened on any of them had they not taken treatment. The principal stratum causal effect that is actually estimated, $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$, is sometimes now referred to as the local average treatment effect (LATE) or the complier average causal effect (CACE). We will return to questions of non-compliance in our discussion of instrumental variables in Section 8.3. Note, however, that here, unlike the previous subsection on principal stratum effects for truncation by death or post-infection outcomes, in the context of noncompliance we actually can identify one of the principal stratum causal effects from the data. We can do this because of the assumption of "no defiers" and the assumptions of no effect of treatment assignment A on the outcome Y except through the actual treatment taken (compliance status) S.

8.1.3. Principal Stratification and Surrogate Outcomes

This principal stratification approach has also been proposed to assess questions of surrogate outcomes (Frangakis and Rubin, 2002; Gilbert and Hudgens, 2008; Wolfson and Gilbert, 2010; Li et al., 2010b). We will consider surrogate outcomes in greater detail in Section 8.2. Here we just briefly point out the connections with principal stratification. The motivation for considering surrogate outcomes is that in certain randomized trials it may be very expensive or require considerable follow-up to assess the outcomes of interest. If measurements on a surrogate that is closely related to the outcome are easier or cheaper to obtain, then one might

analyze the effect of the treatment on the surrogate rather than the effect of the treatment on the primary outcome of interest. Frangakis and Rubin (2002) gave a definition of a "principal surrogate" that they argued was important in finding a good surrogate. With a binary outcome, they said that S is a principal surrogate for the effect of A on Y if for all s, $\mathbb{E}[Y_1 - Y_0|S_0 = S_1 = s] = 0$; that is, for the principal strata in which treatment does change the surrogate ($S_0 = s, S_1 = s$), the treatment should have no effect on the outcome. This property is sometimes referred to as one of "causal necessity."

Building on the work of Frangakis and Rubin (2002), Gilbert and Hudgens (2008) defined the "causal predictiveness surface" as $\mathbb{E}[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$. This is simply the effect of treatment on the outcome in each of the principal strata. In some sense this "causal predictiveness surface" could be used to think about how the effect of the treatment on the surrogate relates to the effect of treatment on the outcome. As was the case with the outcomes truncated by death and post-infection outcomes, identification of such principal strata effects can be difficult but various methodological approaches have been proposed (Gilbert and Hudgens, 2008; Wolfson and Gilbert, 2010; Li et al., 2010b). Again we will consider surrogate outcomes at greater length in the following section.

8.1.4. Principal Stratification and Mediation

There has been some interest in applying ideas of principal stratification to questions of mediation (Gallop et al., 2009; Elliott et al., 2010). Frangakis and Rubin (2002) discuss two types of principal strata causal effects. They call an effect of treatment A on outcome Y within the principal strata in which A does not change S a "dissociative effect," that is, $\mathbb{E}[Y_1 - Y_0|S_0 = S_1 = s]$, and they call an effect of treatment A on the outcome Y within the principal strata in which A does change S an "associative effect," that is, $\mathbb{E}[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$ when $s_0 \neq s_1$. Let us first examine the dissociative effect. The dissociative effect is the effect of treatment on outcome within the principal strata in which treatment doesn't change the intermediate. If, within these principal strata, the treatment doesn't change the intermediate, then its effect cannot operate through the intermediate; it must be "direct." We might thus also call the dissociative effect a "principal strata direct effect" (PSDE). For a binary intermediate we would have two principal strata direct effects, namely, $PSDE(0) = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 0]$ and $PSDE(1) = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 1]$. If one of these were nonzero, then we would conclude that there were some pathway from treatment to outcome not through the intermediate (VanderWeele, 2008). This much seems relatively unproblematic. We still may have the same difficulties with the identification of these principal strata effects from observed data, but sensitivity analysis techniques can be used to address principal strata direct effects (Sjölander et al., 2009; VanderWeele, 2010a; Chiba, 2010) and Bayesian methods have also been employed (Gallop et al., 2009; Elliott et al., 2010).

The principal stratification framework then has a coherent notion of a direct effect. One might then be tempted to take the associative effects (the effect of

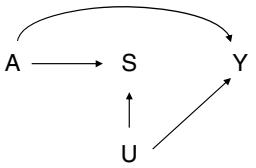

Figure 8.1 Example with randomized treatment A, surrogate S, and outcome Y, with no effect of the surrogate on the outcome but an unmeasured common cause of the surrogate and outcome.

treatment on the outcome when treatment does change the intermediate)—that is, $\mathbb{E}[Y_1 - Y_0|S_0 = s_0, S_1 = s_1]$ with $s_0 \neq s_1$—as a measure of an indirect effect. As we will see, however, this does not work. The problem is that these "associative effects" include the overall effect of treatment within the relevant principal strata. Whatever we might call the "direct effect" and the "indirect effect," the associative effect will pick up both of them. We might in fact have very large associative effects with no "indirect effects" whatsoever.

Consider the setting depicted in Figure 8.1 and suppose that S serves as a very good proxy for Y but has no actual effect on Y whatsoever. In this case, none of the effect would be mediated by S because S has no effect on Y; however, the associative effect $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$ might be large because whenever treatment changes S from 0 to 1, treatment is likely to also have an effect on Y since S serves as a good proxy for Y (VanderWeele, 2011d).

Further insight into why associative effects cannot be used as measures of indirect effects can be gained by comparing them to the natural direct and indirect effects we have discussed in previous chapters (Robins and Greenland, 1992; Pearl, 2001). As before, we consider hypothetical interventions on the treatment A and the intermediate S so that it is possible to define the potential outcome Y_{as}, the value of Y for each individual that would be observed if A were set to a and S were set to s. The average natural direct effect may then be defined as $\mathbb{E}[Y_{1S_0} - Y_{0S_0}]$—that is, the comparison of the outcome with versus without treatment in both scenarios setting the intermediate to the level it would have been without treatment. The average natural indirect effect can be defined as $\mathbb{E}[Y_{1S_1} - Y_{1S_0}]$—that is, a comparison of the outcomes under treatment when setting the intermediate to the level it would have been with versus without treatment. The natural indirect effect will thus only be nonzero if, for some individual, treatment changes the value of the intermediate, and this change in the value of the intermediate changes the value of the outcome. With a nonzero natural indirect effect we have mediation: The treatment changes the outcome by changing the intermediate. As we have seen before, a total effect can be decomposed into a natural direct and indirect effect, $\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_{1S_1} - Y_{1S_0}] + \mathbb{E}[Y_{1S_0} - Y_{0S_0}]$, even in models with interactions and nonlinearities (Pearl, 2001). Previous chapters have considered methods and sensitivity analysis techniques for assessing these natural direct and indirect effects.

Let us now return to considering the associative effect. We can express the relationship between the associative effect and natural direct and indirect effects as

follows:

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1] &= \mathbb{E}[Y_{1S_1} - Y_{0S_0}|S_0 = 0, S_1 = 1] \\ &= \mathbb{E}[(Y_{1S_1} - Y_{1S_0}) + (Y_{1S_0} - Y_{0S_0})|S_0 = 0, S_1 = 1] \\ &= \mathbb{E}[(Y_{11} - Y_{10}) + (Y_{10} - Y_{00})|S_0 = 0, S_1 = 1]\end{aligned}$$

From the second line we see that the associative effect is the sum of the natural direct and indirect effects within the principal strata $(S_0 = 0, S_1 = 1)$. From the final line we see that even if there were no effect of S on Y so that $(Y_{11} - Y_{10}) = 0$, we could still have a substantial associative effect if the direct of A, i.e. $Y_{10} - Y_{00}$, were non-zero. Again, associative effects do not correspond to indirect effects and thus cannot be used to assess mediation. There is nothing within the principal stratification framework that corresponds to a measure of an indirect effect. The "associative" and "dissociative" may be of interest in their own right but one is not assessing mediation in these cases; one is not assessing whether treatment affects the outcome through the intermediate. As discussed above, one can use principal strata direct effects to assess whether there is a pathway from treatment to the outcome other than through the intermediate - and so "direct effects analysis" may still be an appropriate description; but one cannot assess, using principal strata effects, whether there is a pathway through the intermediate itself.

As an example to illustrate these points, Page (2012) uses the principal stratification framework within the context of a randomized Career Academies intervention with subsequent earnings as the outcome and with "exposure to the world of work" as an intermediate. Using a principal stratification approach she finds that the program had substantial effect on subsequent earnings for those for whom the program would change exposure to the world-of-work, but not for those for whom it would not change exposure to the world-of-work. Here the dissociative effects, the effects of the treatment on the outcome when the treatment does not change the intermediate, $PSDE(0) = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 0]$ and $PSDE(1) = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 1]$, seem to be zero, where as the associative effect, the effect of the treatment on the outcome when the treatment does change the intermediate, $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$, appears to be non-zero. One possible explanation of these results would be that the treatment changes the intermediate which in turn changes the outcome; i.e. that we have mediation. But that is not what these principal strata effects assess. Another possible explanation of Page's results are that the entirety of the effect is direct (i.e. not by changing the intermediate e.g. the intermediate has no effect at all on the outcome) but that the effect of the treatment on the intermediate serves as a good proxy for the effect of the treatment on the outcome. Suppose, for example, different students had different levels of baseline pretreatment motivation. It may be the case that students with low motivation $(U = 0)$ are unresponsive to treatment both in terms of exposure to the world of work (either $S_0 = 0, S_1 = 0$ or $S_0 = 1, S_1 = 1$) and in terms of earning $(Y_1 = Y_0)$. In this case, both of the principal strata direct effects will be zero. Suppose that students with high motivation $(U = 1)$ may be responsive to treatment both in terms of exposure to the world of work (so that $S_0 = 0, S_1 = 1$) and in terms of earnings $(Y_1 > Y_0)$; then

the associative principal stratum effect $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$ will be nonzero. We may have zero principal strata direct effects and a nonzero associative principal strata effect, as in Page (2012), even if exposure to the world of work has no effect on earnings and thus does not serve as a mediator at all. This may occur because it may be only the high motivation students for whom their earnings are responsive to treatment even if all of this effect is direct. It is not necessarily the case that exposure to the world of work is not a mediator, only that the analyses of Page (2012) are perfectly consistent with it not being a mediator. The principal stratification framework does not assess mediation.

Analyses of effects within principal strata are in fact more akin to subgroup analyses than to mediation analyses. As discussed in Chapter 2, social scientists generally distinguish between "mediators" and "moderators." A mediator is a variable on the pathway from the treatment to the outcome. A moderator is a variable such that the effect of the treatment on the outcome differs for different levels of the moderator variable. In a randomized trial, for instance, we might estimate the effect of the treatment on the outcome within strata of baseline pretreatment covariates. This would be a moderator analysis. Within the principal stratification framework, the principal strata are effectively viewed as pretreatment characteristics of an individual (Frangakis and Rubin, 2002). Assessing the effect of treatment on the outcome across different principal strata is essentially looking at whether the effect of treatment—the total effect—is different for different subgroups. In the terminology of social scientists, this would be a "moderator" analysis. Effects within principal strata are akin to moderator or subgroup analyses, not mediation analyses.

Additional assumptions have sometimes been employed within the principal stratification framework to better assess mediation (cf. Jo et al., 2011). An example of such an assumption is that the "direct effects" are the same across all principal strata. If this were the case and we found that the principal strata direct effects, $PSDE(0) = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 0]$ and $PSDE(1) = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 1]$, were zero and the associative effect, $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$, were nonzero, then one might think we had evidence for mediation. However, such assumptions are difficult even to state within the principal stratification context (because just using principal stratification we cannot even define the "direct effect" in the stratum $S_0 = 0, S_1 = 1$), and this begs the question why one does not simply use other concepts, like natural direct and indirect effects, discussed in Chapters 2–7, in assessing mediation. As already discussed, one of the central advantages of the principal stratification framework is that potential outcomes are only defined with respect to the exposure (not the intermediate). This is attractive when it is difficult to think of well-defined interventions on the intermediate. However, in the context of mediation, this is also one of the central drawbacks of the principal stratification framework. When we are speaking of mediation, we are speaking of (a) an exposure changing the mediator and (b) this change in the mediator going on to change the outcome. In other words, we are speaking of effects of the mediator. To rigorously discuss effects of the mediator, we need to define potential outcomes with respect to the mediator. Attempts to discuss mediation without potential outcomes for the effect of the mediator end up being convoluted and ultimately unsuccessful. Principal stratification is simply not the right tool for mediation.

8.1.5. Principal Stratification: Uses and Limitations

The principal stratification has shed considerable light on noncompliance, the analysis of instrumental variables, settings in which the outcome is censored by death, and settings in which a post-infection outcome is in view. It is also a potentially useful concept in analysis of surrogate outcomes and it can be of some use for thinking about whether there may be a pathway from treatment to outcome other than through a particular intermediate. However, it is not of use in mediation analysis itself, conceived of as assessing whether there is an effect of the treatment on the outcome that operates through the intermediate. The framework does not assess indirect effects.

A couple of further caveats are also worth mentioning. First, as noted above, even after we have used statistical and sensitivity analysis techniques to assess principal strata effects, the principal strata themselves in general remain unidentified. We do not know who is in which stratum, and this makes the framework somewhat more difficult to use in informing policy questions. Second, in describing the various applications, we have considered the setting of a binary intermediate. Although the framework is not, in principle, restricted to the setting of a binary intermediate (Frangakis and Rubin, 2002), the analysis becomes much less tractable with intermediates with more categories. If the intermediate is continuous, there may be no more than one individual in any of the principal strata. Dichotomization of a continuous or ordinal intermediate, as is often done, can give rise to misleading inferences (Robins et al., 2007). Identification difficulties are also compounded when the intermediate has more than two levels. The notions of principal stratification in practice thus seem most useful in settings in which the intermediate is in fact binary such as all-or-nothing compliance, censoring-by-death versus survival, and the analysis of post-infection outcomes.

8.2. SURROGATE OUTCOMES

In randomized trials it is sometimes expensive or difficult (e.g., it takes a very long follow-up) to measure the outcome of interest. In such cases it might be desirable to measure instead a surrogate for the outcome that is closely related to it. If the surrogate is easier or cheaper to measure than the outcome and is closely related to the outcome, then investigators could perhaps assess the effect of the treatment on the surrogate rather than the effect of treatment on the outcome. This setting raises important questions with regard to how to best evaluate surrogates and what sorts of conclusions one can draw from surrogates. Surrogates are somewhat related to questions of mediation—we might expect most good surrogates to be on the pathway from the treatment to the outcome—but, as will be seen below, a good surrogate need not, in fact, be a mediator.

Unfortunately, when using surrogates rather than the outcome itself, a phenomenon can arise that is sometimes referred to as the surrogate paradox. It may be the case that the treatment has a positive effect on the surrogate and that the surrogate and outcome are strongly positively associated but that the treatment itself has a negative effect on the outcome! This was illustrated dramatically is the case

of trials evaluating the effects of certain drugs on ventricular arrhythmia, taken as a surrogate for mortality. Ventricular arrhythmia is strongly associated with mortality; several drugs were tested in randomized trial, were found to lower ventricular arrhythmia, and were approved by the Food and Drug Administration. However, in follow-up it became clear that the drugs increased, rather than decreased, mortality (Moore, 1995; Fleming and DeMets, 1996). According to some estimates, as many as 50,000 persons died because of the use of these drugs (Moore, 1995). One important task then with regard to surrogate outcomes is determining when data concerning the effect of treatment on the surrogate can be used to make decisions about the direction of the effect of the treatment on an outcome.

Here we will describe some of the definitions and criteria that have been proposed for surrogacy. We will also give criteria that suffice to avoid the surrogate paradox, and we will describe some statistical approaches and measures that have been proposed in the context of surrogate outcomes. As will be seen, the conditions and approaches are related to, but also distinct from, the concepts, questions, and methods used to assess mediation.

8.2.1. Definitions for Surrogates

Let A be a treatment of interest that we will assume randomized, let Y be the outcome of interest, and let S be a proposed surrogate. As above, let Y_a and S_a be counterfactual outcomes (or potential outcomes) for Y and S for each individual that would have been obtained if treatment A had, possibly contrary to fact, been set to a. Finally let Y_{as} be the counterfactual outcome for each individual that would have been obtained if A had been set to a and if S had been set to s.

In what is now considered a classic paper, Prentice (1989) suggested that a surrogate should be such that a test of the null of no effect of the treatment A on surrogate S should serve as a valid test of the null of no effect of the treatment A on outcome Y. Prentice proposed two central criteria for assessing this, and a variable satisfying such criteria has subsequently been referred to as a "statistical surrogate" (Frangakis and Rubin, 2002). His criteria indicated that S would be a good (statistical) surrogate for the effect of A on Y if (i) Y is independent of A conditional on S and (ii) S and Y are correlated. The criteria are suggested by the diagram in Figure 8.2. Suppose there is no controlled direct effect of A on Y; then if there is no effect of A on S, it then follows also that there will be no effect of A on Y. Moreover, in this diagram if there is no direct effect of A on Y, then A will be independent of Y conditional on S. But the criteria do not give the desired result if there are unmeasured confounders of S and Y as in Figure 8.3. There could be correlation between A and Y conditional on S due to U even if A has no direct effect on Y. The Prentice criterion might only be a reasonable requirement if we could control for the common causes of S and Y.

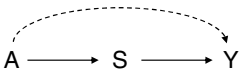

Figure 8.2 Example with randomized treatment A, surrogate S, and outcome Y.

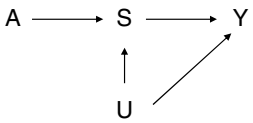

Figure 8.3 Example with randomized treatment A, surrogate S, and outcome Y, with no direct effect of the treatment on the outcome but an unmeasured common cause of the surrogate and outcome.

Prompted perhaps in part by these concerns, Frangakis and Rubin (2002) proposed an alternative criterion, already suggested in Section 8.1.3, to evaluate surrogates and referred to a surrogate that satisfied this criterion as a "principal surrogate." Frangakis and Rubin (2002) defined S to be a principal surrogate for the effect of A on Y if for all s, $P(Y_1|S_1 = S_0 = s) = P(Y_0|S_1 = S_0 = s)$. Essentially a principal surrogate requires that whenever the treatment does not change the surrogate ($S_1 = S_0 = s$) there is no difference in the distribution of potential outcomes with versus without treatment. If a surrogate satisfied this property, then an effect of A on Y will be present only if an effect of A on S is present. If Y is binary, then the definition of a principal surrogate is equivalent to $\mathbb{E}(Y_1 - Y_0|S_1 = S_0 = s) = 0$, a condition that may be referred to as no principal strata direct effects (VanderWeele, 2008). This has likewise also been referred to as the property of "average causal necessity" (Gilbert and Hudgens, 2008).

Lauritzen (2004) proposed a slightly stronger definition related to surrogacy that he referred to as a "strong surrogate." Lauritzen defined S is to be a strong surrogate for the effect of A on Y if there were no controlled direct effects of A on Y not through S—that is, if the controlled direct effects $Y_{1s} - Y_{0s} = 0$ for all s as in Figure 8.3. A strong surrogate is also a principal surrogate (Lauritzen, 2004; VanderWeele, 2008), but the reverse implication does not hold because principal surrogacy only requires no direct effects when $S_1 = S_0 = s$ and only requires this in distribution, not for all individuals. Note also that a strong surrogate will be a statistical surrogate if there is no common cause of the surrogate and the outcome as in Figure 8.2, but a strong surrogate need not be a statistical surrogate if there is such a common cause as in Figure 8.3.

Unfortunately, none of these criteria—statistical surrogate, principal surrogate, or strong surrogate—protect against the surrogate paradox. Chen et al. (2007) gave an example showing that a variable could be a principal surrogate or even a strong surrogate and that it could still be the case that the treatment had a positive effect on the surrogate, with the surrogate and the outcome strongly positively correlated, yet with a negative effect of the treatment on the outcome. This is the phenomenon we will refer to as the surrogate paradox. Following this we will say that S is a consistent surrogate for the effect of A on Y if (a) when S and Y are positively associated, a nonpositive (non-negative) average causal effect of A on S implies a nonpositive (non-negative) average causal effect of A on Y or (b) when S and Y are negatively associated, a nonpositive (non-negative) average causal effect of A on S implies a non-negative (nonpositive) average causal effect of A on Y. We will then say that a surrogate that is not a consistent surrogate exhibits the surrogate paradox.

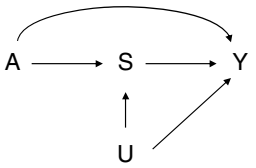

Figure 8.4 Example with randomized treatment A, surrogate S, and outcome Y, with a direct effect of the treatment on the outcome and with an unmeasured common cause of the surrogate and outcome.

Statistical surrogates, principal surrogates, and strong surrogates are all subject to the surrogate paradox. This is somewhat surprising as the notions of a principal surrogate and a strong surrogate are already quite stringent; it is also rather disturbing in that such effect reversal seems to completely undermine the value of a surrogate marker. Fortunately some conditions can be given under which the surrogate paradox is avoided.

8.2.2. Conditions for Consistent Surrogates to Avoid the Surrogate Paradox

Chen et al. (2007) and Ju and Geng (2010) gave sufficient conditions concerning avoiding the surrogate paradox when there was no direct effect of the treatment A on the outcome Y not through S as in Figure 8.3. However, conditions can in fact be given more generally that allow for a direct effect of the treatment on the outcome as in Figure 8.4 (VanderWeele, 2013d). Consider the causal diagram in Figure 8.4. If it is the case that (a) both A and S have a positive effect on Y on average [in the sense that $\mathbb{E}(Y|a,s,u)$ is nondecreasing in a and s for all u] and (b) A positively affects S in distribution (in the sense that $P(S > s|a,u)$ is nondecreasing in a for all s,u), then A has a positive effect on Y; that is, S will be a consistent surrogate (VanderWeele, 2013d).

The result remains true if in both conditions (a) and (b), "nondecreasing" is replaced by "nonincreasing." If in only one of conditions (a) or (b), "nondecreasing" is replaced by "nonincreasing," then the conclusion is changed to that: A has a negative effect on Y, but we will still have a consistent surrogate (and avoid the surrogate paradox) insofar as the direction of the effect of the treatment A on the outcome Y is what we would expect given the effects of A on S and of S on Y. Note that to avoid the surrogate paradox (i.e., to ensure a consistent surrogate) a nonnegative average causal of A on S is not sufficient; rather one needs the effect to be non-negative in the distributional sense that $P(S > s|a,u)$ is nondecreasing in a for all s,u; this is sometimes referred to as "distributional monotonicity" (VanderWeele et al., 2008; VanderWeele and Robins, 2009b, 2010).

These conditions to avoid the surrogate paradox can also be given a fairly intuitive interpretation. Suppose that in a randomized trial we find a positive average causal effect of A on S and we know that S and Y are strongly positively correlated. This is often the setting encountered with surrogate outcomes. In this setting, under what circumstances might the surrogate paradox arise? When might the effect of A on Y be negative rather than positive? The conditions above state that at least

one of three things must occur if we are to get this effect reversal. First, there may be a negative direct effect of A on Y not through S (i.e., the first part of assumption (a) that $\mathbb{E}(Y|a,s,u)$ is nondecreasing in a may be violated). Second, it may be the case that although S and Y are positively correlated, this may not indicate the actual causal relationship of S on Y; the association may be due to confounding by U [i.e., the second part of assumption (a) that once we condition on U, $\mathbb{E}(Y|a,s,u)$ is nondecreasing in s may be violated]. Third, even if neither of these first two phenomenon occur, it may be the case that even though A positively affects S on average and S positively affects Y, A may not positively affect S for all individuals; it may decrease S, and thus decrease Y for some individuals; we may have a lack of transitivity [i.e., assumption (b), the assumption concerning distributional monotonicity which guarantees that this is avoided, may be violated]. In summary, if the surrogate paradox is to occur, we need either (i) a direct effect of A on Y not through S in the opposite direction, (ii) confounding for the effect of S on Y, or (iii) a lack of transitivity so that A does not positively affect S for the same individuals for which S positively affects Y (VanderWeele, 2013d). In thinking about whether the surrogate paradox might occur and whether one ought to draw conclusions concerning an outcome of interest from the analysis of the results concerning a surrogate, an investigator could think through each of these three possibilities. The conditions above state that at least one of them must occur if the surrogate paradox is to arise. If an investigator can, on substantive grounds, rule out each of these three possibilities, then the surrogate can be used instead of the outcome at least to draw conclusions about the correct direction of the effect of the treatment on the outcome of interest.

8.2.3. Mediation and Surrogacy

Several different statistical approaches and measures of surrogacy have been proposed in the literature. Joffe and Greene (2009) discuss four different empirical approaches that have been proposed to measure the extent of surrogacy, and they derived relations between them under linear model assumptions. We will revisit each of these four approaches and discuss the extent to which they ensure consistent surrogates. These four approaches could broadly be described as (i) a "proportion-explained" approach, (ii) an "indirect effects" approach, (iii) a "meta-analytic" approach, and (iv) a "principal stratification" approach. Each of these approaches may tell us something about the role that the surrogate S plays in the relationship between treatment A and outcome Y. In this subsection we will consider the "proportion-explained" approach and the "indirect effects" approach and we will discuss the relationships between mediation and surrogacy. In the next two subsections we then consider the "meta-analytic" approach and "principal stratification" approaches.

Freedman et al. (1992) proposed using a "proportion explained" measure to assess surrogacy. Suppose one were to regress the outcome Y on the exposure A:

$$\mathbb{E}(Y|A=a) = \Phi_0 + \Phi_1 a$$

and then regress the outcome Y on the exposure A and the surrogate S:

$$\mathbb{E}(Y|A=a,S=s)=\theta_0+\theta_1 a+\theta_2 s$$

The proportion of the total effect explained by the surrogate is then taken as

$$(\Phi_1-\theta_1)/\Phi_1 \tag{8.1}$$

which is equivalent to $1-\theta_1/\Phi_1$. Statistical inference for this measure is also described by Lin et al. (1997). This is of course similar to the difference method for mediation, and the measure in (8.1) is similar to the proportion mediated measure also described in Chapter 2. As noted in Chapter 2, in the setting of mediation analysis, this approach is problematic if there is confounding of the effect of S on Y by U; this can occur even if treatment A is randomized since the surrogate S is generally not randomized. Because of this, confounding using the proportion in (8.1) as a measure of mediation would be biased. However, in the context of surrogacy (rather than mediation) if the goal is simply to assess how much of the effect of A on Y can be predicted by the effect of A on S, these concerns about confounding may be less relevant. If U is a common cause of S and Y and, because of U, S gives important information about Y, then S may still be a good surrogate insofar as it may be possible to predict the sign of the effect of A on Y from the sign of the effect of A on S. Although the measure in (8.1) of the "proportion explained" may thus serve as a useful measure, it is not immune to the surrogate paradox. An example is given below in which the average causal effect of A on S is positive, the average causal effect of S on Y is positive, "proportion explained" is 100%, but the effect of A on Y is negative. This can occur because it may be the case that A does not positively affect S for the same individuals for which S positively affects Y. Nothing in the "proportion explained" measure guarantees the distributional monotonicity assumption needed to avoid the surrogate paradox. Thus even if a surrogate is judged to be "good" from the standpoint of having a high proportion explained, this does not guarantee that the surrogate is consistent.

Another alternative approach to assessing surrogacy is what Joffe and Greene (2009) call the "indirect effects" approach. This approach relies on the mediation concepts we have already considered in previous chapters, namely, natural indirect effects (Pearl, 2001). As before, the average natural indirect effect is defined as $\mathbb{E}(Y_{1S_1}-Y_{1S_0})$ and measures the effect comparing setting the treatment to present with the surrogate set to what it would have been with versus without the treatment (Robins and Greenland, 1992; Pearl, 2001; Taylor et al., 2005). For it to be nonzero the treatment must have an effect on the surrogate (i.e., S_1 and S_0 must differ) and then the surrogate must have an effect of the outcome (i.e., the change in the surrogate from S_0 to S_1 must have an effect on Y). This is thus sometimes referred to as a "mediated effect." A measure of surrogacy may then be taken as the "proportion mediated"—that is, the proportion of the natural indirect effect to the total effect:

$$\frac{\mathbb{E}(Y_{1S_1}-Y_{1S_0})}{\mathbb{E}(Y_1-Y_0)} \tag{8.2}$$

We considered the conditions for identification and estimation of the natural direct and indirect effect in Chapter 2, and the reader is referred back to that chapter for further discussion of identification. The advantage of this approach to surrogate measures is that, provided that the natural indirect effect has been correctly identified and estimated, it gives the actual effect of the treatment on the outcome through the surrogate. Likewise, the natural direct effect, $\mathbb{E}(Y_{1S_0} - Y_{0S_0})$, can be used to assess whether there is an effect of the treatment on the outcome not through the surrogate and one could evaluate whether this was in the opposite direction of the direct effect. The natural indirect and direct effects sum to the total effect. Thus, if the natural direct and indirect effects were known, this could be useful in diagnosing the surrogate paradox if these two effects were in opposite directions. The difficulties are, however, effectively transferred to the challenge of identifying and consistently estimating the natural indirect effect, $\mathbb{E}(Y_{1S_1} - Y_{1S_0})$. As noted in Chapter 2, the identification conditions needed to identify this natural indirect effect are quite strong, which constitutes a disadvantage to this approach. However, within the "indirect effects" approach, the criterion generally used to assess whether a surrogate is "good" (whether the proportion mediated is large) unfortunately does not guarantee that a surrogate is consistent. As will be seen in the illustration below, we can in fact have a high proportion mediated (even 100% mediated) in settings in which S exhibits the surrogate paradox. Although the natural direct and indirect effects themselves (if known) could be useful in diagnosing the surrogate paradox, the proportion-mediated criterion itself does not ensure that a surrogate is consistent.

The "indirect effects" approach, taken as a measure of surrogacy, also suffers from another problem. Consider the causal diagram in Figure 8.1 in which the surrogate S has no effect on the outcome Y. Now it may be the case that although S has no effect on Y, it may, because of a common cause U, serve as a very good proxy for Y. Knowing about the value of S may be strongly predictive of what will occur with Y potentially for both the treatment and the control arm of a trial. For example, suppose the treatment turned someone's hair blue (S) if and only if the treatment was going to have an effect on the outcome, then the surrogate S would allow us to know the effect of the treatment on the outcome, and thus S would be a good surrogate, even though S itself had no actual effect on the outcome. In this case, S could still be a very useful and informative surrogate. However, the natural indirect effect, $\mathbb{E}(Y_{1S_1} - Y_{1S_0})$, would be 0 because S has no effect on Y. The measure of surrogacy in (8.2) would be 0 even though S might be a highly informative surrogate. Whereas the "proportion explained" measure may be too liberal for mediation in the presence of unmeasured intermediate–outcome confounding (but may be useful for surrogacy), the "indirect effect" measure is too conservative to assess surrogacy (even though it may be of use in assessing mediation). A good surrogate need not mediate the effect of treatment on the outcome if it is otherwise informative of the effect of treatment on the outcome. Conceived of another way, although confounding is important to consider in evaluating the surrogate paradox, when considering measures of surrogacy confounding is not always simply a problem to be gotten rid of, but can provide valuable relations between S and Y which may be helpful in predicting the effect of A on Y from the effect of A on S. The "indirect effects"

approach by attempting to control for or eliminate confounding essentially misses this potentially important source of information concerning surrogacy. The "indirect effect" measure of surrogacy in (8.2) may be of use when most of the effect of A on Y is in fact mediated through S and when the confounding between S and Y is weak, but, in general, the use of the proportion mediated measure eliminates, rather than incorporates, information that may be of importance for assessing the value of a surrogate.

Some of the literature seems to treat the problems of surrogacy and direct/indirect effects as almost interchangeable problems, and certainly the concepts and methods that have traditionally been employed have overlapped considerably for surrogacy and mediation. The goals, however, are quite different. In mediation analysis, we are interested specifically in whether there is an effect of treatment on the outcome that operates through the intermediate. This setting may also be of interest when assessing the properties of a surrogate; but with surrogate outcomes there are settings, as illustrated in Figure 8.1, in which a variable may serve as a very valuable surrogate even if it does not mediate at all the effect of treatment on the outcome. Whereas mediation concerns the pathways by which effects arise, surrogacy concerns principally whether we are able to predict one effect (of treatment on the outcome) by using another (the treatment on the surrogate). Confounding plays a very different role in questions of mediation versus questions of surrogacy. Whereas it is a problem in assessing mediation, it may be an important source of information in surrogacy. The causal estimands best used to capture mediation and surrogacy also differ. The natural indirect effect (Robins and Greenland, 1992; Pearl, 2001) is arguably the most important counterfactual contrast when assessing mediation. However, as argued above, it may, at least in some settings, be of limited interest in assessing surrogacy. A good surrogate need not mediate the effect. While methods developed for mediation and for surrogacy will undoubtedly inform methodology in the other area, the goals and the questions of each setting should be firmly kept in view in deciding on what concepts, definitions, and methods are most relevant.

8.2.4. Meta-Analytic Approach to Surrogacy

Yet another approach to assessing surrogacy is sometimes called the "meta-analytic" approach because it requires either multiple studies or at least multiple subgroups to employ. Burzykowski et al. (2005), for example, propose using either multiple studies or multiple groups defined by covariates within a study to assess surrogacy. Let Φ_j denote the estimate of the effect of treatment A on the outcome Y in the jth study/group. Let ϕ_j denote the estimate of the effect of treatment on the surrogate in the jth study/group. Note that estimation of Φ_j and ϕ_j relies only on the assumption of randomization. To assess surrogacy visually, we could plot Φ_j against ϕ_j. For a good surrogate, we would hope to find the following: (i) a monotonic relationship between ϕ_j and Φ_j, (ii) when $\phi_j = 0$, then $\Phi_j = 0$; and (iii) in a (possibly nonparametric) regression of Φ_j on ϕ_j, we should find not much variability around the regression line. If the relationship between Φ_j and ϕ_j is approximately linear,

we could run a linear regression of Φ_j on ϕ_j and use the R^2 in this regression as a measure of surrogacy:

$$R^2 = \text{Corr}(\Phi_j, \phi_j) \tag{8.3}$$

For this approach to work, however, there must of course be variation in Φ_j and ϕ_j and there must be multiple studies or subgroups in which to estimate effects.

The meta-analytic approach does not give a criterion that ensures the absence of the surrogate paradox, but it can help diagnose and circumvent it. With the meta-analytic approach, if sample sizes are sufficiently large and estimates and modeling assumptions sufficiently precise, an investigator will be able to identify which studies or subgroups are subject to effect reversal (the surrogate paradox) and, for such subgroups, avoid the use of the surrogate. The meta-analytic approach does not give a criterion for avoiding the surrogate paradox but may be of use in detecting groups for which the surrogate is not consistent.

8.2.5. Principal Stratification for Surrogates

As noted in Section 8.1.3, the concepts of principal stratification can also be employed to attempt to evaluate a potential surrogate. As described in Section 8.1.3, this approach builds on the initial insights of Frangakis and Rubin (2002) and has been developed more fully by Follmann (2006), Gilbert and Hudgens (2008), Wolfson and Gilbert (2010), and Huang and Gilbert (2011). Using notions of principal stratification, Gilbert and Hudgens (2008) define as a measure of surrogacy that they call the "causal effect predictiveness surface" for the effect of the treatment evaluated across strata defined by the effect of the exposure on the surrogate [i.e., defined by the principal stratum (S_1, S_0)]. Formally, the causal predictiveness surface is defined by

$$CEP(s_1, s_0) = \mathbb{E}(Y_1 - Y_0 | S_1 = s_1, S_0 = s_0) \tag{8.4}$$

If we knew $CEP(s_1, s_0)$, then we would know for each principal stratum $(S_1 = s_1, S_0 = s_0)$ what the effect of treatment would be. For a binary outcome, the notion of principal surrogacy of Frangakis and Rubin (2002) considered above is simply that $CEP(s_1, s_0) = 0$ for $s_1 = s_0$. For example, if the treatment is binary, the effects $CEP(0,0)$ and $CEP(1,1)$ are simply the "dissociative effects" and $CEP(1,0)$ (or $CEP(0,1)$) is an "associative effect" already considered above. Principal surrogacy requires that the dissociative effects be zero: $CEP(0,0) = CEP(1,1) = 0$—that is, that when the treatment does not change the surrogate, the treatment will not change the outcome. Principal surrogacy is often taken as a criterion for a "good surrogate." The notion is theoretically appealing. Unfortunately, as already indicated above, a principal surrogate does not prevent the surrogate paradox (Chen et al., 2007). A principal surrogate need not be a consistent surrogate. This is also illustrated in the example below.

If we knew the causal predictive surface $CEP(s_1, s_0)$, for each principal stratum $(S_1 = s_1, S_0 = s_0)$, then this could potentially be useful in diagnosing the surrogate paradox. For example, if we knew we had a principal surrogate [i.e., $CEP(0,0) = CEP(1,1) = 0$] and if we also had monotonicity of the effect of A on S so that

the principal stratum $(S_1 = 0, S_0 = 1)$ was empty, then the direction of the average treatment effect of A on Y would be of the same sign as $CEP(1,0)$. However, the criterion of "principal surrogacy" alone (which itself may be difficult to assess) does not ensure a consistent surrogate. Gilbert and Hudgens (2008) modify the definition of a principal surrogate from that of Frangakis and Rubin (2002) to also require what they call one-sided average causal sufficiency that, for a binary outcome, $S_1 > S_0$ implies $P(Y_1 = 1|S_1 = s_1, S_0 = s_0) > P(Y_0 = 1|S_1 = s_1, S_0 = s_0)$. If a surrogate S has the properties of causal necessity and one-sided average causal sufficiency and if we make the monotonicity assumption that the principal stratum $(S_1 = 0, S_0 = 1)$ is empty, then it is straightforward to verify that S cannot exhibit the surrogate paradox. This modified criteria could then be used for diagnosing the surrogate paradox.

Unfortunately, like the "indirect effects" approach, the "principal stratification" approach also requires strong assumptions for identification of the causal predictiveness surface, and often making these strong assumptions is not desirable in the context of randomized trials designed to assess overall treatment effects. Moreover, even when assumptions have been made to identify effect measures, one still does not know which individuals fall into which strata and thus the measures are difficult to use in making decisions prospectively about which individuals should or should not be treated. Notions of surrogacy based on principal stratification are theoretically appealing but difficult to identify in practice. Alternative designs and additional assumptions (Follmann, 2006; Huang and Gilbert, 2011) can help with identification of these effects.

8.2.6. Illustration

To illustrate some of the difficulties with the various approaches considered, especially in the absence of subgroup data required by the meta-analytic approach, consider the following example. Suppose A is randomized, such that $P(S_1 = 0, S_0 = 0) = P(S_1 = 1, S_0 = 1) = P(S_1 = 2, S_0 = 2) = 0.1$, $P(S_1 = 1, S_0 = 0) = 0.5$, and $P(S_1 = 1, S_0 = 2) = 0.2$, and finally suppose $Y = (0.1) * 1(S = 1) + 1(S = 2) + \epsilon_Y$, where ϵ_Y is a standard normal random variable. Here it can be calculated that $\mathbb{E}(S_{a=1} - S_{a=0}) = 0.3$, $\mathbb{E}(Y_{s=2} - Y_{s=1}) = 0.9$, and $\mathbb{E}(Y_{s=1} - Y_{s=0}) = 0.1$, but $\mathbb{E}(Y_{a=1} - Y_{a=0}) = -0.13$ so that the surrogate paradox is present, with a positive effect of A on S, a positive effect of S on Y, and no direct effect of A on Y not through S, but a negative overall effect of A on Y; S is not a good surrogate. If we apply the "proportion explained" approach, we get a proportion explained estimate of 100%, suggesting that S is a perfect surrogate. If we apply the "indirect effects" approach, the natural indirect effect and total effect are both -0.13, suggesting 100% mediation and thus that S is a good surrogate, which it is not. The surrogate does, moreover, satisfy Prentice's criteria. Finally, using principal strata, we would have $CEP(0,0) = CEP(1,1) = CEP(2,2) = 0$, implying that S is a "principal surrogate" and, by this criterion, thus a good surrogate. In this example, the associative effect $CEP(S_1 = 1, S_0 = 0) = 0.1$, which is of the opposite sign of the overall effect of treatment on the outcome and of the other associative effect, $CEP(S_1 = 1,$

$S_0 = 2) = -0.9$. If we were to use as a criterion for a "good surrogate" either (i) the proportion explained, or (ii) the proportion mediated, (iii) principal surrogacy, or (iv) the Prentice criteria, then all of these approaches would suggest that we have a good surrogate when, in fact the sign of the effect of the treatment on the surrogate is the opposite of the sign of the effect of the treatment on the outcome, even though the surrogate has a positive effect on the outcome and even though there is no direct effect of treatment on the outcome not through the surrogate. In this example, failure of transitivity causes the problem. In other examples, unmeasured confounding or the presence of a direct effect may give rise to the surrogate paradox. Again, one of (i) a direct effect, (ii) surrogate-outcome confounding, or (iii) failure of transitivity, must be present for the surrogate paradox to arise.

8.2.7. Surrogate Measures and Consistent Surrogates

In summary, none of the approaches to surrogate outcomes is entirely immune to the surrogate paradox. For the "proportion explained," "indirect effects," and "principal stratification" approaches, none of the standard criterion guarantee a consistent surrogate. The "proportion explained" may be 100% and yet the surrogate paradox may still arise. Likewise the "proportion mediated" using the ratio of the natural indirect effect to the total effect may be 100% and again the surrogate paradox may arise. Finally, a surrogate may be a "principal surrogate" but not a consistent surrogate—the surrogate paradox may still be present. The "meta-analytic" approach does not provide a criterion to avoid the surrogate paradox, but it can be useful in diagnosing it. Likewise in the "indirect effects" approach, if the natural direct and indirect effects were known, these could be useful in diagnosing the surrogate paradox if it were due to the direct and indirect effects being in opposite directions; and in the principal stratification approach, if the causal predictiveness surface were known, this could likewise be useful in diagnosing the surrogate paradox. Unfortunately, however, both the "indirect effects" approach and the "principal stratification" approach suffer from issues of lack of identification; strong assumptions are in general needed to identify these effects, though alternative study designs or sensitivity analysis techniques can sometimes be useful. In light of the aforementioned issues concerning the problems with the surrogate paradox and difficulties in identification, the "meta-analytic" approach may offer the most promise for assessing surrogate outcomes and for making policy and treatment decisions. The approach in principle relies only on randomization assumptions and does not consider effects that require stronger assumptions to identify; moreover, it allows for easier diagnosis of effect reversal manifested in the surrogate paradox. Nonetheless, it is not without its disadvantages because the sample size requirements for effective implementation may be prohibitively large (Gail et al., 2000).

The surrogate paradox is an important problem. If the effect of the treatment on the surrogate is in the opposite direction of the effect of the treatment on the outcome of interest, then policy and treatment decisions may be severely misguided. In the case of ventricular arrhythmia, this very problem resulted in an estimated 50,000 excess deaths (Moore, 1995). We have considered some conditions that ensure against the surrogate paradox, the phenomenon that the effect of

the treatment on the surrogate may be positive, and the surrogate and outcome being strongly positively associated, but the effect of the treatment on the outcome might still be negative. The conditions are not themselves testable but have the fairly intuitive interpretation that for the surrogate paradox to arise, at least one of the following must be present: (i) a direct effect of the treatment on the outcome not through the surrogate, (ii) confounding of the surrogate–outcome relationship, or (iii) a lack of transitivity so that the treatment does not change the surrogate for the same persons for whom the surrogate changes the outcome.

Here we have focused on the task of determining when data concerning the effect of treatment on the surrogate can be used to make decisions about the direction of the effect of the treatment on an outcome—that is, of assessing whether a surrogate is consistent. We have considered the value of a number of different results and approaches to surrogate outcomes in accomplishing this task. Surrogates may, however, be useful in other tasks. For example, we might be interested in determining the extent to which we can predict the outcome once we observe the treatment and surrogate; or the extent to which we could use treatment, surrogate, and outcome data in one population to predict the effect of treatment on outcomes in another population (or the effect of a different treatment in the same population) for which only data on treatment and surrogate are available. Methodology for surrogacy is still developing, and many of these questions still require further methodological development.

8.3. INSTRUMENTAL VARIABLES

8.3.1. Introduction to Instrumental Variables

Throughout most of the discussion in this book we have been articulating assumptions about confounding control which allow an investigator to estimate various types of causal effects. Often these assumptions about confounding control have been quite strong. We have considered a variety of sensitivity analysis techniques to help assess how sensitive one's conclusions are to violations in the assumptions. An alternative way of addressing this issue of unmeasured confounding is to use what is sometimes called an instrumental variable. The classic notion of an instrument is a variable that affects the exposure, but only affects the outcome through the exposure. If such a variable exists, then it can be useful in assessing what the effect of the exposure on the outcome is, even if the exposure–outcome relationship is confounded. The basic intuition behind using an instrument is that if the effects of the instrument itself on the exposure and on the outcome are unconfounded, then one can estimate the effect of the instrument on the exposure and the effect of the instrument on the outcome and combine these to, in some sense, back out the effect of the exposure on the outcome. The critical assumption with an instrumental variable is that the instrument affects the outcome only through the exposure. This assumption is generally referred to as the exclusion restriction. Often it is difficult to clearly establish that this assumption holds, but without this assumption the instrumental

variable estimators described below will be biased. Whereas in Chapters 2–7, when we were considering mediation, we were attempting to assess the extent to which the effect of the exposure on the outcome was through an intermediate or direct, within the instrumental variable context we are assuming that all of the effect of the instrument A on the outcome Y is through the exposure S. We are assuming there is no direct effect of the instrument A on the outcome Y not through the exposure S. Thus instrumental variable analysis is a very distinct approach to that which we have been considering in previous chapters. In previous chapters we made strong assumptions about confounding but allowed for direct and indirect effects and for the estimation of these. The instrumental variable context does not make assumptions about exposure–outcome confounding (it allows such confounding), but it makes an assumption that all of the effect of the instrument on the outcome is mediated by the exposure—that is, that none of it is direct. Which approach is to be taken will depend on the setting, the data available, which assumptions might be plausible, and, most importantly, the question of interest.

8.3.2. Instrumental Variables with a Binary Exposure

In general, the exclusion restriction alone is not sufficient to estimate causal effects, but the exclusion restriction in conjunction with some other assumptions does suffice to estimate causal effects. Consider first the setting of a binary exposure. Suppose we wanted to estimate the effect of the binary exposure (which we will call S here), on some outcome Y, but we thought that the effect of S on Y was strongly confounded. Suppose that there were a variable A that was an instrument for the effect of S on Y—that is, which was such that A affected S but affected Y only through S. Suppose further that conditional on covariates C the effect of A on Y was unconfounded and the effect of A on S was unconfounded. Finally, suppose that the effect of A on S is monotonic in the sense that an increase in A will always increase or leave unchanged the value of S. In counterfactual notation, this monotonicity requires that if $a^* \leq a$, then $S_{a^*} \leq S_a$. Under these assumptions, even if there is unmeasured confounding for the effect of S on Y, it is possible to estimate the effect of S on Y for a certain subpopulation. Specifically, under these assumptions, if, in comparing two levels of the instrument, a and a^*, one takes the ratio of the effect of A on Y and the effect of A on S within strata of C, that is,

$$\frac{\mathbb{E}[Y|a,c] - \mathbb{E}[Y|a^*,c]}{\mathbb{E}[S|a,c] - \mathbb{E}[S|a^*,c]}$$

then this ratio can be interpreted as the effect of exposure S on outcome Y amongst the subpopulation for whom a change in the instrument from a^* to a would suffice to change the exposure S from $S = 0$ to $S = 1$ (Angrist et al., 1996; cf. Tan, 2006). In counterfactual notation, this effect is $\mathbb{E}[Y_{s=1} - Y_{s=0}|S_a < S_{a^*}, c]$ or $\mathbb{E}[Y_{s=1} - Y_{s=0}|S_a = 1, S_{a^*} = 0, c]$; and under the assumptions above, it is equal to $\frac{\mathbb{E}[Y|A=a,c]-\mathbb{E}[Y|A=a^*,c]}{\mathbb{E}[S|A=a,c]-\mathbb{E}[S|A=a^*,c]}$, which can be estimated from the data. The effect $\mathbb{E}[Y_{s=1} - Y_{s=0}|S_a < S_{a^*}, c]$ is sometimes referred to as a "local average treatment effect" or LATE. It is "local" in that it is an effect, not for the entire population but

only for those for whom a change in the instrument from a^* to a would change the exposure from 0 to 1—that is, for whom $S_{a^*} < S_a$. This is very important in the interpretation of these effects because if a different instrument is used, a different effect is obtained. It is possible to use two different instruments, for them to give two different estimates of a causal effect, and for both estimates to be valid. This is because they are estimates for the effect in two different subpopulations.

A classic example of an instrument is that which was discussed in Section 8.1.2, where, within a randomized trial, treatment assignment A may be an instrument for the effect of treatment taken S on the outcome Y. Here both the instrument A and the exposure S are binary. The effects of the instrument A on the treatment taken S and on the outcome Y are both unconfounded by the randomization of treatment assignment. It follows then that provided that there is no effect of treatment assignment on the outcome that is not through the actual treatment taken, and provided that there is no one who would not take treatment if assigned to treatment but who would take treatment if assigned to control, then, as noted in Section 8.1.2, we would have that the ratio estimator, $\frac{\mathbb{E}[Y|A=1]-\mathbb{E}[Y|A=0]}{\mathbb{E}[S|A=1]-\mathbb{E}[S|A=0]}$, would in fact be a consistent estimator for the treatment effect among the compliers, $\mathbb{E}[Y_{s=1} - Y_{s=0}|S_0 = 0, S_1 = 1]$. Treatment assignment in this setting would be an instrument for the effect of treatment taken on the outcome. However, the notions of instrumental variable estimators apply much more generally to settings in which the instrument is not directly randomized, to settings in which the instrument is not binary, and, as we will see below, also to some settings in which the exposure of interest is not binary.

8.3.3. Instrumental Variables with a Continuous Exposure

Suppose now that we have a continuous exposure S. Assume again that we have an instrument A such that the exclusion restriction holds—that is, such that the effect of A on Y is only through S. Suppose further that the effect of A on S and on Y is unconfounded conditional on covariates C (though, as before, the effect of S on Y may be confounded conditional on C). A common approach to instrumental variable analysis with a continuous outcome Y and a continuous exposure S is to use what is often referred to as "two-stage least squares." In this approach, one first regresses the exposure S on the instrument A and the covariates C:

$$\mathbb{E}[S|a, c] = \beta_0 + \beta_1 a + \beta_2' c.$$

Let $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2'$ be the estimates obtained from this regression model. For each individual i, one then uses the regression model and the parameter estimates to obtain a predicted exposure value for each individual

$$\widehat{S}_i = \widehat{\beta}_0 + \widehat{\beta}_1 a_i + \widehat{\beta}_2' c_i$$

where a_i and c_i are the instrument and covariate values for individual i. Finally one regresses the outcome on the predicted exposure $\widehat{S}_i$ and the covariates:

$$\mathbb{E}[Y|\widehat{S}, c] = \gamma_0 + \gamma_1 \widehat{S} + \gamma_2' c$$

The estimate of the coefficient γ_1 is the two-stage least squares estimator of the effect of the exposure S on the outcome Y. It will be a valid estimate of the effect of exposure S on the outcome Y, provided that the exclusion restriction holds, the effect of A and S and on Y are unconfounded conditional on C, and the effect of S on Y is homogeneous (cf. Hernán and Robins, 2006). The estimate of the standard error for the estimate of γ_1 in the second-stage regression cannot be used as an estimate of the standard error for the effect because it does not take into account the first-stage estimation. Below we will give SAS code that will allow for valid estimates of the standard errors as well. The approach is numerically equivalent to using a ratio estimator with a continuous exposure in which one regresses S on A, $\mathbb{E}[S|a,c] = \beta_0 + \beta_1 a + \beta_2' c$, to obtain an estimate of the effect of A on S of $\widehat{\beta}_1$, and one then regresses Y on A, $\mathbb{E}[Y|a,c] = \theta_0 + \theta_1 a + \theta_2' c$, to obtain an estimate of the effect of A on Y of $\widehat{\theta}_1$, and one then uses the ratio $\widehat{\theta}_1/\widehat{\beta}_1$ as the estimate of the effect of S on Y.

This ratio approach is sometimes also used with logistic regression when the outcome is binary. One might regress S on A using linear regression, $\mathbb{E}[S|a,c] = \beta_0 + \beta_1 a + \beta_2' c$, to obtain an estimate of the effect of A on S of $\widehat{\beta}_1$ and then regresses Y on A using logistic regression, $\text{logit}(Y = 1|a,c) = \theta_0 + \theta_1 a + \theta_2' c$, to obtain estimate of the effect of A on Y of $\widehat{\theta}_1$. The ratio θ_1/β_1 is approximately equal of the effect of S on Y on the log-risk ratio scale if the outcome Y is relatively rare (Didelez et al., 2010); that is, under this rare outcome assumption, we have that $\exp(\theta_1)^{1/\beta_1}$ is approximately equal to the effect of S on Y on the risk ratio scale.

8.3.4. SAS Implementation for Instrumental Variable Estimators

Suppose we had a dataset called "mydata" and with a continuous outcome ("y") and continuous exposure ("s"), with covariates ("c1 c2 c3") and we wanted to use an instrument "a." The following SAS code implements the two-stage least squares for a continuous outcome described in Section 8.3.3.

```
proc syslin data=mydata 2sls;
   endogenous s;
   instruments a c1 c2 c3;
   model s = a c1 c2 c3;
   model y = s c1 c2 c3;
run;
```

The same code could be used for the ratio estimator in Section 8.3.2 with a binary exposure and either a binary or continuous outcome; but if any of the covariates "c1 c2 c3" are continuous, then the models for the exposure or the outcome may not fit well or converge. Alternatively to implementing the ratio estimator of Section 8.3.2, one could fit separate models for $\mathbb{E}[Y|a,c]$ and $\mathbb{E}[S|a,c]$ using any SAS regression procedure and then use these models to compute $\frac{\mathbb{E}[Y|a,c]-\mathbb{E}[Y|a^*,c]}{\mathbb{E}[S|a,c]-\mathbb{E}[S|a^*,c]}$ for two values of the instrument, a and a^*, and obtain standard errors by bootstrapping.

8.4. MENDELIAN RANDOMIZATION

8.4.1. Concept of in Mendelian Randomization

There has been increasing interest in and applications of an approach to assessing causal effects which has come to be known as Mendelian randomization (Katan, 1986; Davey Smith and Ebrahim, 2003, 2004). The basic idea of this approach is to use a genetic marker that affects a particular exposure, but is thought to affect an outcome of interest only through the exposure. In other words, one uses the genetic marker as an instrument for the effect of the exposure on the outcome. The effect of the genetic marker on the exposure and on the outcome is then used in order to back out what the effect of the exposure itself on the outcome might be. Under assumptions about instrumental variables described in the previous section, this can be done even if the exposure–outcome relationship itself is confounded. Again, this approach essentially uses the genetic marker as an instrumental variable for the effect of the exposure on the outcome. The approach is appealing because in many settings the confounding between the exposure and the outcome may be thought to be intractable so that causal effects cannot be estimated from observational data using standard regression techniques.

As is hopefully clear from Section 8.3, the Mendelian randomization approach itself makes a number of assumptions. It is not always the case that these assumptions are carefully evaluated. These assumptions about using a genetic marker as an instrument are sometimes stated as follows: (a) The genetic marker is associated with the exposure, (b) the genetic marker is independent of the outcome given the exposure and all confounders (measured and unmeasured) of the exposure–outcome association (i.e., the exclusion restriction), and (c) the genetic marker is independent of factors (measured and unmeasured) that confound the exposure–outcome relationship. As noted above, the second assumption is sometimes referred to as the exclusion restriction; if the genetic marker affects the exposure, then the assumption is also sometimes stated as follows: The genetic marker affects the outcome only through the exposure. In other words, for the assumptions of Mendelian randomization to hold, we need a setting like that of Figure 8.5a or 8.5b. The Mendelian randomization assumptions are violated in a diagram like that of Figure 8.6a or 8.6b because the genetic marker affects the outcome through pathways other than through the exposure.

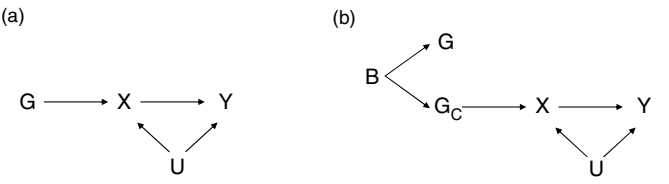

Figure 8.5 Diagrams illustrating a genetic variant (G) which is an instrument for the effect of exposure (X) on outcome (Y) when the exposure-outcome relationship is confounded by unmeasured factors (U).

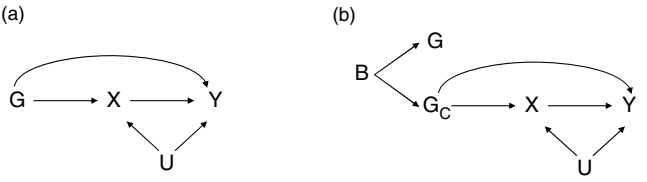

Figure 8.6 Diagrams illustrating violations of the exclusion restriction with the genetic variant (G) not independent of outcome (Y) conditional on exposure (X) and confounders (U) because of a direct effect of G on Y not through X.

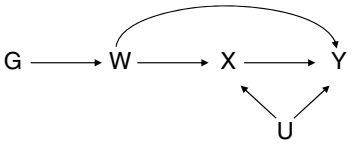

Figure 8.7 Setting in which the variant (G) affects an intermediate (W) on the pathway to exposure (X), violating the exclusion restriction.

If these assumptions do hold, then methods from the instrumental variable literature can be applied to use data on the genetic marker, the exposure and the outcome to estimate what the effect of the exposure on the outcome is. It is not necessary to have the exposure–outcome confounders themselves to be able to do this. If assumption (a) holds but the association between the genetic marker and the exposure is quite weak, then the genetic marker is sometimes referred to as a "weak instrument." Using a weak instrument will often amplify biases due to any violations in the other two Mendelian randomization assumptions.

Although there may be general awareness that these assumptions must hold in a Mendelian randomization analysis, it is often the case that insufficient attention is paid to evaluating these assumptions and their implications. In this section will discuss these assumptions and consider what is entailed by them (VanderWeele et al., 2014a). We will argue that in a number of genetic contexts these assumptions are not plausible. We will also present three ways forward to help address some of these assumptions and their potential violations, which might help strengthen inferences from Mendelian randomization analyses.

8.4.2. Exclusion Restriction Violations: Pathways, Time, Interaction, Measurement Error, Reverse Causation, Sample Selection, and Linkage Disequilibrium

We will begin by highlighting a number of ways in which the exclusion restriction might be violated. First, if the genetic marker is not to affect the outcome except through the exposure then the exposure that is used in the analysis must capture all of the ways that the genetic marker may affect the outcome. Consider a setting such as that in Figure 8.7 where G denotes the genetic marker, X the exposure of interest used in the analysis and Y the outcome. Let W be some variable that is on the pathway from the genetic marker G to the exposure X. In this case, if X is

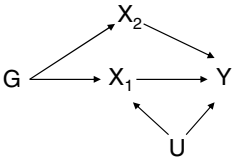

Figure 8.8 Setting in which the exclusion restriction holds when (X_1, X_2) are taken together as the exposure but does not hold for X_1 alone.

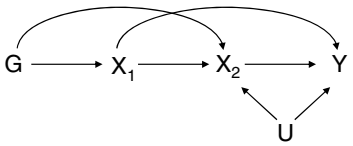

Figure 8.9 Setting in which the exposure is time-varying and in which the exclusion restriction holds when (X_1, X_2) are taken together as the exposure but does not hold for X_1 alone.

used as the exposure in the analysis, the Mendelian randomization approach will be biased. This is because the genetic marker affects the outcome through a pathway not through the exposure X, namely, $G \rightarrow W \rightarrow Y$. For the assumptions of the Mendelian randomization analysis to hold, then the exposure X would essentially have to be the first and only variable on the pathway from G to Y. If there were any variable W on the pathway from G to X that also affected Y, the Mendelian randomization analysis would be biased.

Consider a closely related scenario, namely, the one in Figure 8.8. Suppose that the genetic marker affects the outcome only through a particular exposure or phenotype but that this itself consists of two components (X_1, X_2). If only one of these two components, X_1 say, were used in the Mendelian randomization analysis, then there could be substantial bias in the MR estimates of the effect of the exposure on the outcome because once again the exclusion restriction is violated: The genetic marker affects the outcome via pathways other than the exposure X_1 used in the analysis. The results of the MR analysis using only X_1 in the analysis will be biased for the effects of X_1 on Y and also for the effects of (X_1, X_2) on Y.

A related case in this scenario might be if (X_1, X_2) constitute the exposure of interest at two different times as in Figure 8.9, both of which can affect the outcome Y. If only the measurement at one of the times is used in the analysis, then once again the Mendelian randomization analysis will be biased.

Many of these points have been acknowledged in the past (Davey Smith and Ebrahim, 2003; Thomas and Conti, 2004; Didelez et al., 2010), but it is not clear that these issues are really being acknowledged, evaluated, and discussed when Mendelian randomization analyses are carried out in practice. It is sometimes claimed that, although the Mendelian randomization assumptions may not hold exactly, the bias resulting from them is likely to be small in comparison with the bias from simply examining the associations between the exposure and the outcome (Davey Smith and Ebrahim, 2003). While this claim may hold in some scenarios, it certainly will not hold in all. Violations of the exclusion restriction, such as we have

discussed, can in fact lead to very substantial biases in Mendelian randomization analysis. We illustrate with an example.

It has recently been established that genetic variants on chromosome 15 affect both smoking behavior (e.g., cigarettes per day) and lung cancer. Subsequently there has been interest as to whether the effects of the variants on lung cancer operate entirely through smoking or whether they might also operate through other pathways. As discussed in Chapter 2, using lung cancer case–control data from Massachusetts General Hospital (MGH) to examine whether or not there were pathways not through increasing cigarettes per day gives substantial evidence that such pathways were indeed present (cf. VanderWeele et al., 2012a). The techniques employed in this analysis in Chapter 2 were mediation techniques rather than Mendelian randomization techniques; that is, they did not assume (as Mendelian randomization analysis does) that there were no pathways of the variants on lung cancer except through cigarettes per day. Rather the techniques were designed to evaluate this assumption. What the results of these analyses established was that there were pathways other than through increasing cigarettes per day and that in fact most of the effect of the variants on lung cancer were not due to increasing cigarettes per day. Extensive sensitivity analyses for unmeasured confounding and measurement error confirmed this, and the results were also replicated in three other studies (VanderWeele et al., 2012a).

Note, however, that the mediator that was evaluated was only cigarettes per day, not all aspects of smoking. The results of the analysis is thus still consistent with the possibility that smoking behavior, taken as a whole, mediates the entirety of the association. Indeed it has been suggested (Le Marchand et al., 2008) that a central means by which the variants affect lung cancer is by increasing the amount of nicotine and toxins extracted per cigarette—for example, through deeper inhalation. This then would be a scenario similar to that depicted in Figure 8.8 where perhaps two aspects of the smoking phenotype—for example, cigarettes per day X_1 and depth of inhalation X_2—perhaps mediate all of the effect. What happens then if one were to apply Mendelian randomization using only cigarettes per day in this analysis? Here again, the Mendelian randomization assumptions would be violated with cigarettes per day taken as the exposure because there are pathways from the variants to lung cancer other than through cigarettes per day. In this case, as we will see below, the biases that result are substantial.

Using the same MGH lung cancer data, the association between smoking one pack of cigarettes per day and lung cancer gives an odds ratio of 2.6 (95% CI: 2.3, 3.0), controlling also for age, sex, and college education. If we apply a Mendelian randomization approach to these data, taking cigarettes per day as our exposure, we obtain an odds ratio for the effect of a pack of cigarettes per day on lung cancer of 2180 (VanderWeele et al., 2014a), which is vastly larger than that from examining smoking–lung cancer association directly. Although the smoking–lung cancer association may be confounded, it is arguably not confounded to the degree that would be required for an effect estimate odds ratio of 2180. In this case, the bias from the Mendelian randomization analysis is substantial and again arises from a failure of the exclusion restriction. The MR estimate is so severely exaggerated because the exposure variable used—cigarettes per day—only captures a

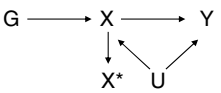

Figure 8.10 Diagram illustrating that when the exposure (X) is measured with error (X^*) the exclusion restriction will not in general hold for the mismeasured exposure X^*.

small part of the effect of genetic variants on the outcome. Moreover, the association between the variant and cigarettes per day is relatively weak, resulting in weak instrument problems, which essentially amplify the bias due to the exclusion restriction violation. The MR estimate for the effect of cigarettes per day on lung cancer effectively divides the variant–lung cancer association by the variant–cigarettes association; and because the latter is so small (since cigarettes per day is only one aspect of smoking by which the variants may affect lung cancer), the final MR estimate is biased dramatically upwards.

This example also suggests another way in which the exclusion restriction can be violated, which is sometimes not discussed when Mendelian randomization is employed in practice. One way to understand the effects of the variants on lung cancer is that the variants essentially amplify the effect of each cigarette smoked, because those with the variant extract more nicotine and toxins per cigarette (Le Marchand et al., 2008). This is then essentially a form of gene–environment interaction, and indeed there is strong statistical evidence for such gene–environment interaction (between the variants and cigarettes per day) on both the additive and multiplicative scales (VanderWeele et al., 2012a). However, gene-environment interaction itself (when both the genetic factor and the environmental factor cause the outcome and interact in their effects) constitutes a violation of the exclusion restriction. This is because under such gene–environment interaction, the genetic variants affect the outcome not simply through the value the exposure takes. Such interaction corresponds to, for example, Figure 8.6a not to Figure 8.5a. Figure 8.5a requires that once we know the exposure X, the genetic marker G gives us no further information about the outcome Y. In the presence of gene–environment interaction, this does not hold, and Figure 8.6a would describe the actual relations. The exclusion restriction is thus again violated and Mendelian randomization analyses are biased. Importantly, here, whether there is gene–environment interaction may depend on how the exposure is defined. If the exposure were defined as all aspects of smoking behavior (or a composite measure of nicotine and toxins extracted) there may be no gene–environment interaction. The genetic variants may affect lung cancer only through the composite of all aspects of smoking behavior. However, when the exposure is defined as cigarettes per day, then gene–environment interaction is present and the exclusion restriction is violated. The use of standard instrumental variable estimators would in general be violated in the presence of gene–environment interaction.

This brings us to yet another potential violation of the exclusion restriction. Consider the diagram in Figure 8.10 in which the exclusion restriction holds for the true underlying exposure X, but the investigator only has access to X^*, a measure of the exposure subject to measurement error. Here again, the exclusion restriction will in general be violated. Some specific exceptions can arise such as when the exposure

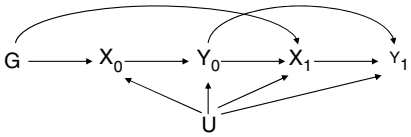

Figure 8.11 Setting in which prior outcome (X_0) can affect subsequent exposure (X_1) so that the variant (G) is not independent of final outcome (Y_1) conditional on exposures (X_0, X_1) and confounders (U), violating MR assumptions, due to reverse causation.

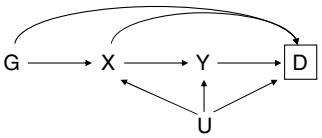

Figure 8.12 Setting in which variant (G) is associated with outcome (Y) conditional on exposure (X) and confounders (U), violating the MR assumptions, because of using data from a case-control study of another outcome (D).

X is continuous and subject to classical nondifferential measurement (Wooldridge, 2002; Davey Smith and Ebrahim, 2003; Lawlor et al., 2008; Pierce and Vander-Weele, 2012). However, in other cases, such as with a dichotomous exposure or differential measurement error, Mendelian randomization analyses will be biased. It is moreover possible that the disease itself distorts the measurement of the exposure with an arrow from Y to X^* in Figure 8.10, a form of differential measurement error. This would introduce yet further bias. And this also brings us to yet another possible violation of the exclusion restriction.

The exclusion restriction might also be violated via reverse causation or feedback. Consider the diagram in Figure 8.11 in which the genetic variant has no effect on the outcome Y_1 that does not go through the exposure (X_0, X_1), but the exposure X_1 is affected by the prior outcome Y_0. If in a Mendelian randomization analysis (X_0, X_1) is used as the exposure with the outcome Y_1, then our Mendelian randomization assumptions will be violated. In this case the genetic marker G is not independent of the outcome Y_1 conditional on the exposure (X_0, X_1) and the confounders U, since conditional on (X_0, X_1, U) the variant G is still associated with Y_1 through the path $G \rightarrow X_1 \leftarrow Y_0 \rightarrow Y_1$ having conditioned on X_1 (Pearl, 2009). We would thus get biased estimates when using instrumental variable techniques to estimate the effect of (X_0, X_1) on Y_1. We note that the situation described in Figure 8.11 is especially concerning for retrospective case–control studies, where data on the exposure is collected after diagnosis of the outcome.

Violations of the exclusion restriction can also arise due to the selection of the sample. For example, if interest lies in the effect of X on Y but the sample was originally collected as part of a case–control study of another disease D and both Y and X are associated with D as in Figure 8.12, then although there is no effect of G on Y that does not go not through X, G is still associated with Y conditional on exposure X and confounders U because of the conditioning on the status of D. Instrumental variable estimates of the effect of X on Y will again be biased. Existing case–control studies are increasingly used for Mendelian randomization analyses of

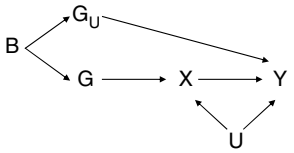

Figure 8.13 Setting in which variant (G) is associated with outcome (Y) conditional on exposure (X) and confounders (U), violating MR assumptions, because of linkage disequilibrium with another variant (G_U) that affects the outcome.

traits other than the diseases they were originally designed to study, so this scenario is practically relevant. Methods for handling Mendelian randomization analyses with such secondary phenotypes are becoming available (Tchetgen Tchetgen et al., 2013), but if the standard methods are used ignoring the design, then this will yield biased estimates.

Finally, consider the diagram in Figure 8.13 in which another genetic marker G_U which also affects the outcome Y is in linkage disequilibrium with the genetic marker G used in the analysis. This again constitutes a violation of the exclusion restriction because G is not independent of Y conditional on X and U. That such linkage disequilibrium violates the assumptions of a Mendelian randomization analysis has likewise been noted before (Davey Smith and Ebrahim, 2003; Didelez and Sheehan, 2007). However, it is not clear that the full implications of this are usually considered when Mendelian randomization is employed in practice. In this case, we could restore the Mendelian randomization assumptions if we could control for the other genetic marker G_U. However, to avoid such violations, like that in Figure 8.13, entirely, it would essentially be required that there be nothing on the same chromosome as the genetic marker used in the analysis that also affects the outcome or that control had been made for all such variables. This is a strong assumption and potential violations are numerous. Recent large-scale fine mapping studies of autoimmune disease, metabolic traits, cancers, and anthropometric traits have all identified risk loci containing multiple correlated alleles that remain statistically significantly associated with these outcomes after mutual adjustment (Evre et al., 2012; Morris et al., 2012; Scott et al., 2012; Lango et al., 2010; Speliotes et al., 2010). Moreover, there are many loci that harbor distinct, linked alleles associated with a wide range of phenotypes (e.g., the HLA region and the ABO locus, Schunkkert et al, 2011; Hindoff et al., 2009). As a concrete example, Garcia-Closas et al. (2013) report evidence for a variant for BMI in the FTO gene that is strongly correlated with a different variant for breast cancer. If this variant for BMI were used in a Mendelian randomization analysis for the effect of BMI on breast cancer, the Mendelian randomization assumptions would be violated.

Similar violations of the exclusion restriction occur in the presence of population stratification. Although this issue can be addressed using the standard methods of genomic control employed in GWAS studies or by using family-based studies (Davey Smith and Ebrahim, 2003, 2004; Lawlor et al., 2008), it is not clear that such methods for genomic control are being employed in Mendelian randomization studies. Failure to do so would result in a causal diagram similar to Figure 8.13, once again biasing Mendelian randomization results.

Although these points have been acknowledged in the literature, they are often not adequately considered when Mendelian randomization is employed in practice and violations can potentially lead to substantial biases. When reporting results of a Mendelian randomization analysis, investigators should clearly state the assumptions being made in the context of their specific application and should discuss the plausibility of the assumptions being made. Investigators should consider and comment on (i) whether the exposure used in the analysis completely captures the phenotype that may mediate the association between the variant and the outcome, (ii) whether the exposure is time-varying, (iii) whether there may be gene–environment interaction, (iv) whether the exposure may be measured with error, (v) whether there may be reverse causation, and (vi) whether there may be other genetic markers on the same chromosome that affect the outcome and are correlated with the marker used in analysis. If a good argument can be made against these possibilities, then investigators can more reliably assume that the exclusion restriction holds. These are, of course, all very strong assumptions. Is there another way forward?

8.4.3. Approach 1: A Return to Katan

The origin of Mendelian randomization analysis is often traced to a letter by Katan (1986). Katan did not propose using instrumental variable statistical methods to estimate the effect of the exposure on the outcome. Rather he simply proposed examining the association between the genetic variants and the outcome to test for an effect of the exposure on the outcome. He proposed not using the exposure data at all!

At first this may be surprising. It would seem that we should be better off making use of the exposure data. However, as discussed in the previous section, one of the challenges in ensuring that the exclusion restriction holds is defining and measuring the exposure in such a way to avoid incomplete phenotype information (time-varying or otherwise), measurement error, and gene–environment interaction. This will in general not be easy to do. However, if all we are interested in is testing for an effect of the exposure on the outcome (rather than estimating the magnitude of this effect), then no exposure information is needed at all. We need data only on the genetic variant and the outcome. We still need to make the exclusion restriction assumption—that the variant affects the outcome only through the exposure—but we do not necessarily need to fully specify what that exposure is or collect data on the exposure thus defined. To see this, consider again the diagrams in Figure 8.5a and 8.5b. If there were no effect of the exposure on the outcome (i.e., no arrow from X to Y), then the genetic variant G and the outcome Y should be unassociated. If there is no arrow from G directly to Y (i.e., if the exclusion restriction holds) and G and Y are correlated, then there must be an effect of the exposure X on the outcome Y. And we can test for this without using data on X. This was the approach initially proposed by Katan.

Of course it seems more desirable to obtain estimates of the effect of the exposure on the outcome rather than just to test for the presence of an effect, and that

is what the instrumental variable methods used in Mendelian randomization analysis purportedly do. However, doing so requires much stronger assumptions. This is essentially because for valid tests we need the Mendelian randomization assumptions to hold under some definition of the exposure. For valid estimates we need the Mendelian randomization assumptions to hold for the measurement of the exposure for which data are available. The Mendelian randomization assumptions are much less likely to hold for a single particular measurement of the exposure than they are for a complete characterization of all aspects of the exposure and exposure history. If we had data on the entire history of all aspects of the exposure, then we could potentially again proceed to obtain estimates using causal assumptions no stronger than those involved in the tests, but in the absence of such complete exposure data, the tests will rely on weaker assumptions than the estimates.

Let us revisit all of the potential violations of the exclusion restriction in Figures 8.8 to 8.13. Suppose we were to follow the original approach proposed by Katan, rather than using the more contemporary approach of employing instrumental variables methods to estimate effects. Consider Figure 8.8 and suppose we had data only on X_1. As discussed above, our Mendelian randomization estimates would be biased. However, if we were to simply examine whether there was association between G and Y, we would still have a valid test of whether there was an effect of phenotype (X_1, X_2) on the outcome Y. If there were no effect, then we should observe no association between G and Y. Likewise in Figure 8.9, with time-varying exposures, we could test whether there were an effect of the exposure (X_1, X_2) on the outcome Y by testing whether G and Y were associated. Once again, if there were no effect, then we should observe no association between G and Y. The difficulties with only observing part of the relevant phenotype vanish. We do not need to make use of the exposure/phenotype information at all, and we are protected against the biases described in the previous section because we are not estimating, only testing.

Similarly, in the case of gene–environment interaction, if there were some definition of the phenotype such that there were no gene–environment interaction (i.e., once the phenotype is known, the genetic marker gives no further information on the outcome) so that Figure 8.5a or 8.5b holds for that definition of phenotype rather than Figure 8.6a or 8.6b, then we could once again test for an effect of phenotype on outcome simply by looking at the association between the genetic marker and the outcome. If there were no effect, we should find no association. Note that here we do not even need to know under what definition of the phenotype there would be no gene–environment interaction; we only need to assume that there is one and that, with respect to this definition, the exclusion restriction holds.

Similar conclusions hold in the case of measurement error. If we use the original approach suggested by Katan, measurement error does not cause a problem. In Figure 8.10, if there is no effect of the true exposure X on Y, we should find no association between G and Y. We can do testing without data on X. We do not need to worry about measurement error in X. And similarly again with Figure 8.11, if X_1 and Y_1 were associated only because of reverse causation—that is, because Y_0 affects both X_1 and Y_1 as in Figure 8.11 (rather than because X affects Y)—then in

Figure 8.11 we would find no association between G and Y_1 if there were no arrow from (X_0, X_1) to (Y_0, Y_1). The approach of Katan is again applicable.

The approach is thus robust to many of the sources of bias and exclusion restriction violations considered in the previous sub-section. This is because we do not need data on the exposure to proceed with testing. The price of this robustness is that we only get tests, not estimates. The approach of just using the genetic marker and the outcome to do testing for an effect of the exposure on the outcome is, however, of course, not robust to all forms of exclusion restriction violations. In Figures 8.6 or 8.7, the exclusion restriction would be violated for the exposure X and both our tests (using the approach of Katan) and estimates, using instrumental variable methods, would be biased. Likewise in Figure 8.12, in which there is selection bias, or in Figure 8.13, in which another variant that affects the outcome that is in linkage disequilibrium with the variant used in the analysis, the assumptions are violated and both tests and estimates will be biased. However, in a number of scenarios, such as those in Figures 8.8 to 8.11, the tests are protected whereas the estimates are not. Perhaps a return to the approach originally proposed by Katan should be considered.

An interesting analogy can also be drawn in comparing Mendelian randomization with randomized trials in the presence of noncompliance. In trials with non-compliance, we can still get unbiased estimates of the intent-to-treat (ITT) effect by simply comparing outcomes according to treatment assignment—this is the effect of treatment assignment. However, if we want the effect of actually taking treatment, then we must take noncompliance into account. This is sometimes done using instrumental variable methods with treatment assignment used as an instrument for compliance/treatment taken as described in Sections 8.1 and 8.3. Doing so requires making the exclusion restriction assumption (no effect of assignment on the outcome not through measured treatment taken). We may be worried that such an assumption is violated (due to mismeasured compliance, partial compliance, or psychological effects of treatment assignment) and so standard practice in randomized trials is to take the ITT effect as the primary estimate of the study and to make an instrumental variable analysis (or any other analysis using compliance data) as a secondary analysis. This ensures the greatest validity for the primary analysis. The equivalent to the ITT analysis in Mendelian randomization is to simply look at the G–Y association, the original approach of Katan. Current Mendelian randomization practice is to report instrumental variable estimates as primary, but perhaps this should be reversed, with the approach of Katan presented as primary and the instrumental variable estimate as secondary, thereby paralleling practices in randomized trials.

The other thing that the approach of Katan makes clear is the strength of the assumptions. We are using a test of association between the genetic marker and the outcome to draw a conclusion about the effect of the exposure on the outcome. The assumptions being made, those outlined above, are doing the rest of the work. These are strong assumptions. The approach of Katan, just testing for association between the genetic marker and the outcome, makes this clear. It should also be noted that the instrument variable methods themselves could in theory be used simply for the purposes of testing. Testing whether the instrumental variable estimate

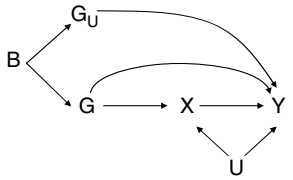

Figure 8.14 Setting illustrating violations of the exclusions restriction and the confounding assumptions for Mendelian randomization.

is different from zero constitutes a valid test of the null of no association between the genetic variant and the outcome, and thus under the MR assumptions a valid test of the null of no effect of the exposure on the outcome (Didelez and Sheehan, 2007). Such a test makes no stronger causal assumptions than the approach of Katan. It is when the instrument variable techniques are used to obtain estimates, rather than just tests, that the stronger assumptions (concerning the validity of the exclusion restriction for the actual exposure measured) are required. However, if the instrumental variable methods are just being used for testing, it is not clear why they would be preferred to the approach advocated by Katan. In using instrumental variable methods, it can seem that the sophistication of the analysis is somehow resulting in the conclusion, whereas it is, in fact, again, relying on the assumptions, and these are even stronger for estimation than for testing.

8.4.4. Approach 2: Sensitivity Analysis for the Exclusion Restriction and Other Violations

Although the approach of Katan can be helpful in addressing violations of the Mendelian randomization assumptions in Figures 8.8 to 8.11, some of the other violations, such as those in Figures 8.6, 8.7, 8.12, and 8.13, are still subject to bias. Another approach to help address violations of the Mendelian randomization assumptions is to use either the approach of Katan or the use of instrumental variable methods in conjunction with sensitivity analysis to assess the extent to which different exclusion restriction violations or violations of assumptions due to linkage disequilibrium with another variant affecting the outcome (or population stratification) would bias results and to try to correct for such bias. The techniques can be adapted in a straightforward manner from existing sensitivity analysis techniques in the literature. Here we present sensitivity analysis techniques for the approach of Katan; in the Appendix we discuss how these can also be applied also to instrumental variable estimation.

Consider Figure 8.14, in which we have both a violation of the exclusion restriction (an arrow from G to Y not through X) and linkage disequilibrium (another genetic marker G_U that was in linkage disequilibrium with G and affected the outcome Y). Suppose first that the outcome Y were dichotomous. Suppose we are comparing two levels of the genetic marker G. For simplicity we denote these by $G = 0$ and $G = 1$, respectively. Suppose then that we estimated a risk ratio between G and Y possibly conditional on measured baseline covariates (which could be

approximated by an odds ratio using logistic regression if the outcome is rare). Suppose that we assume G_U is binary and does not interact with G on the risk ratio scale in its effects on Y and that G_U increases the risk of Y by a factor of γ on the risk ratio scale. Suppose also that although the exclusion restriction is violated (i.e., there is a path from G to Y not through X), the effects of G and X on Y do not themselves interact on the risk ratio scale, and also suppose that the effect of the direct path from X to Y is to increase the risk of Y by a factor of λ on the risk ratio scale. Under these assumptions, we can obtain a "corrected" estimate and confidence interval for the association between G and Y (i.e., what we would have obtained had we been able to control for G_U and had we been able to isolate the $G \rightarrow X \rightarrow Y$ path) by dividing our observed estimate and confidence interval by $\lambda\{1 + (\gamma - 1)\pi_1\}/\{1 + (\gamma - 1)\pi_0\}$, where π_1 and π_0 are the prevalences of G_U amongst those with $G = 1$ and those with $G = 0$ respectively (Conley et al., 2012; VanderWeele and Arah, 2011; VanderWeele et al., 2014a). We could specify values of γ, λ, π_1 and π_0 based on prior studies or we could consider a range of these values and vary them in a sensitivity analysis. Doing so would allow one to assess the degree of confounding by another genetic variant and the strength of the exclusion restriction violations that would be required to explain away the association between the genetic variant G and the outcome Y that is used to test for an effect of the exposure on the outcome. Values of the sensitivity analysis parameters for which the null was still not rejected would be those for which there was still evidence of a causal effect of exposure X on outcome Y, even though the assumptions themselves were violated.

Likewise for a continuous outcome Y, suppose we estimated the effect of G on Y on the difference scale, possibly conditional on measured baseline covariates. Suppose we assume that G_U is binary and does not interact with G on the additive scale in its effects on Y and that G_U increases Y on average by a difference of γ. Suppose also that although the exclusion restriction is violated, the effects of G and X on Y do not themselves interact on the difference scale and that the effect of the direct path from X to Y is to increase the average value of Y by a difference of λ. Under these assumptions, we can obtain a "corrected" estimate and confidence interval (i.e., what we would have obtained had we been able to control for G_U and had we been able to isolate the $G \rightarrow X \rightarrow Y$ path so that the exclusion restriction was not violated) by subtracting from our observed estimate and confidence interval the factor $\gamma\,\delta + \lambda$, where δ is the difference in the prevalences of G_U amongst those with $G = 1$ and those with $G = 0$, respectively (cf. Conley et al., 2012; VanderWeele and Arah, 2011; VanderWeele et al., 2014). If G_U is specified as continuous rather than binary, then γ can be modified to the effect of a one unit increase in G_U on Y, and δ can be modified to the difference in means of G_U amongst those with $G = 1$ versus $G = 0$. We could again specify values of γ, δ, and λ based on prior studies or we could consider a range of these values and vary them in a sensitivity analysis. Again, doing so would allow one to assess the degree of confounding by another genetic variant and the strength of the exclusion restriction violations that would be required to explain away the association between the genetic variant G and the outcome Y that is used to test for an effect of the exposure on the outcome. Values of the sensitivity analysis parameters for which the null was still not rejected would be

those for which there was still evidence of a causal effect of exposure X on outcome Y, even though the assumptions themselves were violated.

While the approach of Katan is subject to bias by linkage disequilibrium with another variant that affects the outcome and to some of the exclusion restriction violations, the extent of this bias can be assessed through sensitivity analysis. The sensitivity analysis technique discussed in this section could likewise be applied to instrumental variable analyses of Section 8.3, outside of the genetics context.

8.4.5. Approach 3: Negative Results May Be More Plausible

The assumptions required for Mendelian randomization are strong, whether for estimation or for testing. The number of applications of Mendelian randomization has been quickly expanding. More attention needs to be paid to the assumptions—to assessing whether incomplete exposure information, time-varying exposures, gene–environment interaction, measurement error, reverse causation, or linkage disequilibrium might bias the analysis. Employing the approach of Katan for testing, rather than estimation, circumvents some of these biases but not others. It is in general difficult to establish the exclusion restriction that is required in these analyses with much certainty. Although empirical tests exist to falsify these assumptions (Wooldridge, 2002; Glymour et al. 2012), the assumptions cannot be fully empirically verified.

In light of the strength of the assumptions being made, it may turn out that Mendelian randomization analysis is in fact more important for establishing negative results. If the approach of Katan is employed by simply examining the association between the genetic variant and the outcome and there appears to be no association, and if it has already been established that the variant affects the exposure of interest, then this may provide evidence that there is in fact no effect of the exposure on the outcome. Of course, relatively large sample sizes would be needed to establish a null association with confidence and to reliably establish conclusions that it would be necessary to use genetic variants with strong associations with the exposure to avoid weak instrument problems (Davey Smith and Ebrahim, 2004). However, research consortia are making it more and more possible to achieve such large sample sizes. Negative results, like positive results, are likewise potentially subject to biases arising from violations in the assumptions. However, with positive results, if there is in fact no effect of the exposure on the outcome, any bias resulting from a deviation from the null will lead to the wrong conclusion. With negative results, if the estimate is close to zero with a very narrow confidence interval, although this could still be subject to bias, the biases would have to all align perfect to move the effect estimate to zero when there is in fact a true effect. There is the whole range of nonzero values for false positives to take; there is only one zero value that a false negative may take.

A null association, with narrow confidence interval, between a genetic variant and an outcome may thus arguably provide more robust evidence of the negative conclusion of no effect or that if there is an effect it is very very small. Such null results can also address important clinical and etiological questions. One recent

example is the finding that CRP-associated variants are not associated with coronary heart disease, suggesting that CRP is not causally associated with heart disease (Elliott et al., 2009). Another example of this are results that indicate that BMI-associated variants are associated with circulating vitamin D levels while variants associated with vitamin D levels are not associated with BMI, suggesting that BMI causally affects with vitamin D levels but not vice versa (Vimaleswaran et al., 2013). In light the strength of the assumptions and the asymmetry between false negative and false positives, Mendelian randomization may in the end prove most valuable in establishing that certain exposures do not affect outcomes, rather than that they do.

8.5. DISCUSSION

In this chapter we have discussed a number of topics concerning intermediates that are related to, but also quite distinct from, mediation. We described the principal stratification framework and discussed that although it is useful for assessing causal effects in the presence of outcomes missing due to death, for assessing post-infection outcomes, and for analyzing causal effects in the presence of noncompliance, it is not particularly helpful in assessing mediation. The principal stratification framework does not make reference to potential outcomes indexed by the mediator; and thus it is difficult to capture, in this framework, effects of the mediator or to analyze mediation. For mediation, it is better to turn to the concepts of natural direct and indirect effects as discussed in Chapters 2–7. In this chapter we also discussed issues of surrogacy. We noted that although we might often expect a good surrogate to also be a mediator, it is in fact possible to have a good surrogate that does not mediate the effect of treatment on the outcome at all. Notions of surrogacy and mediation, although related, are distinct. We also discussed criteria to ensure that the use of a surrogate is at least able to identify the correct direction of the effect of a treatment on the primary outcome of interest. We further discussed the use of instrumental variables techniques and how such techniques assume the absence of direct effects of the exposure on the outcome not through the intermediate, rather than assessing an assumption like this. If an investigator is willing to make such an "exclusion restriction" assumption, then, by using the instrument, it is often possible to identify the effect of the exposure of interest on the outcome even if this relationship is confounded by unmeasured common causes of these two variables. Finally, we discussed how this instrumental variable approach can be applied in the context of genetic studies by using a genetic instrument. Although the approach holds some promise, it is also subject to a number of limitations. The assumptions required for such a "Mendelian randomization" analysis are often much stronger than investigators realize. We unpacked in more detail all that these assumptions entail in the genetic context and have discussed some approaches to using Mendelian randomization ideas even in the presence of violations of the assumptions. As with instrumental variables generally, so also in the Mendelian randomization context, the approach is quite distinct from mediation. Mendelian randomization assumes away, rather than assesses, direct effects, but

if this assumption is reasonable it can be used to deal with issues of otherwise intractable unmeasured confounding.

When direct and indirect effects are truly in view, the methods of Chapters 2–7 can shed light on questions of mediation. However, as we have seen in this chapter, there are other contexts and other questions for which intermediates can be of importance. It is essential that investigators keep in mind the substantive question that is being addressed and appropriately select concepts and methods to handle the question at hand.

PART TWO

Interaction Analysis

9

An Introduction to Interaction Analysis

In this chapter and for the next several we turn to our second major topic of causal explanation: interaction analysis. Whereas mediation analysis essentially tries to assess the pathways by which an effect comes about, interaction analysis is concerned primarily for whom the effect occurs and why it is of a particular magnitude. In this chapter we will present introductory concepts of interaction analysis. We will touch on questions of causation, but much of the material presented in this chapter will be an overview of a number of statistical aspects of interaction. Subsequent chapters will have a greater focus on aspects of the analysis of interaction relevant for causation and explanation.

In this chapter we will discuss interaction on both additive and multiplicative scales using risks, and we discuss their relation to statistical models (e.g., linear, log-linear, and logistic models). We discuss and evaluate arguments that have been made for using additive or multiplicative scales to assess interaction. We describe inferential procedures for interaction when logistic models are fit to data but when additive and not just multiplicative measures of interaction are desired. We discuss issues of confounding for interaction analyses and how whether control has been made for only one or both of two exposures affects whether interaction estimates can be interpreted as causal interaction between the two exposures or only as effect heterogeneity. We further discuss approaches to presenting interaction analyses and approaches to interaction for continuous exposures and outcomes and for time-to-event outcomes. Finally, we discuss methods for identifying subgroups for which to target treatment; qualitative interactions; and methods for attributing effects to interaction.

9.1. MEASURES OF INTERACTION AND SCALE OF INTERACTION

It is not uncommon for the effect of one exposure on an outcome to depend in some way on the presence or absence of another exposure. When this is the case,

Table 9-1. RISK OF LUNG CANCER BY SMOKING AND ASBESTOS STATUS

	No Asbestos	**Asbestos**
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

we say that there is interaction between the two exposures. Recent years have seen increasing interest in interaction between genetic and environmental exposures, but interaction can also occur between two (or more) environmental exposures, or two different genetic exposures, or with various behavioral exposures. The processes giving rise to various outcomes is often inherently quite complex. Interaction between exposures is one manifestation of this complexity.

As a motivating example, consider data presented in Hilt et al. (1986) concerning the effect of smoking on lung cancer and how this varied by previous exposure to asbestos. The risk of lung cancer comparing smokers and non-smokers varied by asbestos exposure as presented in Table 9.1. It seems as though lung cancer risk is much higher when both smoking and asbestos exposure are present together. This is an example of what we might call an interaction.

Let Y denote a binary outcome. Let G and E denote two binary exposures of interest. These might be a genetic factor and an environmental factor, respectively, but our discussion will not be restricted to gene–environment interaction and G and E could represent any two factors; later in the chapter we will also discuss interaction when the factors are not binary. Let $p_{ge} = P(Y = 1|G = g, E = e)$ be the probability of the outcome when G is value g and E is value e. A natural way to assess interaction is to measure the extent to which the effect of the two factors together exceeds the effect of each considered individually. This could be measured by

$$(p_{11} - p_{00}) - [(p_{10} - p_{00}) + (p_{01} - p_{00})] \tag{9.1}$$

Here $(p_{11} - p_{00})$ would be interpreted as the effect of both factors together compared to the reference category of both factors absent. The expressions $(p_{10} - p_{00})$ and $(p_{01} - p_{00})$ would be the effects of the first factor alone or the second factor alone, respectively. We would then consider the contrast between the effects of both factors together versus the sum of each considered separately. If this difference were nonzero, we might say that there was interaction on the difference scale. For now, we will assume that the probabilities of the outcome under different exposure combinations correspond to the actual effects of the exposures on the outcome; we will consider issues of confounding and covariate adjustment in interaction analyses further below.

The measure in (9.1) is sometimes referred to as a measure of interaction on the additive scale. The measure in (9.1) can be rewritten as

$$p_{11} - p_{10} - p_{01} + p_{00} \tag{9.2}$$

If $p_{11} - p_{10} - p_{01} + p_{00} > 0$, the interaction is sometimes said to be positive or "super-additive." If $p_{11} - p_{10} - p_{01} + p_{00} < 0$, the interaction is said to be negative or "sub-additive."

For the data in Table 9.1, we have:

$$p_{11} - p_{10} - p_{01} + p_{00} = 0.0450 - 0.0095 - 0.00670 + 0.0011 = 0.0299$$

We would have evidence here of positive or "super-additive" interaction.

Sometimes, instead of using risk differences to measure effects, one might use risk ratios or odds ratios. For example, we could define the risk ratio effect measures as

$$RR_{10} = p_{10}/p_{00}$$
$$RR_{01} = p_{01}/p_{00}$$
$$RR_{11} = p_{11}/p_{00}$$

A measure of interaction on the multiplicative scale for risk ratios could then be taken as

$$\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}p_{00}}{p_{10}p_{01}} \tag{9.3}$$

This quantity measures the extent to which, on the risk ratio scale, the effect of both exposures together exceeds the product of the effects of the two exposures considered separately. If $RR_{11}/(RR_{10}RR_{01}) > 1$, the multiplicative interaction is said to be positive. If $RR_{11}/(RR_{10}RR_{01}) < 1$, the multiplicative interaction is said to be negative. Note that we compare the measure $RR_{11}/(RR_{10}RR_{01})$ to 1 rather than to 0 here since $RR_{11}/(RR_{10}RR_{01})$ is a ratio. If the ratio is 1, then the effect of both exposures together is equal to the product of the effects of the two exposures considered separately, that is, there is no interaction on the multiplicative scale for risk ratios. This measure of multiplicative interaction can also be rewritten as $\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}}{p_{01}}/\frac{p_{10}}{p_{00}}$, that is, as the ratio of (i) the relative risk for G when $E = 1$ versus (ii) the relative risk for G when $E = 0$. Likewise, it can be written as $\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}}{p_{10}}/\frac{p_{01}}{p_{00}}$, that is, as the ratio of (i) the relative risk for E when $G = 1$ versus (ii) the relative risk for E when $G = 0$.

Using the data in Table 9.1, we have that the measure of multiplicative interaction is given by

$$\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{(0.0450/0.0011)}{\{(0.0095/0.0011) \times (0.0067/0.0011)\}} = \frac{40.9}{8.6 \times 6.1} = 0.78$$

We would have evidence here of negative multiplicative interaction.

This example also demonstrates that whether an interaction is positive or negative may depend on the scale. We may have a positive interaction on the additive scale but a negative interaction on a multiplicative scale. Said another way, the effect of both exposures together on the risk difference scale may exceed the sum of the effects on the risk difference scale of each considered separately, while it also being

Table 9-2. RISK OF OUTCOME BY CROSS-CLASSIFIED EXPOSURE STATUS

	$E = 0$	$E = 1$
$G = 0$	0.02	0.05
$G = 1$	0.07	0.10

Table 9-3. RISK OF OUTCOME BY CROSS-CLASSIFIED EXPOSURE STATUS

	$E = 0$	$E = 1$
$G = 0$	0.02	0.05
$G = 1$	0.04	0.10

the case that the risk ratio for both exposures together is less than the product of the effects of the two exposures considered separately.

Likewise, interaction may be present on one scale but absent on another. Consider the data in Table 9.2. Here there is no additive interaction since $p_{11} - p_{10} - p_{01} + p_{00} = 0.10 - 0.07 - 0.05 + 0.02 = 0$ but there is a negative multiplicative interaction since $RR_{11}/(RR_{10}RR_{01}) = (0.10/0.02)/\{(0.07/0.02)(0.05/0.02)\} = 5/(3.5 \times 2.5) = 0.57 < 1$. Likewise in other settings we might have additive interaction but no multiplicative interaction. Consider the data in Table 9.3. Here the additive interaction is positive since $p_{11} - p_{10} - p_{01} + p_{00} = 0.10 - 0.04 - 0.05 + 0.02 = 0.03 > 0$, but there is no multiplicative interaction since $RR_{11}/(RR_{10}RR_{01}) = (0.10/0.02)/\{(0.04/0.02)(0.05/0.02)\} = 5/(2 \times 2.5) = 1$. In fact it can be shown (cf. Greenland et al., 2008) that if both of the two exposures have an effect on the outcome, then the absence of interaction on the additive scale implies the presence of multiplicative interaction for relative risks; likewise, the absence of multiplicative interaction for relative risks implies the presence of additive interaction. In other words, if both of the two exposures have an effect on the outcome, then there must be interaction on some scale. This raises the question of why interaction is of interest and which scale is to be preferred. In a subsequent section, we will turn to the arguments for and interpretation of additive versus multiplicative interaction. In general, however, either the presence or absence of additive or multiplicative interaction may be of interest, and so it is good practice to evaluate both additive and multiplicative interaction.

One reason why additive interaction is important to assess (rather than only relying on multiplicative interaction measures) is that it is the more relevant public health measure (Blot and Day, 1979; Saracci, 1980; Rothman et al., 1980; Greenland et al., 2008). Consider again the outcome probabilities in Table 9.3. Suppose that the outcome probabilities represent the probability of a disease being cured for a drug (E) stratified by genotype status (G). The effect of E on the risk difference scale amongst those with $G = 0$ is $0.05 - 0.02 = 0.03$, while the effect of E amongst those with $G = 1$ is $0.10 - 0.04 = 0.06$. If we had only 100 doses of the drug and

we had to decide which group to treat, we could cure three additional persons if we used all of the drug supply amongst those with $G = 0$, but we could cure six additional persons if we used all of the drug supply amongst those with $G = 1$. All other things being equal, we would clearly want to give the drug supply to those with $G = 1$. The additive interaction measure, $p_{11} - p_{10} - p_{01} + p_{00} = 0.03 > 0$, allows us to see this. The multiplicative interaction measure, $RR_{11}/(RR_{10}RR_{01}) = 1$, does not.

In fact, the multiplicative scale can indicate the wrong subgroup to treat. Suppose in Table 9.3 we replace the final probability of cure 0.10 with 0.09. Then the effect on the difference scale of E amongst those with $G = 0$ is $0.05 - 0.02 = 0.03$; the effect of E amongst those with $G = 1$ is $0.09 - 0.04 = 0.05$. Thus, on the difference scale, the effect size is larger for the $G = 1$ subgroup, indicating that this is the subgroup we would like to treat if resources are limited. However, on the risk ratio scale, the effect for those with $G = 0$ is $0.05/0.02 = 2.5$ and for those with $G = 1$ it is $0.09/0.04 = 2.25$; the risk ratio effect size is larger for the $G = 0$ subgroup; however, this is not the subgroup we would want to allocate limited resources to. If we had only 100 doses of the drug, we could cure three additional persons if we used all of the drug supply amongst those with $G = 0$, but we could cure five additional person if we used all of the drug supply amongst those with $G = 1$. All other things being equal, we would clearly want to give the drug supply to those with $G = 1$. The possibility of positive additive interaction but negative or null multiplicative interaction is not simply a theoretical possibility. This was precisely the situation with the lung cancer data in Table 9.1, where we had a positive additive interaction but a negative multiplicative interaction. It was likewise the case in analyses of the joint effects of *Helicobacter pylori* and use of NSAIDs in causing peptic ulcer (Kuyvenhoven et al., 1999) with slightly positive additive interaction but negative multiplicative interaction. Similarly, in analyses of interaction between factor V Leiden mutation and oral contraceptive use in causing venous thrombosis, the multiplicative interaction was found to be close to null, but there was a positive additive interaction (Vandenbroucke et al., 1994). Using the multiplicative scale in any of these cases to determine on which subgroups to intervene would have given the wrong conclusion.

More generally, $p_{11} - p_{10} - p_{01} + p_{00} > 0$ implies the public health consequence of an intervention on E would be larger in the $G = 1$ group, while $p_{11} - p_{10} - p_{01} + p_{00} < 0$ implies the public health consequence of an intervention on E would be larger in the $G = 0$ group. Thus, while it may be of interest to assess multiplicative interaction, additive interaction should also in general be examined, if for no other reason than to assess public health relevance.

In some case–control study designs, only odds ratios can be evaluated and thus effect measures and interaction measures are evaluated on an odds ratio scale. The effects for each of the exposures considered separately and both considered together on the odds ratio scale are defined respectively by

$$OR_{10} = \frac{p_{10}/(1 - p_{10})}{p_{00}/(1 - p_{00})}$$

$$OR_{01} = \frac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})}$$
$$OR_{11} = \frac{p_{11}/(1-p_{11})}{p_{00}/(1-p_{00})}.$$

A measure of interaction on the multiplicative scale for odds ratio could then be taken as

$$\frac{OR_{11}}{OR_{10}OR_{01}} \tag{9.4}$$

This quantity measures the extent to which, on the odds ratio scale, the effect of both exposures together exceeds the product of the effects of the two exposures considered separately. If $OR_{11}/(OR_{10}OR_{01}) > 1$, the multiplicative interaction is said to be positive. If $OR_{11}/(OR_{10}OR_{01}) < 1$, the interaction is said to be negative. For the data in Table 9.1, we have

$$\frac{OR_{11}}{OR_{10}OR_{01}} = \frac{42.79}{8.71 \times 6.13} = 0.80$$

The measure of multiplicative interaction on the odds ratio scale is negative. The measure is very close to what was obtained for the multiplicative interaction on the risk ratio scale, that is, 0.78. In general, measures of multiplicative interaction on the odds ratio and risk ratio scale will be very close to one another whenever the outcome is rare. When the outcome is rare, both $(1 - p_{ge})$ and $(1 - p_{00})$ will be close to 1 and thus the odds ratios approximate risk ratios since

$$OR_{ge} = \frac{p_{ge}/(1-p_{ge})}{p_{00}/(1-p_{00})} \approx \frac{p_{ge}}{p_{00}} = RR_{ge}$$

Multiplicative interaction on the odds ratio and risk ratio scales can depart more substantially from one another when the outcome is common.

We may also be interested in assessing additive interaction from data when only relative risks are available or reported. Although we may not be able to estimate the additive interaction in (9.2)—that is, $p_{11} - p_{10} - p_{01} + p_{00}$—directly, we can still proceed as follows. If we divide (9.2) by p_{00}, we obtain the following:

$$RR_{11} - RR_{10} - RR_{01} + 1 \tag{9.5}$$

This quantity is sometimes referred to as the "relative excess risk due to interaction" or RERI (Rothman, 1986). It is also sometimes referred to as the "interaction contrast ratio" or ICR (Greenland et al., 2008). This gives us something similar to additive interaction but using risk ratios rather than risks. Subsequently, we will refer to this quantity in (9.5) as $RERI_{RR}$. We have that $RERI_{RR} > 0$ if and only if for the additive interaction in (9.2), $p_{11} - p_{10} - p_{01} + p_{00} > 0$; likewise $RERI_{RR} < 0$ if and only if $p_{11} - p_{10} - p_{01} + p_{00} < 0$; and $RERI_{RR} = 0$ if and only if $p_{11} - p_{10} - p_{01} + p_{00} = 0$. Thus, we can assess whether additive interaction is positive, negative, or zero using risk ratios and $RERI_{RR}$. It should be noted

that although $RERI_{RR}$ gives the direction (positive, negative, zero) of the additive interaction, we cannot in general use $RERI_{RR}$ to make statements about the relative magnitude of the underlying additive interaction for risks, $p_{11} - p_{10} - p_{01} + p_{00}$, unless we know p_{00}. We may have $RERI_{RR}$ larger in one of two subpopulations, but the additive interaction for risks, $p_{11} - p_{10} - p_{01} + p_{00}$, may be larger in the other; this is because the baseline risks, p_{00}, may differ and $RERI_{RR}$ depends on the baseline risk (Skrondal, 2003).[1] However, again, only the direction, rather than the magnitude, of $RERI_{RR}$ is needed to draw conclusions about the public health relevance of interaction. If we are trying to decide which subgroup of G to target for an intervention when resources are limited, $RERI_{RR} > 0$ implies the public health consequences of an intervention on E would be larger in the $G = 1$ group, while $RERI_{RR} < 0$ implies the public health consequences of an intervention on E would be larger in the $G = 0$ group.

Two other measures of additive interaction using data from risk ratios or odds ratios are sometimes employed. The so-called synergy index (Rothman, 1986) is defined as

$$S = \frac{RR_{11} - 1}{(RR_{10} - 1) + (RR_{01} - 1)}$$

It measures the extent to which the risk ratio for both exposures together exceeds 1, as well as whether this is greater than the sum of the extent to which each of the risk ratios considered separately each exceed 1. Suppose the denominator of S is positive, then if $S > 1$ we will have $RERI_{RR} > 0$ and thus $p_{11} - p_{10} - p_{01} + p_{00} > 0$; and if $S < 1$, then we will have $RERI_{RR} < 0$ and thus $p_{11} - p_{10} - p_{01} + p_{00} < 0$. Thus, the synergy index can likewise be used to assess additive interaction. As with $RERI_{RR}$, the risk ratios in the synergy index are often replaced, and approximated, by odds ratios when it is used in practice because of case–control study designs. The interpretation of the synergy index becomes difficult in settings in which one or both of the exposures is preventive rather than causative (Knol et al., 2011).[2] Another

1. For example, suppose that the risks for G and E, stratified by gender, are: for males $p_{00} = 0.02$, $p_{01} = 0.03$, $p_{10} = 0.03$, $p_{11} = 0.06$ and for females $p_{00} = 0.01$, $p_{01} = 0.02$, $p_{10} = 0.02$, $p_{11} = 0.05$. Then the additive interaction for risks for males is $p_{11} - p_{10} - p_{01} + p_{00} = 0.02$ and for females it is also $p_{11} - p_{10} - p_{01} + p_{00} = 0.02$. However, if we examine $RERI_{RR}$ for males, we get $RERI_{RR} = (p_{11} - p_{10} - p_{01} + p_{00})/p_{00} = 1$ but for females we obtain $RERI_{RR} = (p_{11} - p_{10} - p_{01} + p_{00})/p_{00} = 2$. We have a higher $RERI_{RR}$ for females than for males even though the underlying additive interaction for risks is the same. We obtain a higher $RERI_{RR}$ for females because the baseline risk for females $p_{00} = 0.01$ is lower than for males, $p_{00} = 0.02$. Again $RERI_{RR}$ can be used to assess the direction (positive, negative, zero) of the additive interaction for risks but not the magnitude of the additive interaction for risks. If the magnitude (rather than just the sign) of $RERI_{RR}$ is going to be interpreted, then it must be kept in mind that this magnitude is on the excess relative risk scale, and this does not necessarily correspond to the relative magnitude of additive interaction for risks. Once again, this is because the baseline risks may differ across groups.

2. When one or both of the exposures is preventive, rather than causative (i.e., $RR_{10} < 1$ and/or $RR_{01} < 1$), such that the denominator of S, $(RR_{10} - 1) + (RR_{01} - 1)$, is less than 0, then with an inequality like $S > 1$, multiplying both sides of this inequality by $(RR_{10} - 1) + (RR_{01} - 1)$, which is negative, will reverse the sign of the inequality, because of multiplication by a negative number, to give $RR_{11} - 1 < (RR_{10} - 1) + (RR_{01} - 1)$ or $RERI_{RR} < 0$; and thus when the denominator of S is

measure of additive interaction that is sometimes used is called the attributable proportion and is defined as

$$AP = \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11}}$$

and essentially measures the proportion of the risk in the doubly exposed group that is due to the interaction itself. The attributable proportion is essentially a derivative measure of the relative excess risk due to interaction: $AP > 0$ if and only if $RERI_{RR} > 0$, and $AP < 0$ if and only if $RERI_{RR} < 0$. A variant on the attributable proportion may also be potentially of interest. The attributable proportion measure above, $AP = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}} = \frac{p_{11}-p_{10}-p_{01}+p_{00}}{p_{11}}$, essentially measures the proportion of *risk* in the doubly exposed group that is due to interaction. Alternatively, we might consider the proportion of the joint *effects* of both exposures together that is due to interaction (Rothman, 1986; VanderWeele, 2013e; VanderWeele and Tchetgen Tchetgen, 2014). This measure is given by $AP^* = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}-1} = \frac{p_{11}-p_{10}-p_{01}+p_{00}}{p_{11}-p_{00}}$. Its properties will be considered later in the chapter in Section 9.12 on attributing effects to interactions.

All of these measures can be used in cohort studies, but these measures are also of interest and can be employed in case–control studies as well. Suppose that we only have estimates for odds ratios but that the outcome is rare (or that the controls are selected from the entirety of the underlying population rather than just from the non-cases; cf. Knol et al., 2008) so that odds ratios approximate risk ratios. We could then replace each of the risk ratios in $RERI_{RR}$, the synergy index S, or the attributable proportion measures, with odds ratios to obtain approximations to each of these measures of additive interaction. For example, for the relative excess risk due to interaction, we can define $RERI_{OR} = OR_{11} - OR_{10} - OR_{01} + 1$, which is the odds ratio analogue of $RERI_{RR}$. If the outcome is rare, then we have that

$$\begin{aligned} RERI_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &\approx RR_{11} - RR_{10} - RR_{01} + 1 = RERI_{RR} \end{aligned}$$

Thus, provided that the outcome is rare (or the controls are selected from the entirety of the underlying population rather than just from the non-cases) so that odds ratio approximate risk ratios, we can assess additive interaction, at least approximately, even if only estimates of odds ratios are available from case–control study designs. Note that for this argument to apply using the assumption of a rare outcome (10% is often used as a threshold in practice), the outcome must be rare in each stratum defined by the two exposures. Sampling controls for the entire underlying population rather than only the non-cases removes the need for this rare outcome assumption (cf. Knol et al., 2008).

As an example, Figueiredo et al. (2004) studied the effects of XRCC3-T241M polymorphisms and alcohol consumption on breast cancer risk using a case–control

negative, $S < 1$ becomes the condition for positive additive interaction, which can be confusing. In general it is thus best not to report S unless the denominator, $(RR_{10} - 1) + (RR_{01} - 1)$, is positive.

Table 9-4. ODDS RATIOS FOR BREAST CANCER BY STRATA OF ALCOHOL CONSUMPTION AND XRCC3-T241M

	No Alcohol	**Alcohol**
T/T or T/M	1	1.12
M/M	1.21	2.09

study design. The genetic risk factor was considered the M/M genotype versus a reference of the T/T or T/M genotype. They obtained the odds ratios. in Table 9.4 from their case–control study.

Although we cannot assess additive interaction directly using risks, $p_{11} - p_{10} - p_{01} + p_{00}$, from the odds ratios in Table 9.4, we can still estimate

$$\begin{aligned} RERI_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &= 2.09 - 1.21 - 1.12 + 1 = 0.76 > 0 \end{aligned}$$

and so we would have evidence of positive additive interaction. Breast cancer is a relatively rare outcome, and so odds ratios will closely approximate risk ratios in this study. Likewise, we could calculate the synergy index $S = \frac{RR_{11}-1}{(RR_{10}-1)+(RR_{01}-1)} = 3.30 > 1$, again indicating positive additive interaction. And we can calculate the proportion of risk in the doubly exposed group attributable to interaction, $AP = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}} = 36.4\%$, or the proportion of the joint effects of both exposures attributable to interaction, $AP^* = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}-1} = 69.7\%$.

9.2. STATISTICAL INTERACTIONS AND STATISTICAL INFERENCE

In practice, interactions are often evaluated by using statistical models by including a product term for the two exposures in the model. A statistical model on the linear scale accommodating interaction might take the form

$$P(Y = 1|G = g, E = e) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg \qquad (\mathbf{9.6})$$

Under this model it is easy to verify that $\alpha_0 = p_{00}$, $\alpha_1 = p_{10} - p_{00}$, $\alpha_2 = p_{01} - p_{00}$, and $\alpha_3 = p_{11} - p_{10} - p_{01} + p_{00}$. The coefficient α_3 is thus equal to our measure of additive interaction based on risks; for this reason, α_3 is sometimes referred to as a statistical interaction on the additive scale.

Similarly, one might use a log-linear model for risk ratios, including a product term:

$$\log\{P(Y = 1|G = g, E = e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg \qquad (\mathbf{9.7})$$

Here we have that $e^{\beta_0} = p_{00}$, $e^{\beta_1} = RR_{10}$, $e^{\beta_2} = RR_{01}$, and $e^{\beta_3} = RR_{11}/(RR_{10}RR_{01})$. The so-called "main effects," β_1 and β_2, when exponentiated, simply give the risk ratios for each of the two exposures when each is considered

alone. The coefficient β_3, when exponentiated, gives our measure for multiplicative interaction for risk ratios, $RR_{11}/(RR_{10}RR_{01})$. The coefficient β_3 is thus often referred to as a statistical interaction for a log-linear model. Likewise, one might use a logistic model for odds ratios, including a product term:

$$logit\{P(Y=1|G=g,E=e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg \qquad (\mathbf{9.8})$$

Here we have that $e^{\gamma_0} = p_{00}/(1-p_{00})$, $e^{\gamma_1} = OR_{10}$, $e^{\gamma_2} = OR_{01}$, and $e^{\gamma_3} = OR_{11}/(OR_{10}OR_{01})$. The main effects, γ_1 and γ_2, when exponentiated, simply give the odds ratios for each of the two exposures. The coefficient γ_3, when exponentiated, gives our measure for multiplicative interaction for odds ratios, $OR_{11}/(OR_{10}OR_{01})$. Thus, γ_3 is referred to as a statistical interaction for a logistic model. The equality $e^{\gamma_0} = p_{00}/(1-p_{00})$ will only hold with cohort data. However, all the other equalities, $e^{\gamma_1} = OR_{10}$, $e^{\gamma_2} = OR_{01}$, and $e^{\gamma_3} = OR_{11}/(OR_{10}OR_{01})$, will hold for both cohort data and case–control data. We can thus assess both of the main effects of the exposure and the multiplicative interaction between the exposures on an odds ratio scale using case–control data.

When the outcome and both exposures are binary, and no further covariates are included, it is straightforward to fit these models to the data using standard software. The estimate and confidence intervals obtained by maximum likelihood estimation and given by such software for α_3 will constitute an estimate and confidence interval for the additive interaction $p_{11} - p_{10} - p_{01} + p_{00}$. The estimate and confidence intervals obtained by maximum likelihood estimation and given by such software for β_3 and γ_3, when exponentiated, will constitute an estimate and confidence interval for the multiplicative interaction on the risk ratio and odds ratio scales, respectively. Statistical inference for interaction is thus straightforward in these cases.

Often we may want to control for other covariates in models (9.6)–(9.8). For example, we may want to fit the following analogous models which include an additional vector of covariates, C:

$$\begin{aligned} P(Y=1|G=g,E=e,C=c) &= \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg + \alpha_4' c \\ \log\{P(Y=1|G=g,E=e,C=c)\} &= \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg + \beta_4' c \\ \text{logit}\{P(Y=1|G=g,E=e,C=c)\} &= \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c. \end{aligned}$$

Unfortunately, the linear and log-linear models, when fit to data, will often run into convergence problems in the maximum likelihood algorithms used to fit the models, especially when there are continuous covariates in C, because the models do not ensure that the predicted probabilities lie between 0 and 1. The logistic model with covariates does not suffer from this problem. For this reason, the most common approach to assessing interaction in practice has become fitting the logistic model with covariates and assessing the estimate and confidence interval for the product term coefficient, γ_3, in this model. This approach is also popular because it can be implemented in a straightforward way with case–control data as well. The coefficient, γ_3, is an important and useful measure of interaction, and proceeding with this strategy is recommended.

However, as discussed throughout this chapter, it is also recommended that investigators assess additive interaction as well. This can be more challenging when covariates are in the model. Additional strategies to fit linear and log-linear models with covariates using data from cohort studies have been described elsewhere (cf. Yelland et al., 2011; Knol et al., 2012, for overviews of several different methods) but the use of these is still not common. In the next section, however, we will describe what has now become a fairly standard approach (Hosmer and Lemeshow, 1992) to estimating additive interaction, with covariate control, which consists of using a logistic regression with additional covariates and transforming the parameter estimates to obtain estimates and confidence intervals for the relative excess risk due to interaction (*RERI*).

9.3. INFERENCE FOR ADDITIVE INTERACTION

Suppose the following model is fit to the data:

$$\text{logit}\{P(Y=1|G=g,E=e,C=c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c \qquad \textbf{(9.9)}$$

We then have that

$$\begin{aligned} RERI_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &= e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1 \end{aligned}$$

Thus we can estimate a measure of additive interaction, $RERI_{OR}$, using the parameters of a logistic regression. This approach has the advantage that the logistic regression in (9.9) can more easily be fit to data when there are continuous covariates than the corresponding linear or log-linear models for binary outcomes given in the previous section. This approach with logistic regression also has the advantage that it can be employed even with case–control data. Standard errors for $RERI_{OR}$, as estimated above, can be obtained using the delta method (Hosmer and Lemeshow, 1992). Software options are now available to estimate these standard errors (e.g., Lundberg et al., 1996; Andersson et al., 2005).[3] Below, we provide some simple SAS and Stata code to estimate $RERI_{OR}$ and its standard error using the delta method (see also Ai and Norton, 2003; Norton et al., 2004). Finally, as an online supplement to VanderWeele and Knol (2014), there is an Excel spreadsheet that can be used in conjunction with standard output from logistic regression (output on parameter estimates and either the covariance or correlation estimates) using any software package.

3. Easy to implement software (Richardson and Kaufman, 2009; Kuss et al., 2010) is also available for estimating $RERI_{OR}$ using so-called linear odds models (cf. Skrondal, 2003). This approach, however, can have difficulty handling continuous covariates *C*. Such covariates can be handled in linear odds models by using a weighting approach for covariate control (VanderWeele and Vansteelandt, 2011) and this approach can be employed in case–control data as well. All of these approaches for additive interaction with a linear odds model, however, do require a rare outcome so that $RERI_{OR}$ approximates $RERI_{RR}$.

The approach described above works well if the outcome is rare so that $RERI_{OR}$ approximates $RERI_{RR}$. If the outcome is common, $RERI_{OR}$ may not be an adequate measure of additive interaction. In such cases, for cohort data, one could estimate $RERI_{RR}$ by replacing the logistic model in (9.9) with a log-linear model and then calculating the relative excess risk due to interaction by $RERI_{RR} = e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1$ from the coefficients of model (9.7) above, or its analogue with covariates, though such log-linear models with continuous covariates C may not always converge. An approach for risk ratios using modified Poisson, rather than logistic regression, has also been proposed that can be used with a common outcome (Zou, 2008). Alternatively, with cohort data with a common outcome, one may use a weighting approach to estimating additive interaction (VanderWeele et al., 2010).

To estimate standard errors for $RERI_{OR}$ using logistic regression, in addition to the delta method described by Hosmer and Lemeshow (1992) and implemented with SAS or Stata code below, one may also use bootstrapping, which can have more accurate standard errors when the sample size is small (Assmann et al., 1996); other resampling-based approaches are available when some of the outcome counts for particular exposure combinations are low (Nie et al., 2010). Bayesian approaches to $RERI_{OR}$ are also now available (Chu et al., 2011). When sample sizes are relatively large, the approaches to estimating $RERI_{OR}$ will give fairly comparable confidence intervals; when sample sizes are small, the resampling-based approach may be more accurate. However, in general, fairly large sample sizes are required to detect interaction; thus, for the most part, in those very settings in which it is possible and reasonable to test for interaction, the various approaches to estimate $RERI_{OR}$ are likely to give comparable estimates and standard errors. We discuss issues of power and sample size in Chapter 13.

Our discussion thus far has focused on binary exposures. A similar approach can be used with ordinal or continuous exposures. The logistic regression model above in (9.9) could be fit to the data if the two exposures G and E were ordinal or continuous. However, when additive interaction is carried out for ordinal or continuous exposures using this approach based on logistic regression, two things must be kept in mind, one analytical and one interpretative. First analytically, for ordinal and continuous exposures, it is important to consider the magnitude of the change in the exposures for which one is examining interaction. If one is considering a change for the value of G from g_0 to g_1 and a value of E from e_0 to e_1, then instead of using $e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$ as an estimate of $RERI_{OR}$, one uses

$$RERI_{OR} = e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} - e^{(g_1-g_0)\gamma_1+(g_1-g_0)e_0\gamma_3} - e^{(e_1-e_0)\gamma_2+(e_1-e_0)g_0\gamma_3} + 1$$

This needs to be taken into account when using the software and Excel spreadsheets so that estimates and covariance matrices are multiplied by the appropriate factors. This is described in more detail in the following section. Similar expressions could be given using categorical exposures: Under any specific statistical model and for any two levels of each of the two exposures, one simply calculates the three relative risks comparing the various exposure combinations to the reference group and

one subtracts from the risk ratio of the doubly exposed group the two risk ratios for each of the singly exposed groups and adds 1. The second, more interpretative point, when ordinal, continuous, or categorical exposures are being employed, is that it is important to keep in mind that the $RERI_{OR}$ measure does vary according to the levels being compared and can vary in sign as well. The additive interaction measure for a change in E from 10 to 20 and in G from 0 to 1 may be different from the additive interaction measure for a change in E from 20 to 30 and in G from 0 to 1, but again as noted above, the $RERI_{OR}$ measure should be interpreted as giving the direction of additive interaction (positive, negative, or zero) and its relative magnitude does not necessarily correspond to the relative magnitude of the additive interaction for absolute risks. See also Knol et al., (2007) for further discussion.

9.4. SAS AND STATA CODE FOR ADDITIVE INTERACTION FROM LOGISTIC REGRESSION

Here we will present SAS and Stata code for estimating additive interaction from logistic regression. This code could fairly easily be adapted for other software packages. Other SAS code has also been provided elsewhere (Lundberg et al., 1996; Andersson et al., 2005). However, most of this code is given for a parameterization of logistic regression that has separate indicator variables for each combination of the exposure levels whereas the parameterization that is used most often in practice is one in which each exposure has a main effect and then a product term is also included in the logistic regression. Here we will give code for this more common parameterization.

9.4.1. SAS Code for Additive Interaction from Logistic Regression for Binary Exposures

Suppose we have a dataset named 'mydata' with outcome variable 'd', exposure variables 'e' and 'g' and three covariates 'c1', 'c2' and 'c3'. To calculate the relative excess risk due to interaction we can run a standard logistic regression in SAS using proc logistic where we add 'outest=myoutput covout' to the procedure statement and then we also run the code that follows. The output will include the estimate of $RERI_{OR}$, its standard error, and a 95% confidence interval. If the "class" statement is used in SAS for categorical covariates then the exposures should not be included in the class statement or SAS will recode the exposures and this will lead to the wrong computation for $RERI_{OR}$. Also it is important that in the model statement, the two exposures and their interaction are listed as the first three covariates; the confounding variables can be included in any order after that.

```
proc logistic descending data=mydata outest=myoutput covout;
  model d=g e g*e c1 c2 c3;
run;

data rerioutput;
  set myoutput;
```

```
  array mm {*} _numeric_;
  b0=lag4(mm[1]);
  b1=lag4(mm[2]);
  b2=lag4(mm[3]);
  b3=lag4(mm[4]);
  v11=lag2(mm[2]);
  v12=lag(mm[2]);
  v13=mm[2];
  v22=lag(mm[3]);
  v23=mm[3];
  v33=mm[4];
  k1=exp(b1+b2+b3)-exp(b1);
  k2=exp(b1+b2+b3)-exp(b2);
  k3=exp(b1+b2+b3);
  vreri=v11*k1*k1 + v22*k2*k2 + v33*k3*k3 + 2*v12*k1*k2
  + 2*v13*k1*k3 + 2*v23*k2*k3;
  reri=exp(b1+b2+b3)-exp(b1)-exp(b2)+1;
  se_reri=sqrt(vreri);
  ci95_l=reri-1.96*se_reri;
  ci95_u=reri+1.96*se_reri;
  keep reri se_reri ci95_l ci95_u;
  if _n_=5;
run;

proc print data=rerioutput;
  var reri se_reri ci95_l ci95_u;
run;
```

9.4.2. SAS Code for Additive Interaction for Ordinal and Continuous Exposures

We can adapt this code also to calculate RERI for exposures which are ordinal or continuous. Suppose we wish to calculate the relative excess risk due to interaction comparing two different levels of the first exposure "g," say level 0 to level 2, and two different levels of our second exposure "e," say level 5 to level 25. We could then use the code below. Mathematical justification is given in the Appendix. In the code below, the user must input the two levels being compared for both exposures at the beginning of the data step, for example, "g1 = 2; g0 = 0; e1 = 25; e0 = 5" or whatever values are of interest in comparing. Note that if the user fixes "g1 = 1; g0 = 0; e1 = 1; e0 = 0" then this will give the same output as the previous code above for binary exposures. If the "class" statement is used in SAS for categorical covariates, then the exposures should not be included in the class statement or SAS will recode the exposures and this will lead to the wrong computation for $RERI_{OR}$. Also it is important that in the model statement, the two exposures and their interaction are listed as the first three covariates; the confounding variables can be included in any order after that.

```
proc logistic descending data=mydata outest=myoutput covout;
  model d=g e g*e c1 c2 c3;
run;
```

```
data rerioutput;
  set myoutput;
  g1=2;
  g0=0;
  e1=25;
  e0=5;
  array mm {*} _numeric_;
  b0=lag4(mm[1]);
  b1=lag4(mm[2]);
  b2=lag4(mm[3]);
  b3=lag4(mm[4]);
  v11=lag2(mm[2]);
  v12=lag(mm[2]);
  v13=mm[2];
  v22=lag(mm[3]);
  v23=mm[3];
  v33=mm[4];
  k1=(g1-g0)*exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-(g1-g0)*exp((g1-g0)*b1+(g1-g0)*e0*b3);
  k2=(e1-e0)*exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-(e1-e0)*exp((e1-e0)*b2+(e1-e0)*g0*b3);
  k3=(g1*e1-g0*e0)*exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-(g1-g0)*e0*exp((g1-g0)*b1+(g1-g0)*e0*b3)
-(e1-e0)*g0*exp((e1-e0)*b2+(e1-e0)*g0*b3);
  vreri=v11*k1*k1 + v22*k2*k2 + v33*k3*k3 + 2*v12*k1*k2
  + 2*v13*k1*k3
+ 2*v23*k2*k3;
  reri=exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)-
  exp((g1-g0)*b1+(g1-g0)*e0*b3)
-exp((e1-e0)*b2+(e1-e0)*g0*b3)+1;
  se_reri=sqrt(vreri);
  ci95_l=reri-1.96*se_reri;
  ci95_u=reri+1.96*se_reri;
  keep reri se_reri ci95_l ci95_u;
  if _n_=5;
run;

proc print data=rerioutput;
 var reri se_reri ci95_l ci95_u;
run;
```

9.4.3. SAS Code for Additive Interaction for Categorical Exposures

For categorical exposures, to obtain estimates and confidence intervals for additive interaction, one can restrict attention to two specific levels of each of the two variables and calculate measures of additive interaction using the code for binary exposures above. It is possible to proceed in this manner for each possible comparison of two levels of each of the two exposures. For example, if there were two categorical variables, A and B, and A had three levels (A1, A2, A3) and B had four levels (B1, B2, B3, B4), then one could assess additive interaction comparing A = A1 and A = A2 and B=B1 and B = B4 by ignoring the observations with A = A3 and also ignoring those with B = B2 or B = B3 and then using the code for binary

exposures above. Suppose the name of the dataset with the categorical variables was mycatdata. We could then use the following SAS code:

```
data mydata;
   set mycatdata;
   if A='A1' then g=0;
   if A='A2' then g=1;
   if B='B1' then e=0;
   if B='B4' then e=1;
   if A='A1' or A='A2';
   if B='B1' or B='B4';
run;
```

The code deletes the observations with A = A3, and those with B = B2 or B = B3 and creates a new dataset only with values of A which are A1 or A2 and with values of B which are B1 or B4. The code for additive interaction for binary exposures can then be used directly. We could similarly proceed with any other comparison. We could compare (A1, A2) and (B1, B2); or (A1, A2) and (B1, B3); or (A1, A3) and (B1, B2); and so on.

9.4.4. Stata Code for Additive Interaction for Binary Exposures

Suppose we have a dataset with outcome variable "d," exposure variables "g," and "e," and three covariates "c1," "c2," and "c3." To calculate the relative excess risk due to interaction, we can create an interaction variable "Ige," then run a standard logistic regression in Stata using the logit command, and then use the Stata "nlcom" command in the code that follows. The output will include the estimate of RERI, its standard error, and a 95% confidence interval.

```
generate Ige = g*e
logit d g e Ige c1 c2 c3
nlcom exp(_b[g]+_b[e]+_b[Ige])-exp(_b[g])-exp(_b[e])+1
```

9.4.5. Stata Code for Additive Interaction for Ordinal and Continuous Exposures

We can also calculate RERI using Stata for exposures that are ordinal or continuous. Suppose we wish to calculate the relative excess risk due to interaction comparing two different levels of the first exposure "g," say level 0 to level 2, and two different levels of our second exposure "e," say level 5 to level 25. We could then use the code below. In this code the user must specify, in the first four lines of code, the levels of both exposures that are being compared (in the code below the two levels for "g" are 2 and 0 and the two levels for "e" are "25" and "5," but these can be changed). If the user fixes g1 = 1, g0 = 0, e1 = 1, and e0 = 0, then the code will give the same output as the previous code above for binary exposures. The next two lines of code generate an interaction variable between "g" and "e" and fit the logistic regression model allowing for interaction. The final line of code uses the "nlcom" command

in Stata to obtain RERI. The output will include the estimate of RERI, its standard error, and a 95% confidence interval.

```
generate g1=2
generate g0=0
generate e1=25
generate e0=5

generate Ige = g*e
logit d g e Ige c1 c2 c3

nlcom exp((g1-g0)*_b[g]+(e1-e0)*_b[e]+(g1*e1-g0*e0)*_b[Ige])
  -exp((g1-g0)*_b[g]+(g1-g0)*e0*_b[Ige])
  -exp((e1-e0)*_b[e]+(e1-e0)*g0*_b[Ige])+1
```

9.4.6. Stata Code for Additive Interaction for Categorical Exposures

For categorical exposures, to obtain estimates and confidence intervals for additive interaction, one can restrict attention to two specific levels of each of the two variables and calculate measures of additive interaction using the code for binary exposures above. It is possible to proceed in this manner for each possible comparison of two levels of each of the two exposures. For example, if there were two categorical variables, A and B, and A had three levels (A1, A2, A3) and B had four levels (B1, B2, B3, B4), then one could assess additive interaction comparing A = A1 and A = A2, and B = B1 and B = B4, by ignoring the observations with A = A3 and also ignoring those with B = B2 or B = B3 and then using the code for binary exposures above. We could create the restricted dataset using the following Stata code:

```
generate g=0 if A==’A1’;
replace g=1 if A==’A2’;
generate e=0 if B==’A1’;
replace e=1 if B==’B4’;
```

The code for additive interaction for binary exposures can then be used directly. The code creates variables g and e only for those observations with values of A which are A1 or A2 and with values of B which are B1 or B4. When the code for additive interaction for binary exposures is used, it will only analyze the observations with values of A which are A1 or A2 and with values of B which are B1 or B4 since those with values of A which are A3 or with values of B which are B2 or B3 will have their values of g and of e missing.

We could similarly proceed with any other comparison. We could compare (A1, A2) and (B1, B2); or (A1, A2) and (B1, B3); or (A1,A3) and (B1,B2); and so on.

9.5. ADDITIVE VERSUS MULTIPLICATIVE INTERACTION

The fact that interaction can be assessed on different scales and that interaction is scale-dependent raises the question on which scale interaction should be

assessed: additive or multiplicative or some other. In general it is almost always best to present both additive and multiplicative measures of interaction (Botto and Khoury, 2001; Vandenbroucke et al., 2007; Knol and VanderWeele, 2012). In practice, measures of multiplicative interaction, using logistic regression, are most frequently reported. This is very likely simply done because of convenience, rather than because careful thought has been given to which measure is to be preferred. Standard software using logistic regression will automatically give an estimate and confidence interval for multiplicative interaction. As noted in the previous section, additional work is required in most current software packages to obtain measures of additive interaction, and for this reason it is not often done. In a recent review of a random sample of 25 cohort and 50 case–control studies from the five most highly ranked epidemiological journals, Knol et al. (2009) noted that although 61% of the studies included at least as secondary analyses an assessment of effect modification or interaction, only one reported a measure of additive interaction. In our view, it is in general a mistake to not report additive interaction. As noted above, additive interaction is always relevant for assessing the public health significance of an interaction. Although we believe both additive and multiplicative interaction should in general be reported, we nonetheless review some of the reasons that have been put forward for using one scale versus the other.

The difference scale is useful for assessing the public health importance of interventions and the public health significance of interaction (Blot and Day, 1979; Saracci, 1980; Rothman et al., 1980; Greenland et al., 2008). As noted above, if the effect of an intervention is larger on the difference scale in one subgroup versus another, then this indicates that there would be larger numbers for whom the disease was prevented/cured in giving 100 individuals in the first subgroup the intervention versus giving 100 individuals in the second subgroup the intervention. Such information is useful for targeting subpopulations for which the intervention is most effective. This will be relevant whenever resources are constrained, and thus relevant also for cost-effectiveness (Greenland, 2009). As discussed above, the additive, not the multiplicative, scale gives this information. A second reason sometimes given for using additive interaction is that it more closely corresponds to tests for mechanistic interaction, rather than merely statistical interaction (Greenland et al., 2008; VanderWeele and Robins, 2007b, 2008; VanderWeele, 2010b,c). As discussed more extensively below, tests for additive interaction can sometimes be used to detect synergism in Rothman's (1976) sufficient cause framework. Conceived of another way, assessing additive interaction can sometimes be used to assess whether there are persons for whom the outcome would occur if both exposures were present but not if only one or the other of the exposures were present. As discussed below and more extensively in the next chapter, this ends up being a different, and in many cases stronger, notion of interaction than merely a statistical interaction. Such mechanistic interaction has sometimes also been interpreted as "biological interaction." However, it is now been pointed out on numerous occasions and in different ways that statistical interaction and even "mechanistic interaction" need not imply that the two exposure interact in any biological or physical way (Siemiatycki and Thomas, 1981; Thomas, 1991; VanderWeele and Robins, 2007b; Phillips, 2008; Cordell, 2009). Again we will return to this issue in the next chapter.

Nonetheless, detecting mechanistic interaction may be another reason to use the additive scale. Finally, as will be discussed in Chapter 13, tests for additive interaction are sometimes more powerful than tests for multiplicative interaction; thus for the purposes of discovery and detection, the additive scale may be preferred as well.

Several reasons are also often also put forward for using the multiplicative scale. First, as noted above, it is easier to fit multiplicative models (such as logistic regression), and the multiplicative scale is the most natural scale on which to assess interaction for such models; moreover, when using such models, measures of multiplicative interaction are readily obtained from standard software. Second, it is sometimes claimed that there is in general less heterogeneity on the multiplicative scale. Studies of meta-analyses have suggested that in terms of statistical significance, the risk ratio and odds ratio are less heterogeneous than the risk difference (Engels et al., 2000; Sterne and Egger, 2001; Deeks and Altman, 2003).[4] However, it is not entirely clear the extent to which this is simply due to difference in power across the different scales or whether there is genuinely less heterogeneity. Nevertheless, if it is indeed the case that the multiplicative scales (odds ratio or risk ratio) are "less heterogeneous," then detecting an interaction on a multiplicative scale may be of greater import than detecting interaction on the additive scale. A third reason sometimes given for using the multiplicative scale for overall effects (but also potentially applicable to interaction), stated in some epidemiology textbooks, is that the relative effect measures are better suited to "assessing causality." According to Poole (2010), this notion can be traced back to a paper by Cornfield et al. (1959) showing that smoking was strongly related to lung cancer but not to other diseases on a relative risk scale, while smoking seemed similarly related to lung cancer and also to other diseases on an absolute risk scale. Because specificity of effect was seen as a criterion of causality (Hill, 1965), the relative risk scale was seen as superior over the absolute risk scale in assessing causality. As noted by Poole (2010), whether the relative or absolute measure is more useful for "assessing causality" will, however, vary by setting. In some cases, such as that considered by Cornfield et al., (1959), the multiplicative scale may indeed prove to be more useful.

Arguments can be given in favor of each of the two scales. However, nothing prohibits investigators from reporting measures of interaction on both additive and multiplicative scale; and, in most settings, we think that this approach is the best because both can be informative (Botto and Khoury, 2001; Vandenbroucke et al., 2007; Knol and VanderWeele, 2012). The presence or absence of interaction on either scale is of interest.

4. Engels et al. (2000) found that for 107 of 125 meta-analyses (86%) the *p*-value for heterogeneity for risk differences was less than that for the odds ratios. With a *p*-value cut-off of 0.10, they found that 59 (47%) meta-analyses were heterogeneous for the risk difference and 44 (35%) were heterogeneous for the odds ratio. Deeks and Altman (2003) likewise report that the risk difference was more heterogenous than the odds ratio or risk ratio using 1889 meta-analyses. Sterne and Egger (2001) reviewed 78 meta-analyses and found that the *p*-value for heterogeneity was less than 0.05 in 29%, 27%, and 35% of these meta-analyses, for the odds ratio, risk ratio and risk difference, respectively.

9.6. CONFOUNDING AND THE INTERPRETATION OF INTERACTION: INTERACTION VERSUS EFFECT HETEROGENEITY

Thus far, we have considered measures of interaction using risk differences, risk ratios, and odds ratios. In general, however, we want to know whether our effect estimates correspond to causal effects rather than mere associations. In observational studies we thus attempt to control for confounding. Analytically, this is often done through regression adjustment for other covariates. In interaction analyses we have two exposures and thus potentially two sets of confounding factors to consider. The causal interpretation of interaction measures depends on whether control has been made for one or both sets of confounding factors, or neither.

Suppose we have made control for one set of confounding factors, those for the relationship between our primary exposure of interest and the outcome, but that we have possibly not controlled for confounding of the relationship between the secondary factor defining subgroups and the outcome. We would in this case still be able to obtain valid estimates of the effect of the primary exposure within strata defined by our secondary factor. For example, suppose we found substantial interaction between a drug and hair color when examining some health outcome. If we had controlled for the confounding factors for the drug–outcome relationship, or if the drug were randomized, we could interpret our interaction measure as a measure of heterogeneity concerning how the actual causal effect of the drug varied across subgroups defined by hair color. If we found that the effect of our primary exposure varied by strata defined by the secondary factor in this way, then we might call this "effect heterogeneity" or "effect modification." This might be useful, for example, in decisions about which subpopulations to target in order to maximize the effect of interventions. Provided that we have controlled for confounding of relationship between the primary exposure and the outcome, these estimates of effect modification or effect heterogeneity could be useful even if we have not controlled for confounding of the relationship between the secondary factor and the outcome. What we would not know, however, is whether the effect heterogeneity were *due to* the secondary factor itself, or something else associated with it. If we have not controlled for confounding for the secondary factor, the secondary factor itself may simply be serving as a proxy for something that is causally relevant for the outcome (VanderWeele and Robins, 2007c). For example, if we found that the effect of the drug varied by strata defined by hair color, this may simply be due to the fact that hair color is associated with genotype, and it is this that is causally relevant for modifying the effect of the drug on the outcome. If we were simply to dye someone's hair, this would not change the effect of the drug.

If we are interested principally in assessing the effect of the primary exposure within subgroups defined by a secondary factor then simply controlling for confounding for the relationship between the primary exposure and the outcome is sufficient. However, if we want to intervene on the secondary factor in order to change the effect of the primary exposure then we need to control for confounding of the relationships of both factors with the outcome. When we control for

confounding for both factors we might refer to this as "causal interaction" in distinction from mere "effect heterogeneity" mentioned above (VanderWeele, 2009c).

As another example, VanderWeele and Knol (2011a) consider a randomized trial for a housing intervention program for homeless adults to reduce the number of hospitalizations. Suppose that the effect of the housing program were examined within strata defined by whether the participants had at least part time employment. Here, the housing program is randomized, but employment status is not. If it were found that the housing intervention had a larger effect for those with part-time employment than for those without, this could be used as a valid estimate for the effect of the intervention within these different subgroups and could be useful in subsequently targeting the intervention towards the subgroups for which it would be most effective. By randomization, we have controlled for confounding for the housing intervention, but we have not necessarily controlled for confounding for employment status. Thus, while we could get valid estimates of effects of the housing intervention within strata defined by employment status, we could not draw conclusions on what would happen if we intervened on employment status as well to try to improve the effect of the intervention. Again, employment status has not been randomized. Employment status may, for instance, be serving as a proxy for mental health, and it may be that mental health is in fact what is relevant in altering the effects of the intervention. It is possible that if we intervened on employment status, without changing mental health, then this would not alter at all the effect of the housing intervention. We would only be able to assess what the effect of interventions on employment status in altering the effect of the housing intervention would be if we had controlled for confounding of the relationship between the factor defining subgroups, namely employment status, and the outcome.

In summary, if we are interested in identifying which subpopulations it is best to target with a particular intervention, then assessing effect heterogeneity is fine and only one set of confounding factors need be considered (though even here it is sometimes argued that control for other factors can help with external validity and extrapolation to other settings). If we are interested in potentially intervening on the secondary factor to change the effects of the primary intervention (or if we are interested in assessing mechanistic interaction, described below), then we want measures of causal interaction and we would need to control for confounding for the relationships between both factors and the outcome.

In practice, typically a regression model is simply fit to the data of the outcome on the two exposures, a product term, and possibly other covariates, such as the regression models considered in Section 9.2. However, whether the regression coefficient for the product term can be interpreted as a measure of effect heterogeneity or causal interaction or both or neither depends on what confounding factors have been controlled for. For effect modification we only have one set of confounding factors to consider, just those for the relationship between the primary exposure and the outcome. For causal interaction we have two sets of confounding factors to consider, those for the primary exposure and the outcome and those for the secondary factor and the outcome. Epidemiologists make an effort to control for confounding and think carefully about confounding in observational studies for overall causal

effects. However, too often issues of confounding have been neglected in interaction analyses. Careful thought needs to be given to interaction analyses in interpreting associations as causal and in distinguishing between whether attempt is being made to control for one or both sets of confounding factors; and to whether "effect heterogeneity" (also sometimes called "effect modification") or "causal interaction" is of interest; this will of course depend upon the context.[5]

The terms "interaction" and "effect modification" in practice are often used interchangeably. In some sense, what we have called "effect modification" is still a type of interaction analysis; and what we have called "causal interaction" could almost be viewed as "effect modification" by intervening on a secondary variable (VanderWeele, 2009c, 2010c). There is some ambiguity in terminology, and it would be difficult to insist on a particular set of rules for terminology. However, even if the terms themselves are used interchangeably, it is important to keep in mind that there are still two distinct concepts present. The distinction again has to do with whether one or two potential interventions are in view. Failure to take the distinction into account could lead to incorrect policy recommendations. In writing papers, researchers can make clear which of the two concepts is in view (without having to adopt a strict terminological stance) by clarifying, in a Methods section, whether confounding control is intended for one or both exposures and by commenting, in a Discussion section, whether interventions on one or both exposures are being considered when interpreting the implications of the results.

9.7. PRESENTING INTERACTION ANALYSES

Careful thought should be given to the presentation of interaction analyses. Very often when interaction or effect modification is of interest, effect measures are presented for each stratum separately using separate reference groups. Suppose, for example, we had data as in Table 9.1 and that effect measures were computed on the risk ratio scale. We let $E = 1$ denote asbestos exposure and $E = 0$ the absence of asbestos exposure, and we let $G = 1$ denote smoking and $G = 0$ non-smoking. It is not uncommon for papers to present, for example, the (adjusted) risk ratio effect measures for say, the exposure E separately across strata of the other factor G. For example, the effect measures might be presented as in Table 9.5.

While this information can be useful to see that the risk ratio in the non-smoking $(G = 0)$ stratum is larger than the risk ratio in the smoking $(G = 1)$ stratum, and for calculating multiplicative interaction $(4.74/6.09 = 0.78$ as above), there

5. Additional subtlies also arise in distinguishing between interaction and effect modification. For example, VanderWeele (2009a) showed that there can be cases in which effect modification is present but not interaction or when interaction is present but not effect modification. Likewise, there are also cases in which effect modification measures are identified from the data, but interaction measures are not; there are more subtle cases in which interaction measures are identified from the data but effect modification measures are not. Finally, VanderWeele (2009a) also discusses how the analytic procedures required to fit marginal structural models (Robins et al., 2000a) for effect modification differ from those required to fit marginal structural models for interaction.

Table 9-5. Risk Ratios with Separate Reference Groups (Uninformative Presentation)

	No Asbestos $(E = 0)$	**Asbestos** $(E = 1)$
Non-smoker $(G = 0)$	1 (reference)	RR = 6.09
Smoker $(G = 1)$	1 (reference)	RR = 4.74

Table 9-6. Risk Ratios with a Common Reference Group (Informative Presentation)

	No Asbestos $(E = 0)$	**Asbestos** $(E = 1)$
Non-smoker $(G = 0)$	1 (reference)	6.09
Smoker $(G = 1)$	8.64	40.91

are several other comparisons for which Table 9.5 is uninformative. For example, by presenting the analyses with separate reference groups (for each of the $G = 0$ and $G = 1$ strata), we will not know from such a presentation whether the $(G = 0, E = 1)$ subgroup or the $(G = 1, E = 0)$ subgroup is at higher risk for the outcome. In fact, simply from the information in Table 9.5, we would not know whether the $(G = 1, E = 1)$ subgroup or the $(G = 0, E = 1)$ subgroup is at higher risk for the outcome, or whether the $(G = 1, E = 0)$ subgroup or the $(G = 0, E = 0)$ subgroup is at higher risk for the outcome. Nor do we know from Table 9.5 what the sign is for measures of additive interaction. Because of these reasons, the current guidelines (Vandenbroucke et al., 2007; Knol and VanderWeele, 2012) recommend that interaction and effect modification analyses be presented with a single common reference group, say the $(G = 0, E = 0)$ subgroup, or that the original data be presented (Botto and Khoury, 2001). If risk ratios with a common reference group were used for the data in Table 9.1, the effects could then be presented in Table 9.6.

From the information presented in Table 9.6, which uses a common reference group, we would know that the ordering of risk across $G \times E$ subgroups was $(G = 0, E = 0)$, then $(G = 0, E = 1)$, then $(G = 1, E = 0)$, and then $(G = 1, E = 1)$. We could still calculate the individual risk ratios for E in the different strata of G as $6.09/1 = 6.09$ for $G = 0$ and $40.91/8.64 = 4.74$ for $G = 1$ (and we could also add these to the table if desired). We could thus also estimate measures of multiplicative interaction. We could moreover estimate the risk ratios for G across strata of E: for example, $3.5/1 = 3.5$ for $E = 0$ and $5.0/2.5 = 2$ for $E = 1$ (and we could present these in the table if desired). And we could moreover estimate measures of additive interaction from the information in Table 9.6: $RERI_{RR} = 40.91 - 8.64 - 6.09 + 1 = 27.18 > 0$. The presentation of interaction analyses in Table 9.6 thus gives the reader far more information (using a single common reference category) than the presentation in Table 9.5 (using multiple reference categories). Presenting interaction analyses using a common reference category such as the presentation in Table 9.6 is thus to be preferred. If the study is a cohort study then it may be even further preferable to present the actual risks, as in Table 9.1, in the cells of the table, rather than the risk ratios (Botto and Khoury, 2001; Knol and VanderWeele, 2012).

Knol and VanderWeele (2012) further suggest that when interaction and effect modification analyses are presented, the following items should all be given in a table: (1) relative risks (RRs), odds ratios (ORs), or risk differences (RDs) for each (G,E) stratum with a single reference category (possibly taken as the stratum with the lowest risk of the outcome); (2) RRs, ORs, or RDs for G within strata of E, and for E within strata of G; (3) interaction measures on additive and multiplicative scales, along with confidence intervals and p-values for these; (4) the exposure–outcome confounders for which adjustment has been made either for one of the exposures (for effect modification/heterogeneity analyses) or for both of the exposures (for interaction analyses) with clear indication of whether attempt is being made to control for one or two sets of confounding factors. Knol and VanderWeele (2012) also consider different layout options for this information and how to further extend such presentations when one or both exposures has more than two levels. If multiple different interaction analyses are conducted in the same paper and presented in the same table, it may be desirable to put all of these items on a single line of a table so that multiple interactions analyses can be presented in the same table.

Careful thought should be given to presenting interaction analyses so that the reader has the maximum amount of information available. In almost all cases, interaction analyses with a single reference group should be presented. Failure to do so will obscure information from the reader.

9.8. SYNERGISM AND MECHANISTIC INTERACTION

Thus far we have been considering different notions of statistical interaction and their interpretation. We noted above that such notions of interaction were scale-dependent. In this section we will consider drawing conclusions about more mechanistic forms of interaction. We might say that a "sufficient cause interaction" is present if there are individuals for whom the outcome would occur if both exposures were present but would not occur if just one or the other exposure were present (VanderWeele and Robins, 2007b, 2008). If we let Y_{ge} denote the counterfactual outcome (the outcome that would have occurred) for each subject if, possibly contrary to fact, G had been set to g and E had been set to e, then a sufficient cause interaction is present if for some individual $Y_{11} = 1$ but $Y_{10} = Y_{01} = 0$. This is in some sense a "mechanistic interaction" insofar as when both exposures are present, the outcome is turned "on," but when only one or the other exposure is present, the outcome is turned "off." Note that a sufficient cause interaction does require some individual with $Y_{11} = 1$ but $Y_{10} = Y_{01} = 0$ but does not require $Y_{00} = 0$ for this individual. Further below we will also consider an even stronger notion of "mechanistic interaction" which requires some individual for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$. However, we will begin our discussion of mechanistic interaction with the slightly weaker notion of a sufficient cause interaction because this is all that is required for synergism between G and E within the sufficient cause framework (Rothman, 1976).

Additive interaction is sometimes used to test for such mechanistic or sufficient cause interaction. However, having positive additive interaction only implies such sufficient cause interaction under additional assumptions. If it can be assumed that both exposures are never preventive for any individual (formally, if Y_{ge} is nondecreasing in g and e for all individuals), then provided that control is also made for confounding of both exposures[6], positive additive interaction, $p_{11} - p_{10} - p_{01} + p_{00} > 0$, suffices for sufficient cause interaction (Greenland et al., 2008; VanderWeele and Robins, 2007b). The assumption that neither exposure can ever be preventive for any individual is sometimes referred to as a positive "monotonicity" assumption; it is a strong assumption. Again it requires that the exposure always be either neutral or causative for every individual in the population; it can never be preventive for any individual. In some contexts monotonicity might be plausible. For example, we would probably never think that smoking is protective for lung cancer for any individual. There may be some persons for whom smoking causes lung cancer, there may be others for whom smoking is neutral, but we would never think that smoking prevents lung cancer for anyone (i.e., that they would not have lung cancer if they smoked, but that they would have lung cancer if they did not smoke). Thus the positive monotonicity assumption for the effect of smoking on lung cancer may be plausible. But in other cases the assumption may be less plausible. For example, if we were to consider the effect of alcohol consumption on stroke, alcohol may be protective for stroke in some persons but causative for others; the monotonicity assumption would not be plausible here. Positive monotonicity requires that the effect never be preventive for the outcome for any person in the population. Importantly, to assess sufficient cause interaction simply by examining whether additive interaction is positive requires that the effects of both exposures on the outcome be monotonic. This will in many contexts be a strong assumption, and it is an assumption that is not possible to verify empirically; it must be established on substantive grounds.

Fortunately, it is also possible to test for sufficient cause interaction even without such monotonicity assumptions, but the standard tests for positive additive interaction no longer suffice. Alternative tests must be used. VanderWeele and Robins (2007b, 2008) showed that if the effect of the two exposures are unconfounded, then

$$p_{11} - p_{10} - p_{01} > 0$$

would imply the presence of a sufficient cause interaction. This is a stronger condition than regular positive additive interaction in (9.2), which only requires $p_{11} - p_{10} - p_{01} + p_{00} > 0$ because with the condition $p_{11} - p_{10} - p_{01} > 0$ we are no longer adding back in the outcome probability p_{00} for the doubly unexposed group. This condition for a sufficient cause interaction thus does not correspond to, and is stronger than, the regular test for additive interaction, or than simply examining whether interaction is positive in a statistical model (VanderWeele, 2009d). In these various cases, the magnitude of the contrast $p_{11} - p_{10} - p_{01} + p_{00}$ with

6. Formally, we say that the effects of both exposures are unconfounded if the counterfactual outcomes D_{ge} are independent of the actual exposures $\{G, E\}$.

monotonicity or $p_{11} - p_{10} - p_{01}$ without monotonicity in fact gives a lower bound on the prevalence of individuals manifesting sufficient cause interaction patterns (VanderWeele et al., 2010). If data are only available on the ratio scale, then if both exposures have positive monotonic effects on the outcomes, we can test for sufficient cause interaction by re-expressing the condition $p_{11} - p_{10} - p_{01} + p_{00} > 0$ as $RERI_{RR} > 0$. Under this monotonicity assumption, we can thus test for sufficient cause interaction by testing $RERI_{RR} > 0$. Likewise, the condition $p_{11} - p_{10} - p_{01} > 0$ without imposing monotonicity assumptions can be expressed in terms of $RERI_{RR}$ as $RERI_{RR} > 1$; again this is stronger than simply the ordinary condition for additive interaction $RERI_{RR} > 0$. However, $RERI_{RR}$ still can be used in a straightforward way to test for such sufficient cause interaction without imposing the monotonicity assumption by testing whether $RERI_{RR} > 1$ rather than simply $RERI_{RR} > 0$.

Note that when the empirical conditions above are satisfied, the conclusion is that there are some individuals for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$; the conclusion is not that all individuals have this response pattern. Note also that these conditions given here are sufficient but not necessary for sufficient cause interaction; that is, if these conditions are satisfied, then a sufficient cause interaction must be present, but if the conditions are not satisfied, then there may or may not be a sufficient cause interaction—one simply cannot determine this from the data. The conditions given here are the weakest possible empirical conditions to test for sufficient cause interaction without making further assumptions (VanderWeele and Richardson, 2012).

VanderWeele (2010b,c) discussed empirical tests for an even stronger notion of interaction. We might say that there is a "singular" or "epistatic" interaction if there are individuals in the population who will have the outcome if and only if both exposures are present; in counterfactual notation, this is that there are individuals for whom $Y_{11} = 1$ but $Y_{10} = Y_{01} = Y_{00} = 0$. In the genetics literature, when gene–gene interactions are considered, such response patterns are sometimes called instances of "compositional epistasis" (Phillips, 2008; Cordell, 2009) and constitute settings in which the effect of one genetic factor is masked unless the other is present. VanderWeele (2010b,c) noted that if the effects of the two exposures on the outcome were unconfounded, then

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

would imply the presence of such an "epistatic interaction." Again this is an even stronger notion of interaction; in this condition for "epistatic interaction" we are now subtracting p_{00}. The condition $p_{11} - p_{10} - p_{01} - p_{00} > 0$ expressed in terms of $RERI_{RR}$ is equivalent to $RERI_{RR} > 2$.

For epistatic interactions, if the effect of at least one of the exposures is positive monotonic (Y_{ge} is nondecreasing in at least one of g and e), then $p_{11} - p_{10} - p_{01} > 0$ suffices for an epistatic interaction and tests for $RERI_{RR} > 1$ could be used; if the effect of both exposures are positive monotonic, then $p_{11} - p_{10} - p_{01} + p_{00} > 0$ suffices and tests for $RERI_{RR} > 0$ could be used to test for an epistatic interaction (VanderWeele, 2010b,c). These results are summarized in Table 9.7.

Table 9-7. Relations Between the Additive Relative Excess Risk Due to Interaction (*RERI*) and Forms of Mechanistic Interaction Under Different Monotonicity Assumptions

Monotonicity Assumption	$RERI_{RR}>0$	$RERI_{RR}>1$	$RERI_{RR}>2$
No assumptions about monotonicity		S	S, E
One of *G* or *E* have positive monotonic effects		S, E	S, E
Both *G* and *E* have positive monotonic effects	S, E	S, E	S, E

Note: "S" indicates the presence of a sufficient cause interaction; "E" denotes an epistatic interaction.

In assessing additive interaction using $RERI_{RR}$, it thus is useful to examine not only whether the estimate and confidence interval for $RERI_{RR}$ are greater than 0 (i.e., whether there is additive interaction), but also whether the estimate and confidence interval for $RERI_{RR}$ are all greater than 1 or are all greater than 2; this is because $RERI_{RR}$ of this magnitude would provide evidence for mechanistic interaction (sufficient cause or epistatic interaction) without the need for additional assumptions. When the outcome and exposures are binary, we recommend that investigators examine not only additive interaction but also whether $RERI_{RR}$ exceeds the values of 1 or of 2.

For example, using age-standardized measures, Bhavnani et al. (2012) report that risk ratios for diarrheal disease across groups infected with rotavirus and/or *Giardia*. With the doubly unexposed group as the reference category, the risk ratio for rotavirus (in the absence of *Giardia*) is 2.63, the risk ratio for *Giardia* (in the absence of rotavirus) is 1.13, and the risk ratio when both rotavirus and *Giardia* are present is 10.72. This gives an $RERI_{RR}$ of $RERI_{RR} = 10.72 - 2.63 - 1.13 + 1 = 7.96$ (95% CI: 3.13, 18.92). The value of $RERI_{RR}$, and its entire 95% confidence interval exceeds the value 2, suggesting strong evidence for mechanistic interaction (both 'sufficient cause' and 'epistatic' interaction) without needing to make any monotonicity assumptions at all. Extensions of these ideas to exposures with more than two levels (VanderWeele, 2010b–d) and to multi-way interactions between three or more exposures (VanderWeele and Robins, 2008; VanderWeele and Richardson, 2012) as well as for settings with causal antagonism in which the presence of one exposure may block the operation of the other (VanderWeele and Knol, 2011b) are discussed in the next chapter.

Although tests like these give conclusions about mechanistic interaction, it should be noted that even such "mechanistic interaction," does not imply that the two exposures are *physically* interacting in any real sense (Siemiatycki and Thomas, 1981; Thomas, 1991; VanderWeele and Robins, 2007b; Phillips, 2008; Cordell, 2009). We should thus distinguish between (i) statistical interaction on the one hand and (ii) mechanistic interaction (e.g., the outcome occurs if and only if both exposures are present) on the other and, finally, (iii) "biological" or "functional" interaction in which the two exposures physically interact to bring about the outcome (Phillips, 2008; Cordell, 2009; VanderWeele, 2010b; VanderWeele, 2011e). We return to these issues also at greater length in the next chapter.

9.9. INTERACTIONS FOR CONTINUOUS OUTCOMES AND TIME-TO-EVENT OUTCOMES

9.9.1. Continuous Outcomes

When continuous outcomes are in view, linear and log-linear regression can still be used to estimate measures of additive and multiplicative interaction respectively. For additive interaction, a linear regression model for the continuous outcomes could be used:

$$\mathbb{E}[Y|G=g,E=e,C=c]=\alpha_0+\alpha_1 g+\alpha_2 e+\alpha_3 eg+\alpha_4' c$$

and α_3 can be taken as a measure of additive interaction. This parameter is equal to the additive interaction measure:

$$\begin{aligned}\alpha_3 = {} & \mathbb{E}[Y|G=1,E=1,C=c]-\mathbb{E}[Y|G=1,E=0,C=c]\\ & -\mathbb{E}[Y|G=0,E=1,C=c]+\mathbb{E}[Y|G=0,E=0,C=c]\end{aligned}$$

For multiplicative interaction, a log-linear regression model for the continuous outcome could be used:

$$\log\{\mathbb{E}[Y=1|G=g,E=e,C=c]\}=\beta_0+\beta_1 g+\beta_2 e+\beta_3 eg+\beta_4' c$$

and β_3 can be taken as a measure of multiplicative interaction. This parameter is equal to the multiplicative interaction measure:

$$e^{\beta_3}=\frac{\mathbb{E}[Y|G=1,E=1,C=c]/\mathbb{E}[Y|G=1,E=0,C=c]}{\mathbb{E}[Y|G=0,E=1,C=c]/\mathbb{E}[Y|G=0,E=0,C=c]}$$

Note that with a continuous outcome, most of the arguments for preferring one scale to another are no longer applicable. With a continuous outcome we generally no longer run into convergence problems for the additive scale. But the argument for the public health significance of the additive scale is not as applicable for a continuous outcome as we are no longer analyzing discrete events. Moreover, with a continuous outcome, it is not clear that the additive scale gives any insight into mechanistic interaction. Whether additive or multiplicative scales are to be preferred for a continuous outcome will generally depend on the distribution of the outcome data.

9.9.2. Time-to-Event Outcomes

Let T be a time-to-event outcome. Let G and E be two dichotomous exposures of interest and let C be a collection of covariates. Let $S(t;g,e,c)$ and $\lambda(t;g,e,c)$ denote, respectively, the survival function and hazard function at time t conditional on $G=g,E=e,C=c$. Li and Chambless (2007) consider interaction in a proportional

hazards model that takes the form

$$\lambda(t;g,e,c) = \lambda_0(t)e^{\eta_1 g + \eta_2 e + \eta_3 ge + \eta_4' c} \quad (\mathbf{9.10})$$

where $\lambda_0(t)$ is the baseline hazard at time t and $(\eta_1, \eta_2, \eta_3, \eta_4)$ are the model parameters. The coefficient η_3, when exponentiated, gives a measure of multiplicative interaction for hazard ratios—that is, $e^{\eta_3} = \frac{HR(1,1;c)}{HR(1,0;c)\times HR(0,1;c)}$, where $HR(g,e;c)$ is the hazard ratio, conditional on $C = c$, comparing $G = g$, $E = e$ with the reference group $G = 0, E = 0$. Li and Chambless propose using as a measure of additive interaction the relative excess risk due to interaction on the hazard ratio scale:

$$RERI_{HR} = e^{\eta_1+\eta_2+\eta_3} - e^{\eta_1} - e^{\eta_2} + 1$$

which equivalently is

$$HR(1,1;c) - HR(1,0;c) - HR(0,1;c) + 1$$

Li and Chambless discuss estimation, testing, and confidence intervals for this quantity, $RERI_{HR}$. However, once model (9.10) is fit to the data, inference for $RERI_{HR}$ can essentially proceed in the same manner as was discussed above for $RERI_{OR}$. The same software options (Excel spreadsheet and/or SAS code) could be used to give confidence intervals for $RERI_{HR}$. As discussed above, with binary outcomes, $RERI_{OR}$ will indicate the direction of the additive interaction for risks under a rare outcome or by sampling the controls from the entirely underlying population rather than just the non-cases. Likewise, $RERI_{HR}$, will, under the assumption of a rare outcome, give the direction of additive interaction for survival probabilities.

VanderWeele (2011f) discussed causal interactions in the setting of a proportional hazards model. It was shown that if $RERI_{HR} > 1$, then there is some time at which causal interaction is present (i.e., there is a time at which some individuals would have the outcome if both exposures were present but not if just one or the other were present); if the effects of both exposures on the outcome are monotonic, then $RERI_{HR} > 0$ implies that there is some time at which causal interaction is present. VanderWeele (2011f) also discussed inference for the range of times at which there is evidence of causal interaction, and the interested reader is referred to that paper for further discussion.

9.10. IDENTIFYING SUBGROUPS TO TARGET TREATMENT

In this chapter our focus has been on estimating and interpreting interactions; we have focused on binary exposures but have considered also ordinal or continuous exposures. As we had noted above, one motivation for examining interaction is determining whether a particular intervention might be more effective for one subgroup than another. It was noted that assessing interaction on the additive scale was most important for this purpose. This motivation does, however, raise the question as to how to choose the variable or variables that are to define subgroups. Most of our discussion has presupposed that we have a particular secondary variable in mind

which will define subgroups and for which we will examine whether there is effect heterogeneity across subgroups. In some settings, data on many such variables that could potentially define subgroups may be available. One option would then be to use each of these and see if any of them are such that there is evidence for substantial effect heterogeneity. A downside of this approach is that by testing for effect heterogeneity across many variables, we are more likely to find spurious results suggesting effect heterogeneity by chance. We would need to correct for such "multiple testing" to mitigate this possibility; methods to do so is one of the topics in Chapter 12, which will focus on genetic factors (where often there are very many possible tests that might be conducted). An alternative approach and one that is often advocated in the literature is to decide in advance, based on substantive knowledge, which factor or factors are thought most likely to show evidence for effect heterogeneity and test for those alone.

An additional complication arises when the variable that is going to define subgroups is continuous. One might then have to decide what cutoff of the continuous variable one is to use in defining subgroups. One might also be interested in whether there is in some sense an optimal cutoff of such a continuous variable such that whenever the variable is above that level it is best to treat. Methods to address this type of question are now available for a single continuous variable (Bonetti and Gelber, 2000, 2005; Song and Pepe, 2004).

An even more general approach involves forming anticipated "effect scores" for each and every person in a sample or population based on many baseline covariates and then targeting treatment to those above a certain "effect score" threshold. This approach has the advantage of being able to incorporate information from many different covariates in defining subgroups to try to optimize the effect of treatment. One approach to forming such effect scores is to fit a regression model for the outcome on all or several covariates for the treated or exposed subjects and then to fit a separate model for the untreated or unexposed subjects. For each person in the sample, one can then use the models, once they are fit to the data, to get a predicted outcome (or probability of the outcome) under exposure and a predicted outcome (or probability of the outcome) under control. The difference between these two predicted outcomes would then be the individual "effect score." One might then consider targeting treatment to those only above a certain threshold. It would even be possible to compare different models for the outcome under exposed and control condition, or different sets of covariates, in these models, to see which has the "effects scores" that best allow one to predict the outcome and target subpopulations (Zhao et al., 2013).

The approach is appealing and intuitive. Several complications do, however, arise in trying to make inferences in this manner, but methods have been developing to help address these. One complication is "overfitting" if the same data are used to fit the models and to evaluate which of the effect scores and models and covariates have the best predictive properties in forming subgroups. The evaluation of the effect scores and models and covariates may be misleading because the model parameters were specifically estimated to fit the available data as best as possible; and if the same parameters were used to get predicted outcomes in a different sample drawn from the same population, its performance would not be as good.

Zhao et al. (2013) have proposed a cross-validation procedure that involves splitting the sample into a training dataset (which is used to fit the models) and an evaluation dataset (which is used to evaluate and compare effect scores and models and covariates) to address this problem. Based on simulations, they recommend using 4/5 of the data to fit the models and 1/5 to evaluate the models.

Another complication that can arise with this approach is that if the models to get predicted outcomes are not correctly specified, then the inferences about the effects for different subgroups defined by the effect score may be misleading. Cai et al. (2011) have proposed a two-stage approach which helps address this issue. They recommend fitting a parametric regression model for the treated and control subjects to form the effect scores and then to use nonparametric regression to estimate the effects of the treatment on the outcome across subgroups defined by these effect scores. They describe procedures to carry out inference and form confidence intervals for the effects across subgroups defined by the effect scores that are applicable even if the parametric models used to initially form the effect scores are not correctly specified.

These approaches to identifying subgroups for which to target treatment using multiple covariates are appealing and potentially powerful. More methodological development remains to be done so that these are easy to implement and optimally choose cutoffs; but as these methods develop, it is likely that they will be very useful in both observational and experimental research.

9.11. QUALITATIVE INTERACTION

In some cases, we might think that an exposure has a positive effect for one subgroup and a negative effect for a different subgroup. Such instances are sometimes referred to as "qualitative interactions" or "crossover interactions" (Peto, 1982; Gail and Simon, 1985).[7] Unlike statistical interactions in which the effects within two subgroups are both in the same direction, but simply differ in magnitude, qualitative interactions do not depend on the scale that is being used (de González and Cox, 2007). If there is a qualitative interaction on the difference scale, there will also be a qualitative interaction on the ratio scale, and vice versa.

As an example of such qualitative interaction, Gail and Simon (1985) consider data from a trial of two therapies for breast cancer, one of which does and the other of which does not involve tamoxifen. For young patients under age 50 with low progesterone receptor levels, the treatment without tamoxifen led to higher proportions who were disease-free at three years. However, for all other groups (who were either older, or had higher progesterone receptor levels, or both) the treatment with tamoxifen led to higher proportions who were disease-free at three years. Here we

7. The term "quantitative interaction" is sometimes used exclusively for interactions that are not qualitative interactions (Peto, 1982). However, others use the term "quantitative interaction" to describe a statistical interaction on any scale, and they prefer using "non-crossover interaction" for the presence of interaction that is not a "qualitative interaction" (Gail and Simon, 1985).

would likely want to give young patients with low progesterone receptor levels the treatment without tamoxifen and give others the treatment with tamoxifen.

In an example like this, we see then that qualitative interaction is very important in decision-making. We discussed above that in settings in which treatment is beneficial for everyone but the magnitude of the benefit varies across subgroups, additive interaction can be useful in assessing whether it would be better to target treatment to some subgroups rather than others if resources are limited. However, in such settings, if resources are not limited and the treatment is beneficial for everyone, we may well want to treat all subgroups. Qualitative interaction, in contrast, has implications for treatment decisions even if resources are unlimited. In the presence of qualitative interaction, we do not want to treat all subgroups, because the treatment is in fact harmful in some subgroups. If qualitative interaction is present, it is thus important to be able to detect it.

Several statistical approaches have been developed for testing for such qualitative interaction (e.g., Gail and Simon, 1985; Piantadosi and Gail, 1993; Pan and Wolfe, 1997; Silvapulle, 2001; Li and Chan, 2006). The details of these various approaches and their power properties do vary, but they all essentially coincide when one is simply testing for qualitative interaction between two subgroups. The approaches differ when examining qualitative interaction across three or more subgroups.[8] When testing for qualitative interaction across two subgroups, one particularly simple approach (Pan and Wolfe, 1997) to test for a qualitative interaction at the 5% significance level is to construct 90% confidence intervals for the exposure effect in each of the two subgroups. If, on a difference scale say, one of the 90% confidence intervals lies entirely above 0 and the other lies entirely below 0, then one would reject the null hypothesis of no qualitative interaction. Note that only 90% confidence intervals (not 95%) need to be constructed here—that is, confidence intervals of the form $(\overline{\mu} - 1.65 \times s, \overline{\mu} + 1.65 \times s)$, where $\overline{\mu}$ and s are the estimates of the effect and its standard error. These confidence intervals will be narrower than the usual 95% confidence intervals of the form $(\overline{\mu} - 1.96 \times s, \overline{\mu} + 1.96 \times s)$ and thus will

8. The various approaches do differ when testing for qualitative interaction using more than two subgroups. Pan and Wolfe (1997) describe a fairly straightforward way to carry out such testing. Their approach allows for multiple subgroups and allows also testing for qualitative interaction of at least a certain magnitude (rather than simply whether one of the effects is larger, and the other smaller, than 0); it essentially just requires constructing confidence intervals of various sizes depending on the number of subgroups. Their approach is equivalent to that described by Piantadosi and Gail (1993), sometimes referred to as the "range test," but the implementation described by Pan and Wolfe (1997) is easier to carry out. An alternative approach was proposed by Gail and Simon (1985) which involves not simply constructing confidence intervals for the effects in each subgroup but rather constructing a confidence interval for the sum of the positive versus negative standardized effects across subgroups. The approach of Gail and Simon (1985) tends to perform better when there are several subgroups with effects that are positive and several also with effects that are negative. The approach of Piantadosi and Gail (1993) and Pan and Wolfe (1997) tends to perform better if the effects in most of the subgroups are in one direction and there are only one or very few subgroups for which the effect is in the opposite direction. The motivation for these various approaches involving several subgroups is often having a continuous covariate or multiple covariates of interest that might define subgroups for which a qualitative interaction is thought to be present.

allow one to reject the null hypothesis of no qualitative interaction more often. One could alternatively carry out the analysis on a ratio scale and also construct 90% confidence intervals for the effects in each of the subgroups and examine whether one of these 90% confidence intervals was completely above 1 and whether the other was completely below 1.

A special case or limit case of qualitative interaction is what is sometimes called a pure interaction in which the exposure has no effect whatsoever in one subgroup but does have an effect in a different subgroup. Like qualitative interactions, pure interactions do not depend on the scale being used. An example of such a "pure" interaction might include the setting from genetic epidemiology in Chapter 2 in which the genetic variants on chromosome 15 seemed to only affect lung cancer for individuals who smoked (Li et al., 2010b); otherwise it seemed that the variants had no effect for those who do not smoke. We will consider this example again further below.

9.12. ATTRIBUTING EFFECTS TO INTERACTIONS

9.12.1. Attributing Joint Effects to Interactions

In Section 9.1 we discussed different measures concerning the proportion of risk or effect attributable to interaction. In fact, we can actually decompose the joint effects of the two exposures, G and E, into three components: (i) the effect due to G alone, (ii) the effect due to E alone, and (iii) the effect due to their interaction. On the risk difference scale, this decomposition is

$$p_{11} - p_{00} = (p_{10} - p_{00}) + (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})$$

where the first component, $(p_{10} - p_{00})$, is the effect due to G alone, the second component, $(p_{01} - p_{00})$, is the effect due to E alone; and the final component, $(p_{11} - p_{10} - p_{01} + p_{00})$, is just the standard additive interaction. We could then also compute the proportion of the joint effect due to G alone, $\frac{(p_{10}-p_{00})}{(p_{11}-p_{00})}$, due to E alone, $\frac{(p_{01}-p_{00})}{(p_{11}-p_{00})}$, and due to their interaction, $\frac{(p_{11}-p_{10}-p_{01}+p_{00})}{(p_{11}-p_{00})}$.

We can also carry out a similar decomposition on the ratio scale using excess relative risks. We can decompose the excess relative risk for both exposures, $RR_{11} - 1$, into the excess relative risk for G alone and for E alone and into the excess relative risk due to interaction, $RERI$. Specifically we have (VanderWeele and Tchetgen Tchetgen, 2014)

$$RR_{11} - 1 = (RR_{10} - 1) + (RR_{01} - 1) + RERI_{RR}$$

We could then likewise compute the proportion of the effect due to G alone, $\frac{RR_{10}-1}{RR_{11}-1}$, due to E alone, $\frac{RR_{01}-1}{RR_{11}-1}$, and due to their interaction, $\frac{RERI_{RR}}{RR_{11}-1}$.[9]

9. As discussed in Section 9.1, Rothman (1986) considered a measure of interaction that he called the attributable proportion, defined as $\frac{RERI}{RR_{11}}$; the denominator Rothman used was RR_{11}. The measure was meant to capture the proportion of the *disease* in the doubly exposed group that is due

Under the logistic regression model

$$\text{logit}\{P(Y=1|G=g,E=e,C=c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c \qquad (9.9)$$

for an outcome that is rare, the joint effect attributable to G alone, to E alone, and to their interaction are given approximately by

$$\frac{RR_{10}-1}{RR_{11}-1} \approx \frac{e^{\gamma_1}-1}{e^{\gamma_1+\gamma_2+\gamma_3}-1}$$
$$\frac{RR_{01}-1}{RR_{11}-1} \approx \frac{e^{\gamma_2}-1}{e^{\gamma_1+\gamma_2+\gamma_3}-1}$$
$$\frac{RERI_{RR}}{RR_{11}-1} \approx \frac{(e^{\gamma_1+\gamma_2+\gamma_3}-e^{\gamma_1}-e^{\gamma_2}+1)}{e^{\gamma_1+\gamma_2+\gamma_3}-1}$$

The expressions can be used even when control is made for covariates in the logistic regression. VanderWeele and Tchetgen Tchetgen (2014) provide SAS and Stata code to do this automatically and to calculate standard errors and confidence intervals for the proportions and also discuss extensions to exposures that are not binary.

to the interaction. Rothman (1986) also considered an alternative measure, $\frac{RERI}{RR_{11}-1}$, which captured the proportion of the *effect* of both exposures on the additive scale that is due to interaction. This latter definition is the measure used in the decomposition here (VanderWeele and Tchetgen Tchetgen, 2014). Most of the subsequent literature has focused on the former measure; but the latter measure, (i.e., using $RR_{11}-1$ as the denominator) in fact has some advantages (VanderWeele, 2013e). With Rothman's primary measure, $\frac{RERI}{RR_{11}}$, even if all of the joint effect were due to interaction so that the effect of G alone and E alone were both risk ratios of 1 (i.e., $RR_{10}=1$ and $RR_{01}=1$), we would nevertheless have that Rothman's primary attributable proportion measure would be $\frac{RERI}{RR_{11}} = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}} = \frac{RR_{11}-1-1+1}{RR_{11}} = \frac{RR_{11}-1}{RR_{11}} < 1$; that is, even if the entirety of the joint effect of both exposures were due to interaction, the attributable proportion measure is still less than 100%. The measure $\frac{RERI}{RR_{11}-1}$ does not have this issue. It is 100% when the main effects of G alone and E alone were both risk ratios of 1—that is, when the entirety of the joint effect is due to interaction. The measure $\frac{RERI}{RR_{11}-1}$ captures the proportion of the joint effect attributable to interaction. The attributable proportion of joint effects measure, $\frac{RERI}{RR_{11}-1}$, is also attractive from another standpoint. Skrondal (2003) criticized Rothman's original attributable proportion measure because, in the presence of covariates, if the risks follow a linear risk model that is additive in the covariates, $P(D=1|G=g,E=e,C=c) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 ge + \alpha_4 c$, then, although the additive interaction, $p_{11}-p_{10}-p_{01}+p_{00} = \alpha_3$, does not vary across strata of the covariates, Rothman's primary attributable proportion measure, $\frac{RERI}{RR_{11}} = \frac{\alpha_3}{\alpha_0+\alpha_1+\alpha_2+\alpha_3+\alpha_4 c}$, does vary across strata of the covariates. Skrondal also noted that *RERI* itself, which would be given here by $RERI = \frac{\alpha_3}{\alpha_0+\alpha_4 c}$, likewise depends on the covariates. However, the measure of the proportion of the joint effects attributable to interaction, $\frac{RERI}{RR_{11}-1} = \frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3}$, does not vary with the covariates and thus circumvents Skrondal's criticism. Likewise, the other two components in the decomposition, namely $\frac{RR_{10}-1}{RR_{11}-1} = \frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3}$ and $\frac{RR_{01}-1}{RR_{11}-1} = \frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3}$, also do not depend on the covariates. The decomposition of the joint effect of the two exposures into three components, (i) the effect due to G alone, (ii) the effect due to E alone, and (iii) the effect due to their interaction, thus entirely circumvents Skrondal's critique of *RERI* and Rothman's primary attributable proportion measure, $\frac{RERI}{RR_{11}}$.

We illustrate the decomposition with an example from genetic epidemiology presented by VanderWeele and Tchetgen Tchetgen (2014) using data from a case–control study of lung cancer at Massachusetts General Hospital of 1836 cases and 1452 controls. The study included information on smoking and genotype information on locus 15q25.1. For simplicity, we will code the exposure as binary so that smoking is ever versus never and the genetic variant is a comparison of 0 versus 1/2 T alleles at rs8034191. Analyses were restricted to Caucasians, and covariate data include age (continuous), gender, and educational history (college degree or more, yes/no). If we proceed with the decomposition of the joint effect, then the proportions attributable to G alone, to E alone, and to their interaction are

$$\frac{RR_{10}-1}{RR_{11}-1} \approx 0.8\% \text{ (95\% CI: } -6.2\%, 7.7\%)$$
$$\frac{RR_{01}-1}{RR_{11}-1} \approx 51.4\% \text{ (95\% CI: } 33.4\%, 69.4\%)$$
$$\frac{RERI}{RR_{11}-1} \approx 47.8\% \text{ (95\% CI: } 33.3\%, 62.3\%)$$

Almost none of the joint effect (comparing both G and E present to both absent) is due to the effect of genetic variant G in the absence of smoking E, about 51% is due to smoking E is the absence of G, and about 48% is due to the interaction between G and E.

Assessing the importance of an interaction can be difficult to do merely from the coefficient estimate of an interaction term in a statistical model, or from a p-value. The measures here that assess what proportion of the effects are attributable to interaction can be a much easier-to-interpret measure for assessing the importance of an interaction than simply using $RERI$ measure or a ratio measure of multiplicative interaction such as $OR_{11}/(OR_{10}OR_{01})$. Furthermore, as will be seen in Chapter 13, power to detect interaction is often quite low, so it is possible to have a substantial interaction that is not "statistically significant." It may be tempting in such circumstances to conclude that the interaction is not important, because it is not statistically significant; but often this would be a mistake because, again, a statistically insignificant interaction can in fact be quite substantial in magnitude. By assessing what proportion of effects are attributable to interaction using the measures above, one can have some sense as to the potential importance of an interaction without relying on tests of statistical significance.

9.12.2. Attributing Total Effects to Interactions

If the distribution of the two exposures G and E are independent in the population, then we can also decompose the total effect of one of the exposures (e.g., total effect of E) into two components. If we let p_e denote $P(Y=1|E=e)$, then we have the following decomposition of the total effect of E (VanderWeele and Tchetgen Tchetgen, 2014):

$$(p_{e=1}-p_{e=0}) = (p_{01}-p_{00}) + (p_{11}-p_{10}-p_{01}+p_{00})P(G=1)$$

This decomposes the overall effect of E on Y into two pieces: The first piece is the conditional effect of E on Y when $G = 0$, and the second piece is the standard additive interaction, $(p_{11} - p_{10} - p_{01} + p_{00})$, multiplied by the probability that $P(G = 1)$. In some sense then, we can attribute the total effect of E on Y to the part that would be present still if G were 0 (this is $p_{01} - p_{00}$), as well as to a part that has to do with the interaction between G and E (this is $(p_{11} - p_{10} - p_{01} + p_{00})P(G = 1)$). If we could remove the genetic exposure (i.e., set it to 0), we would remove the part that is due to the interaction and would be left with only $p_{01} - p_{00}$. Since we can do this decomposition, we might define a quantity $pAI_{G=0}(E)$ as the proportion of the overall effect of E that is attributable to interaction, with a reference category for the genetic exposure of $G = 0$, as

$$pAI_{G=0}(E) := \frac{(p_{11} - p_{10} - p_{01} + p_{00})P(G = 1)}{(p_{e=1} - p_{e=0})}$$

The remaining portion $(p_{01} - p_{00})/(p_{e=1} - p_{e=0})$ is the proportion of the effect of E that would remain if G were fixed to 0. VanderWeele and Tchetgen Tchetgen (2014) provide SAS and Stata code to do this automatically and handle more general cases and models. Note that the three-way decomposition above for joint effects did not require that the exposures be statistically independent of one another. However, the two-way decomposition for a total effect given here does require the exposure to be statistically independent. VanderWeele and Tchetgen Tchetgen (2014) also discuss similar, but more complex, decompositions when the two exposures, G and E are correlated. See also Chapter 14.

As already discussed, one of the motivations for studying interaction is to identify which subgroups would benefit most from intervention when resources are limited. In settings in which it is not possible to intervene directly on the primary exposure of interest, one might instead be interested in which other covariates could be intervened upon to eliminate much or most of the effect of the primary exposure of interest. The methods here for attributing effects to interactions can be useful in assessing this and identifying the most relevant covariates for intervention.

The reader who has already read Part I (Mediation Analysis) of this book may wonder about the relationships between the decompositions here for interaction and those in Part I for mediation. In Chapter 14 we discuss the relationships between these different decompositions and in fact give results that unify these decompositions. An interested reader could turn to Chapter 14 and read this discussion before returning to Chapter 10 and the rest of the material on interaction.

9.13. DISCUSSION

Here we have considered a number of, primarily statistical, concepts related to interaction analysis, but we have also touched about the causal interpretation of such interactions. We have seen that we can assess interaction on the additive or multiplicative scales. Multiplicative interaction is perhaps what is most often

assessed and reported in part because it comes easily with standard software output using logistic regression. However, in most cases when the outcome is binary at least, additive interaction is worth assessing on grounds of policy relevance. It can, as we have seen, also be used to shed insight into the presence of mechanistic interaction.

In subsequent chapters we will cover a number of more advanced or specialized topics including more discussion of mechanistic interaction, bias analysis for interaction, topics concerning interaction in genetic studies, and power and sample size calculations for interaction. However, a number of further topics concerning interaction will not be covered in this book, including methods to robustly estimate interaction even if models for the main effects are misspecified (Vansteelandt et al., 2008, 2012b; Tchetgen Tchetgen, 2010; Tchetgen Tchetgen and Robins, 2010) and methods to assess effect modification by time-varying covariates and/or exposures (Petersen et al., 2007; Robins et al., 2007a; VanderWeele et al., 2010; Almirall et al., 2010). Most of these approach require considerable programming to implement. The focus of the this chapter and remaining chapters are methods to assess interaction that can be implemented in a relatively straightforward manner.

10

Mechanistic Interaction

In the previous chapter we briefly discussed the notion of a mechanistic interaction, that an individual would have the outcome if both of two exposures were present but not if only just one or the other were present. In this chapter we will consider notions of mechanistic interaction in more detail. We will discuss the closely related notion of synergism within the sufficient cause framework, the joint presence of two exposures in the same sufficient cause. We will also discuss in greater detail another related notion, that of compositional epistasis in the genetics literature. Whereas the concepts, and methods for mediation focused on explanations regarding *how* particular exposures affected outcomes, and our discussion of interaction in the last chapter focused primarily on *for whom* and *to what extent* exposures affected outcomes, the concepts and methods in this chapter lie somewhere in between the two. As we will see, the notion of a mechanistic interaction, while providing some insight about *for whom* exposures affect outcomes, also sheds light on *why and how* particular exposures affect outcomes because they tell us something about the presence of mechanisms, specifically the joint presence of exposures in a mechanism for the outcome. Later in the chapter we will discuss various extensions to the ideas of a mechanistic interaction, to ordinal and categorical exposures, to settings with three or more exposures of interest, and to analogous ideas concerning antagonism. Many of these extensions are variations on the same theme, and the details can quickly multiply and get somewhat overwhelming and difficult to remember. The reader is thus encouraged to read through Sections 10.1–10.5, which introduce the basic concepts and then perhaps only skim Sections 10.6–10.9 and consult these on an as-needed basis. The reader can safely skip ahead to Section 10.10, which discusses the limits of the conclusions that can be drawn about actual physical or biologic interaction from empirical data. As we will see throughout, considerable progress can be made in testing for various forms of mechanistic interaction, but it is important to understand the limits of such inference. Moreover, as will be seen throughout, the tests for such mechanistic interaction are generally different and often stronger than ordinary tests for different forms of statistical interaction considered in the last chapter.

10.1. SUFFICIENT CAUSES AND SYNERGISM

The sufficient cause framework (Cayley, 1853; Mackie, 1965; Rothman, 1976) conceptualizes causation as a collection of different sufficient conditions or causes for the occurrence of an event. Each sufficient condition or cause is usually conceived of as consisting of various (necessary) components such that if all components are present, the sufficient cause is complete and the event or outcome occurs. Whereas the principal focus of the counterfactual approach to conceiving of causation (Neyman, 1923; Lewis, 1973; Rubin, 1974) is the cause or intervention itself, the sufficient cause framework instead considers primarily the effect. Said another way, the counterfactual framework asks questions of the effects of causes, whereas the sufficient cause framework asks questions of the causes of effects.

Early development of the sufficient cause framework in epidemiology appears in the writings of MacMahon and Pugh (1967), and its early development in philosophy appears in the work of Mackie (1965). Mackie proposed that when we refer to something as a "cause," it is then generally known to be a "an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result." Mackie used the term "INUS condition" as a shorthand for the expression in quotations in the previous sentence, the term "INUS" being derived from the first letter of each of the italicized words.

Although the ideas had been around for some time before, the sufficient cause framework came to popularity in epidemiology principally through the writing of Rothman (1976). Rothman provided a graphical schematic for the sufficient causes that have informally come to be known as "causal pies." If, for example, there were two sufficient causes for an outcome Y, say ABC and EFG, we might present these sufficient causes as two "causal pies" as in Figure 10.1. Rothman conceived of each of the sufficient causes or causal pies as a mechanism for bringing about the outcome. Each sufficient cause consists of multiple components such that if all components were present, then the sufficient cause would be complete, and some process would be set in motion such that the outcome would inevitably occur. Rothman referred to each component of a sufficient cause as a "component cause" or simply as a "cause." Rothman's conceptualization thus essentially matches that of Mackie (1965).

This approach to thinking about causation has also received attention in the legal literature (Wright, 1988). Wright (1988) proposed that the equivalent of an

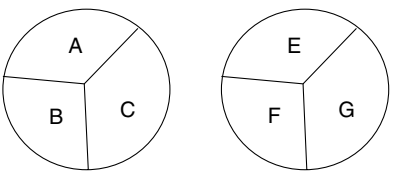

Figure 10.1 Two sufficient causes for the outcome, one requiring conditions *ABC*, the other *EFG*.

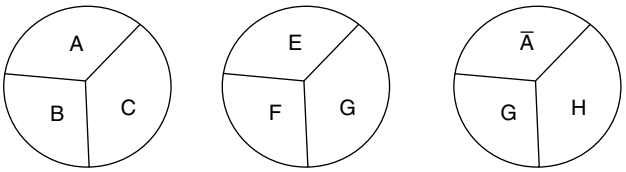

Figure 10.2 Three sufficient causes for the outcome including one which requires $\overline{A}$ i.e. the absence of *A*.

"INUS condition" in Mackie's (1965) language or a "component cause" in Rothman's (1976) be used as a standard for causation in legal reasoning. Wright referred to such a condition as a "NESS factor" where the term "NESS" is derived from the first letter of the italicized words in the expression "*necessary element* for the sufficiency of a *sufficient set*." The notion is again the same. Similar ideas also appear in the psychology literature (Cheng, 1997; Novick and Cheng, 2004). See VanderWeele (2012d) for an overview of the sufficient cause framework as it has emerged in different disciplines.

The sufficient cause framework allowed researchers to conceive of "interaction" or "synergism" in a manner that was not dependent on a particular statistical model. Rothman (1976) conceived of synergism between two factors as their joint presence in a particular sufficient cause. For example, if the only two sufficient causes for an outcome were those in Figure 10.1, then we would say that *A* and *B* exhibit synergism because they are two components of the same sufficient cause. We would say that *A* and *E* do not exhibit synergism because although *A* and *E* are both causes, there is no sufficient cause that has both as components. An exposure may be present in more than one sufficient cause, and it is said to exhibit synergism with another exposure if there is any sufficient cause of which the two are both components. It is also possible that the absence of an exposure is required for the operation of a sufficient cause. We will denote the absence of an exposure by the exposure with an overbar, that is, $\overline{A}$. Suppose, for instance, in Figure 10.2 that there were a third sufficient cause that perhaps required the absence of *A* (i.e., $\overline{A}$), the presence of *G*, and the presence of *H*. We would then have one sufficient cause that required the presence of *A*, namely *ABC*, and one that required the absence of *A*, namely $\overline{A}GH$. When a sufficient cause requires the presence of one exposure and the absence of another, then the two exposures are sometimes said to exhibit "antagonism." Thus, if this third sufficient cause were present, we would say that *A* and *G* exhibited antagonism and that *A* and *H* exhibited antagonism but that *G* and *H* exhibited synergism.

10.2. STATISTICAL INTERACTION WITH NO MECHANISTIC INTERACTION

In this section we will give a couple of examples to show that statistical interaction may be present between two exposures without there being any synergism between them in the sufficient cause framework. These examples will motivate empirical

tests for synergism presented in the next section which provide conditions under which we can conclude from statistical models that synergism is present.

Suppose a particular disease Y can arise only through one of three mechanisms, one involving a genetic factor G, one involving a environmental factor E, and one involving neither the genetic nor the environmental factor as in Figure 10.3.

The genetic factor alone may not be sufficient for developing disease Y, but suppose that the genetic factor G in combination with some other factors, denoted by A_1, will lead to disease. Similarly, suppose that the environmental factor E in combination with some other factors, denoted by A_2, will lead to disease. Finally, let A_0 denote the set of factors necessary for the third mechanism for the disease, which does not require the genetic nor the environmental factor to operate. There are thus three mechanisms for the disease as in Figure 10.3, one involving G and A_1, one involving E and A_2, and one involving just A_0. Note that there is no mechanism requiring both G and E to operate in this example—that is, no synergism between G and E. For simplicity, suppose that the distributions of G, E, A_0, A_1, and A_2 are all statistically independent in the population. Because A_0, A_1, and A_2 are all independent of, and do not affect, G or E, the variables A_0, A_1, and A_2 do not confound effects of G and E on Y. Suppose the probabilities of G and E are 0.2 and 0.5 respectively, and that the probabilities of A_0, A_1, and A_2 are all 0.10. Suppose the target population of interest has 10000 individuals; the expected cross-classification of cases and non-cases by genetic and environmental exposure status is given in Table 10.1.

We have $P(Y = 1|G = 0, E = 0) = P(A_0 = 1) = 0.10$, $P(Y = 1|G = 1, E = 0) = P(A_0 = 1 \text{ or } A_1 = 1) = 1 - (1 - 0.10)(1 - 0.10) = 0.19$, $P(Y = 1|G = 0, E = 1) = P(A_0 = 1 \text{ or } A_2 = 1) = 1 - (1 - 0.10)(1 - 0.10) = 0.19$, and $P(Y = 1|G = 1, E = 1) = P(A_0 = 1 \text{ or } A_1 = 1 \text{ or } A_2 = 1) = 1 - (1 - 0.10)(1 - 0.10)(1 - 0.10) = 0.271$.

If we use p_{ge} to denote $P(Y = 1|G = g, E = e)$, then our measure of additive interaction is $p_{11} - p_{10} - p_{01} + p_{00} = -0.009 < 0$; we have negative additive interaction. Our measure of multiplicative interaction on the risk ratio scale is $p_{11}p_{00}/(p_{10}p_{01}) = 0.7507 < 1$; we have negative multiplicative interaction on the

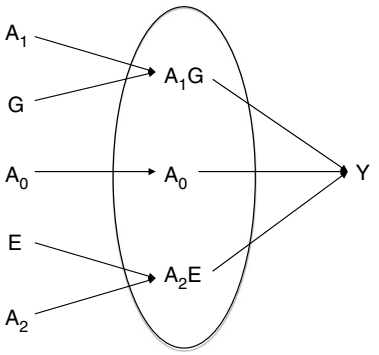

Figure 10.3 Example illustrating negative multiplicative interaction but no mechanistic interaction.

Table 10-1. EXAMPLE OF NEGATIVE STATISTICAL INTERACTION BUT NO SYNERGISM

G	*E*	Number of Individual	Cases	Risk
$G=0$	$E=0$	4000	400	0.10
$G=1$	$E=0$	1000	190	0.19
$G=0$	$E=1$	4000	760	0.19
$G-1$	$E=1$	1000	271	0.271

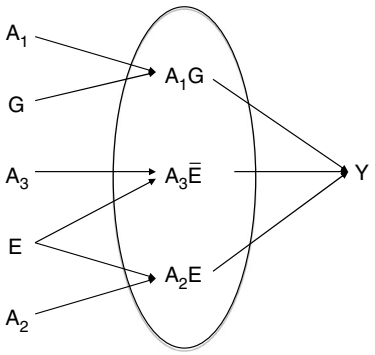

Figure 10.4 Example illustrating positive multiplicative and positive additive interaction but no mechanistic interaction.

risk ratio scale. Our measure of multiplicative interaction on the odds ratio scale is $(\frac{p_{11}}{1-p_{11}}/\frac{p_{10}}{1-p_{10}})/(\frac{p_{01}}{1-p_{01}}/\frac{p_{00}}{1-p_{00}}) = 0.7507 < 1$; we have negative multiplicative interaction on the odds ratio scale. We thus have both negative additive interaction and negative multiplicative interaction; but there is no synergism between G and E in this example, no mechanism that requires both G and E to operate, even though we have statistical interaction on all of our standard scales. Clearly negative statistical interaction does not necessarily allow us to draw conclusions about synergism.

Consider now the example in Figure 10.4. Suppose there are again three causal mechanisms for the outcome Y, one involving G and some other factors A_1, one involving E and some other factors A_2, and one involving the absence of E, which we denote by $\overline{E}$, and some other factors A_3. Note that no mechanism requires both G and E; that is, there is no mechanistic interaction between G and E. Suppose again that the distributions of genetic factor G and the environmental factor E are statistically independent in the population. Let the probability of G and E be 0.4 and 0.5, respectively. Finally, suppose that the distributions of A_1, A_2, and A_3 are independent of G and E and the joint distribution of A_1, A_2, and A_3 is as given in Table 10.2.

It can then be shown that a cohort of 10,000 can be cross-classified into cases and non-cases by genetic and environmental exposure status as given in Table 10.3.

Table 10-2. JOINT PROBABILITIES OF $A_1, A_2,$ AND A_3

$$P(A_1 = 0, A_2 = 1, A_3 = 1) = 0.004$$
$$P(A_1 = 1, A_2 = 0, A_3 = 0) = 0.004$$
$$P(A_1 = 0, A_2 = 1, A_3 = 0) = 0.001$$
$$P(A_1 = 1, A_2 = 1, A_3 = 0) = 0.001$$
$$P(A_1 = 0, A_2 = 0, A_3 = 0) = 0.99$$

Table 10-3. EXAMPLE OF POSITIVE STATISTICAL INTERACTION BUT NO SYNERGISM

G	***E***	**Number of Individual**	**Cases**	**Risk**
$G = 0$	$E = 0$	3000	12	0.004
$G = 1$	$E = 0$	2000	10	0.005
$G = 0$	$E = 1$	3000	18	0.006
$G - 1$	$E = 1$	2000	20	0.010

Our measure of additive interaction is $p_{11} - p_{10} - p_{01} + p_{00} = 0.003 > 0$; we have positive additive interaction. Our measure of multiplicative interaction on the risk ratio scale is $p_{11}p_{00}/(p_{10}p_{01}) = 1.33 > 1$; we have positive multiplicative interaction on the risk ratio scale. Our measure of multiplicative interaction on the odds ratio scale is $(\frac{p_{11}}{1-p_{11}}/\frac{p_{10}}{1-p_{10}})/(\frac{p_{01}}{1-p_{01}}/\frac{p_{00}}{1-p_{00}}) = 1.34 > 1$; we have positive multiplicative interaction on the odds ratio scale. We thus have both positive additive interaction and positive multiplicative interaction, but once again there is no synergism between G and E in this example—that is, no mechanism that requires both G and E to operate—even though we have positive statistical interaction on all of our standard scales. We thus see that not even positive statistical interaction on the additive or multiplicative scale necessarily allows us to draw conclusions about synergism.

We might then wonder whether we can ever draw conclusions about synergism from simply examining statistical interaction.

10.3. EMPIRICAL TESTS FOR SUFFICIENT CAUSE SYNERGISM

Starting with the work of Miettinen (1982) and Greenland and Poole (1988) there has been effort to relate sufficient causes to the counterfactual framework. Following Miettinen (1982), Greenland and Poole (1988) considered a setting with two binary causes, G and E, and a binary outcome Y and considered the potential responses ("counterfactual outcomes") for individuals under different combinations of the exposures so that Y_{ge} denotes the potential outcome that would have occurred had G been set to g and E to e. If G and E are binary, then, because there

Table 10-4. ENUMERATION OF RESPONSE PATTERNS TO FOUR POSSIBLE EXPOSURE COMBINATIONS ACCORDING TO AN INDIVIDUAL'S POTENTIAL OUTCOMES

Type	Y_{11}	Y_{01}	Y_{10}	Y_{00}
1	1	1	1	1
2	1	1	1	0
3	1	1	0	1
4	1	1	0	0
5	1	0	1	1
6	1	0	1	0
7	1	0	0	1
8	1	0	0	0
9	0	1	1	1
10	0	1	1	0
11	0	1	0	1
12	0	1	0	0
13	0	0	1	1
14	0	0	1	0
15	0	0	0	1
16	0	0	0	0

are four possible exposure combinations, each individual has four potential outcomes: Y_{11}, Y_{10}, Y_{01}, and Y_{00}, each of which may be either 0 or 1. Individuals can be classified into one of 16 different types as in Table 10.4.

The four potential outcomes defined what was called an individual's "response type"; because there were four different potential outcomes, there were $2^4 = 16$ different possible response types. For example, response type 8 indicates an individual for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$; similarly, response type 10 indicates an individual for whom $Y_{10} = Y_{01} = 1$ and $Y_{11} = Y_{00} = 0$ and similarly for the other response types.

Within this setting of two binary causes of interest, Greenland and Poole (1988) also noted that one could conceive of nine different possible sufficient causes U_0, U_1G, $U_2\overline{G}$, U_3E, $U_4\overline{E}$, U_5GE, $U_6\overline{G}E$, $U_7G\overline{E}$, and $U_8\overline{G}\,\overline{E}$, each involving the presence of G, or the absence of G (denoted by $\overline{G}$) or being unrelated to G, and likewise each involving the presence of E, or the absence of E (denoted by $\overline{E}$) or being unrelated to E, and finally each also possibly involving an additional background cause U_j. If there were a sufficient cause U_5GE, then we would say that there was synergism between G and E. If there were sufficient causes $U_6\overline{G}E$ or $U_7G\overline{E}$, we would say that there was antagonism between G and E. Greenland and Poole noted that the presence or absence of different sufficient causes (indicated by $U_j = 1$ or $U_j = 0$) would give rise to different response types (cf. Greenland and Brumback, 2002).

In fact, the notion of synergism in the sufficient cause framework is very close related to response types. It can be shown (VanderWeele and Robins, 2008; cf. Appendix) that if there is any individual for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$ (i.e.,

any individual of response type 7 or 8), then there must be a sufficient cause U_5GE; that is, there must be synergism between G and E. Stated another way, the only way we can get can a response pattern that has $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$ is by having a sufficient cause U_5GE.

In Chapter 9 we called individuals for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$—that is, for whom the outcome would occur if both exposures were present but not if just one or the other were present—a mechanistic interaction or a sufficient cause interaction. It was noted there that, provided we had controlled for confounding of the effects of both exposures on the outcome, we could empirically test for such mechanistic interaction.

It was noted in Chapter 9 (cf. VanderWeele and Robins, 2007b, 2008) that if the effect of the two exposures were unconfounded, then

$$p_{11} - p_{10} - p_{01} > 0$$

would imply the presence of a sufficient cause interaction. Recall that this is a stronger condition than regular positive additive interaction which only requires $p_{11} - p_{10} - p_{01} + p_{00} > 0$. The condition $p_{11} - p_{10} - p_{01} > 0$ is stronger because with the condition $p_{11} - p_{10} - p_{01} > 0$ we are no longer adding back in the outcome probability p_{00} for the doubly unexposed group. This condition for a sufficient cause interaction thus does not correspond to, and is stronger than, the regular test for additive interaction (VanderWeele, 2009d). Indeed we saw in the example in Figure 10.4 and Table 10.3 that we could have positive additive interaction with no synergism present. Again, in general, we need the more stringent condition $p_{11} - p_{10} - p_{01} > 0$ to be satisfied to conclude the presence of mechanistic interaction or sufficient cause synergism.

It was also noted in the last chapter that if the effects of both exposures are positive "monotonic" in the sense that the counterfactuals Y_{ge} are nondecreasing in g and e for all individuals (i.e., the exposures never have protective effects on the outcome for any individual), then the standard test for additive interaction $p_{11} - p_{10} - p_{01} + p_{00} > 0$ can be used to test for sufficient cause interaction (VanderWeele and Robins, 2007b, 2008). Expressed in terms of sufficient causes, the monotonicity assumption is essentially that there are no sufficient causes that involve $\overline{G}$ or $\overline{E}$ (i.e., no sufficient causes that require the absence of G or the absence of E). More precisely, if there are no sufficient causes that involve $\overline{G}$ or $\overline{E}$ (i.e., no sufficient causes that require the absence of G or the absence of E), then Y_{ge} must be nondecreasing in g and e. If this monotonicity assumption holds for both exposures, we can test for synergism in the sufficient cause sense by simply testing for positive additive interaction. Note that although there was positive additive interaction in the example in Figure 10.4 and Table 10.3, the monotonicity assumption did not hold (there was a sufficient cause that involved $\overline{E}$, i.e. the absence of E) and thus positive additive interaction was no longer a sufficient condition for synergism and one would need the stronger condition $p_{11} - p_{10} - p_{01} > 0$. Note that these conditions given here are sufficient but not necessary for synergism in the sufficient cause framework; that is, if these conditions are satisfied, then a sufficient cause

interaction must be present, but if the conditions are not satisfied, then there may or may not be a sufficient cause interaction—one simply cannot determine this from the data.

10.4. SUFFICIENT CAUSE INTERACTION AND STATISTICAL INTERACTIONS

The conditions above for sufficient cause interaction can be expressed in terms of various forms of statistical interaction. However, as we will see, without making monotonicity assumption, standard tests for statistical interaction will not suffice to conclude the presence of a sufficient cause interaction; stronger conditions will be needed.

We will begin with the relative excess risk due to interaction. If we define the relative risk by $RR_{ge} = \frac{p_{ge}}{p_{00}}$, then the relative excess risk due to interaction using risk ratios is defined by $RERI_{RR} = RR_{11} - RR_{10} - RR_{01} + 1 = \frac{p_{11}}{p_{00}} - \frac{p_{10}}{p_{00}} - \frac{p_{01}}{p_{00}} + 1$. It was noted in Chapter 9 that the condition $p_{11} - p_{10} - p_{01} > 0$ expressed in terms of the relative excess risk due to interaction $RERI_{RR}$ is equivalent to $RERI_{RR} > 1$. The condition for standard positive additive interaction, $p_{11} - p_{10} - p_{01} + p_{00} > 0$, can likewise be expressed as $RERI_{RR} > 0$. The conditions $RERI_{RR} > 0$ and $RERI_{RR} > 1$ essentially provide thresholds, with and without monotonicity assumptions respectively, such that if the relative excess risk due to interaction exceeds the threshold (and if we have controlled for confounding of both exposures), then synergism in the sufficient cause sense must be present.

Similar thresholds can be established for multiplicative interaction. In particular, it can be shown (VanderWeele, 2009d) that with monotonicity for both exposures a multiplicative interaction for risk ratios, $\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}p_{00}}{p_{10}p_{01}}$, that exceeds 1 suffices for synergism. Without such monotonicity assumptions, a multiplicative interaction that exceeds 2 is sufficient for synergism, provided that both the main effect of G and E on Y are on average non-negative. Note that in the example in Figure 10.4 and Tables 10.2 and 10.3, we had a multiplicative interaction of 1.33 that exceeded 1, but in this example the monotonicity assumption was violated (since E could be preventive, i.e., there was a sufficient cause that involved $\overline{E}$) and thus we would have needed to use the threshold of 2. The multiplicative interaction in this example of 1.33 did not exceed the threshold of 2. The multiplicative interaction threshold of 2 essentially ensures that examples such as that in Figure 10.4 and Table 10.3 (with no synergism, but nonzero multiplicative interaction) cannot arise; if the multiplicative interaction exceeds 2 and the main effects of G and E are non-negative, then there must be synergism, irrespective of whether the individual-level monotonicity holds or not. Note that while the $RERI_{RR}$ thresholds of $RERI_{RR} > 0$ and $RERI_{RR} > 1$ are equivalent to the conditions $p_{11} - p_{10} - p_{01} + p_{00} > 0$ and $p_{11} - p_{10} - p_{01} > 0$, respectively, the thresholds for the multiplicative interaction imply (but are not implied by, i.e. they are stronger than) the conditions on the additive interaction $p_{11} - p_{10} - p_{01} + p_{00} > 0$ and $p_{11} - p_{10} - p_{01} > 0$. In testing for synergism, it is thus

best whenever possible to use the additive interaction conditions or $RERI_{RR} > 0$ and $RERI_{RR} > 1$. Thus when assessing additive interaction using $RERI_{RR}$, it is useful to examine not only whether the estimate and confidence interval for $RERI_{RR}$ are greater than 0 (i.e., whether there is additive interaction), but also whether the estimate and confidence interval for $RERI_{RR}$ are all greater than 1 since this magnitude would provide evidence for sufficient cause without the need for additional assumptions.

These conditions can likewise be expressed in terms of the coefficients of the statistical models considered in Chapter 9. If we were to fit an additive model of the form

$$P(Y = 1|G = g, E = e) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg$$

then the condition for sufficient cause synergism under monotonicity, $p_{11} - p_{10} - p_{01} + p_{00} > 0$, can be expressed in terms of the model parameters as $\alpha_3 > 0$—that is, a positive statistical interaction on the additive scale. The condition for sufficient cause synergism without monotonicity assumptions, $p_{11} - p_{10} - p_{01} > 0$, can be expressed in terms of the model parameters as $\alpha_3 > \alpha_0$. We see that only under monotonicity assumptions does a statistical interaction on the additive scale, $\alpha_3 > 0$, correspond to a sufficient cause interaction, and then only a positive statistical interaction. Otherwise, without monotonicity we must use the more stringent condition, $\alpha_3 > \alpha_0$.

Likewise, a log-linear model for risk ratios that includes a product term takes the form

$$\log\{P(Y = 1|G = g, E = e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg$$

Here, if G and E have positive monotonic effects on Y then the condition $\beta_3 > 0$ implies a sufficient cause interaction (VanderWeele, 2009d). Thus, with monotonicity, we have that a positive statistical interaction on the multiplicative scale implies a sufficient cause interaction. Without monotonicity, provided that both the main effect of G and E on Y are non-negative (i.e., $\beta_1 \geq 0$ and $\beta_2 \geq 0$), the condition $\beta_3 > \log(2)$ implies a sufficient cause interaction (VanderWeele, 2009d). Since $e^{\beta_3} = RR_{11}/(RR_{10}RR_{01})$, this is just equivalent to the condition $RR_{11}/(RR_{10}RR_{01}) > 2$ described above. Thus, once again, without monotonicity assumptions a positive statistical multiplicative interaction, $\beta_3 > 0$, alone does not suffice. Without monotonicity assumptions we need the more stringent condition $\beta_3 > \log(2)$. As above, however, if we can estimate the parameters of the multiplicative model $\beta_1, \beta_2, \beta_3$, then, as described in Chapter 9, we can calculate the relative excess risk due to interaction by $RERI_{RR} = e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1$ and we would be better off testing for sufficient cause synergism using the conditions $RERI_{RR} > 0$ and $RERI_{RR} > 1$ with and without monotonicity respectively, as these conditions are more often satisfied than those for the multiplicative interaction ($\beta_3 > 0$ and $\beta_3 > \log(2)$). The comments here for statistical interaction for risk ratios in a log-linear model pertain also approximately to statistical interaction for odds ratios in a logistic regression model when the outcome is rare.

Example: Infectious Disease Epidemiology. It was noted in Chapter 9 that an analysis by Bhavnani et al. (2012), using age-standardized measures, showed that there was evidence of additive interaction between two different pathogens (rotavirus and *Giardia*) in their effects on diarrheal disease. Bhavnani et al. (2012) report that risk ratios for diarrheal disease across groups infected with rotavirus and/or *Giardia*. With the doubly unexposed group as the reference category, the risk ratio for rotavirus (in the absence of *Giardia*) is 2.63, the risk ratio for *Giardia* (in the absence of rotavirus) is 1.13, and the risk ratio when both rotavirus and *Giardia* are present is 10.72. This gives an $RERI_{RR}$ of $RERI_{RR} = 10.72 - 2.63 - 1.13 + 1 = 7.96$ (95% CI: 3.13, 18.92). The value of $RERI_{RR}$ and its entire 95% confidence interval exceed the value 1, suggesting strong evidence for synergism in the sufficient cause sense—that is, for the presence of a sufficient cause that requires both rotavirus and *Giardia* to operate.

10.5. ''EPISTATIC'' OR SINGULAR INTERACTIONS

In this section we will consider another notion of "causal" or "mechanistic" interaction, which is related to, but as it turns out stronger than, the "sufficient cause interactions" in the previous sections. This even stronger notion of mechanistic interaction we will refer to as an "epistatic" or "singular" interaction for reasons described below. Some of the motivation for this stronger notion of interaction comes from concepts in the genetics literature. Often the term "epistasis" is used in the genetics literature to describe statistical interaction between two genetic factors. Cordell (2002, 2009) has noted that although the word "epistasis" is now essentially used simply to describe a statistical gene–gene interaction, the word originally had a somewhat different sense. Writing in 1909, Bateson used the term "epistasis" to describe instances in which the effect of a particular genetic variant was masked by a variant at another locus so that variation of phenotype with genotype at one locus was only apparent amongst those with certain genotypes at the second locus (Bateson, 1909). Cordell (2002, 2009) argued that the statistical tests that are often used to assess interactions are of limited use in elucidating the type of interaction that Bateson had originally conceived.

Here we will use G_1 and G_2 to denote our two genetic exposures. We will assume for now that these are both binary. However, we will discuss extensions to variables with more than two levels below (since genetic factors are often coded as variables with three levels indicating 0, 1, or 2 variant alleles). Under Bateson's original conception, epistasis would be said to be present if variation of the outcome (often referred to in the genetics literature as the phenotype) with the genetic factor at one locus, G_1, was only apparent amongst those with certain values of the genetic factor at the second locus ($G_2 = 1$). Those with other values at the second locus (e.g., $G_2 = 0$) would show no effect at the first. Suppose then we think of interaction this way (different from the statistical interaction) and ask the question whether there are any individuals for whom the first genetic factor G_1 has no effect on the outcome unless the second genetic factor G_2 is present (e.g., $G_2 = 1$) as in Table 10.5. Table 10.5 describes an outcome pattern for a particular individual such that the

Table 10-5. Example of a Table of Phenotypes for a Particular Individual for the Effects of Different Genotypes at Two Loci Exhibiting Epistasis Under Bateson's (1909) Original Definition

	$G_2 = 0$	$G_2 = 1$
$G_1 = 0$	0	0
$G_1 = 1$	0	1

effect of the first genetic factor G_1 is only present when $G_2 = 1$. If $G_2 = 0$, then $D = 0$ irrespective of whether $G_1 = 0$ or $G_1 = 1$; that is, if $G_2 = 0$, then it looks as though G_1 has no effect on the outcome. However, when $G_2 = 1$, then $D = 1$ when $G_1 = 1$ but $D = 0$ when $G_1 = 0$; thus when $G_2 = 1$ it is clear that G_1 has an effect on the outcome. Essentially then if $G_2 = 0$, the effect of G_1 is masked. This is what Bateson had in view when he introduced the term "epistasis." In this setting we would then say that G_1 is epistatic to G_2. By symmetry in this example, it is also true that the effect of G_2 is only present when $G_1 = 1$ so we would also say that G_2 is epistatic to G_1.

If we use our counterfactual notation, then the response pattern in Table 10.5 is simply an individual for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$. Although this is a conceptually distinct notion from that of a statistical interaction, current terminology in genetics does not distinguish between "statistical gene–gene interaction" and "epistasis" in the sense of Bateson (1909). Cordell (2009) notes that Fisher (1918) used the term "epistacy" for statistical gene–gene interaction, distinguishing it from Bateson's "epistasis." However, the two terms sound very similar and, with time, "epistasis" came to be used synonymously with "epistacy," both indicating only "statistical interaction" between genetic factors. With the greater recognition that there are two distinct concepts, Phillips (2008) proposed using "statistical epistasis" for statistical interaction between two genetic factors and "compositional epistasis" for epistasis in the sense of masking and the terminology has been adopted by others (Cordell, 2009; Moore and Williams, 2009). In what follows we will call this response pattern with $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$ either "compositional epistasis" or an "epistatic interaction." Note that this is somewhat stronger than what we defined above as a sufficient cause interaction that required only $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$. For an "epistatic interaction" we also require that the outcome be 0 when both the exposures are absent, that is, $Y_{00} = 0$. With a sufficient cause interaction we do not require any particular value of Y_{00}; knowing that there are individuals for whom $Y_{11} = 1$ and $Y_{10} = Y_{01}$ suffices to know that there must be a sufficient cause with both G and E; we do not need to know Y_{00}.

Some of the background we have been describing here arises from the genetics literature response pattern with $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$ (i.e., the outcome occurs if and only if both of two exposures are present) may of course be of interest outside of the genetic context. We may be interested in such response patterns between a genetic exposure and an environmental exposure, or between two

environmental exposures, or between two behavioral exposures, and so on. Because such response patterns may be of interest outside of the genetic context, we will also refer to such response patterns with $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$ in these various settings as a "singular interaction."

Cordell (2002, 2009) pointed out that tests for statistical interactions ("statistical epistasis") will generally be of limited use in drawing conclusions about epistasis in the sense of masking ("compositional epistasis") as Bateson had originally conceived of it and gave examples with a statistical interaction but no compositional epistasis. Although tests for ordinary statistical interaction between two genetic factors do not in general allow one to draw conclusions about epistasis, progress can be made in empirically testing for such compositional epistasis or epistatic interaction by using empirical conditions that often differ from standard statistical interaction (VanderWeele, 2010b, c). As was mentioned briefly in Chapter 9, if the effect of the two exposures were unconfounded, then an "epistatic" or "singular" interaction must be present if

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

This is a stronger condition than regular positive additive interaction which only requires $p_{11} - p_{10} - p_{01} + p_{00} > 0$, and it is a stronger condition than even that required for sufficient cause interaction without monotonicity assumptions, which was simply that $p_{11} - p_{10} - p_{01} > 0$. Here we are now subtracting p_{00} in the probability contrast. However, if the effect of both exposures are unconfounded (e.g., in the genetics context if control is made for population stratification), then if we have $p_{11} - p_{10} - p_{01} - p_{00} > 0$, then an epistatic interaction must be present.

As with sufficient cause interactions, somewhat weaker conditions suffice to detect epistatic interaction if we make certain monotonicity assumptions. Specifically, if we can assume that at least one of the two exposures has a positive monotonic effect on the outcome, then $p_{11} - p_{10} - p_{01} > 0$ suffices. If we can assume that both exposures have positive monotonic effects on the outcome, then by simply using the standard condition for the positive additive interaction, we obtain $p_{11} - p_{10} - p_{01} + p_{00} > 0$, an epistatic interaction. Note that these conditions just given are sufficient but not necessary for an epistatic interaction; that is, if these conditions are satisfied, then an epistatic interaction must be present, but if the conditions are not satisfied, then there may or may not be an epistatic interaction—one simply cannot determine this from the data.

These conditions can likewise be expressed using the relative excess risk due to interaction, again assuming that control has been made for confounding. Without any monotonicity assumptions, we can test for an epistatic interaction by testing $RERI_{RR} > 2$. If we assume that at least one of the two exposures has a positive monotonic effect on the outcome, then $RERI_{RR} > 1$ suffices. If we assume that both exposures have positive monotonic effects on the outcome, then $RERI_{RR} > 0$ suffices.

Sufficient conditions for epistatic interaction also hold related to the magnitude of multiplicative interaction if control has been made for confounding. Specifically, provided that the main effects of both exposures are on average non-negative, then without any monotonicity assumptions, a multiplicative interaction, $\frac{RR_{11}}{RR_{10}RR_{01}}$,

Table 10-6. ODDS RATIOS FOR ESOPHAGEAL CANCER FOR THE EFFECTS OF ARG VARIANTS ON ADH2, G_1, AND GLU/GLU VERSUS GLU/LYS ON ALDH2, G_2

	$G_2 = 0$	$G_2 = 1$
$G_1 = 0$	1	3.52
$G_1 = 1$	1.40	7.20

greater than 3 suffices for an epistatic interaction. If we assume that at least one of the two exposures has a positive monotonic effect on the outcome, then $\frac{RR_{11}}{RR_{10}RR_{01}} > 2$ suffices. If we assume that both exposures have positive monotonic effects on the outcome, then $\frac{RR_{11}}{RR_{10}RR_{01}} > 1$ suffices. Expressed in terms of the coefficients of a log-linear model given in Section 10.4 (or a logistic model with rare outcome), these conditions are $\beta_3 > \log(3)$ without monotonicity, $\beta_3 > \log(2)$ if one exposure has a positive monotonic effect, and $\beta_3 > 0$ if both exposures have positive monotonic effects. However, as in Section 10.4, if we can estimate the parameters of the multiplicative model, then we can estimate the relative excess risk due to interaction and we would be better off testing for sufficient cause synergism using the conditions $RERI_{RR} > 2$, $RERI_{RR} > 1$, or $RERI_{RR} > 0$ because these are weaker conditions than those on the multiplicative scale.

Example: Genetic Epidemiology. Yang et al. (2005) examine interaction between two genetic variants in their effects on esophageal cancer. Specifically, they examine the effects of Arg variants on ADH2 (on chromosome 4), which will be our first exposure, G_1; and they studied Glu/Glu versus Glu/Lys on ALDH2 (on chromosome 12), which will be our second exposure, G_2. Odds ratios for esophageal cancer using the case–control data from Yang et al. (2005) are given in Table 10.6.

The outcome is relatively rare, and using these odds ratio we can thus estimate the relative excess risk due to interaction (and calculate the confidence interval using the data provided by Yang et al. (2005): $RERI_{OR} = OR_{11} - OR_{10} - OR_{01} + 1 = 7.20 - 1.40 - 3.52 + 1 = 3.28$ (95% CI: 0.4, 6.16). The estimate $RERI_{OR} = 3.28 > 2$ would suggest compositional epistasis without any assumptions about monotonicity at all. However, the confidence interval contains a value as small as $RERI_{OR} = 0.4 > 0$, which would imply compositional epistasis only if both variants had monotonic effects on esophageal cancer.

10.6. EXTENSIONS TO ORDINAL EXPOSURES

The ideas above concerning epistatic interaction, and also concerning sufficient cause interaction, can in fact be extended to settings with exposures that have more than two levels. We will consider in this section some extensions for epistatic interactions in which the exposures have three levels. This would be perhaps the most

common setting in genetics in which G_1 and G_2 have three levels corresponding to 0, 1, or 2 variant alleles. Similar approaches are also possible for sufficient cause interactions (rather than epistatic interaction) for a setting in which one or both of the exposures has three levels, but the interpretation of these is somewhat more complicated and the interested reader is referred to VanderWeele (2010e) for details. Once again, we will let Y be a binary outcome for each individual in the population and let Y_{ij} denote the outcome that would have occurred for that individual if G_1 had been i and G_2 had been j. We again say that G_1 has a positive monotonic effect on Y if Y_{ij} is nondecreasing in i for all individuals, and we say G_2 has a positive monotonic effect on Y if Y_{ij} is nondecreasing in j for all individuals. Consider the response pattern in Table 10.7.

Note that for the response pattern given in Table 10.7, the effect of G_1 is only apparent when $G_2 = 2$. Thus the response pattern in Table 10.6 would be another instance that would be considered epistasis under Bateson's original conception. We can once again test for epistasis empirically. Let $p_{ij} = P(Y = 1 | G_1 = i, G_2 = j)$. We will suppose throughout the remainder of this section that the effects of G_1 and G_2 on Y are unconfounded. If the effects of G_1 and G_2 on Y are unconfounded conditional on some covariates C, then we will assume that all probabilities are conditional on C. It can be shown that if both G_1 and G_2 have positive monotonic effects on Y, then

$$p_{22} - p_{21} - p_{12} + p_{11} > 0$$

implies the existence of individuals with response type in Table 10.7 (VanderWeele, 2010c). If only G_1 has a positive monotonic effect on Y, then $p_{22} - p_{21} - p_{20} - p_{12} > 0$ suffices. If only G_2 has a positive monotonic effect on D, then $p_{22} - p_{12} - p_{02} - p_{21} > 0$ suffices. Without any monotonicity assumptions, $p_{22} - p_{21} - p_{20} - p_{12} - p_{11} - p_{10} - p_{02} - p_{01} - p_{00} > 0$ suffices.

A number of other response patterns constituting instances of epistasis are also possible. We might more generally say that there is compositional epistasis between G_1 and G_2 in that G_1 is epistatic to G_2, if there exists some individual for whom there is some level k of G_2 such that there is no effect of G_1 on Y when $G_2 = k$ and some other level l of G_2 such that there is an effect of G_1 on Y when $G_2 = l$. In counterfactual notation, this would be some level k of G_2 such Y_{ik} is constant in i and some level l of G_2 such that Y_{il} is not constant in i.

Consider the response type pattern in Table 10.8. Here G_1 has no effect on Y if $G_2 = 1$ (or if $G_2 = 0$) but G_1 does have an effect on Y if $G_2 = 2$, so here we again have compositional epistasis. It can be shown that if both G_1 and G_2 have positive monotonic effects on Y, then

$$p_{12} - p_{21} - p_{02} + p_{01} > 0$$

implies the existence of individuals with response type in Table 10.7 (VanderWeele, 2010c). If only G_1 has a positive monotonic effect on Y, then $p_{12} - p_{21} - p_{20} - p_{02} > 0$ suffices. Tests of this form cannot in general be used to test for epistatic response patterns like those in Table 10.8 if it can only be assumed that G_2 has a positive monotonic effect on Y or if no assumptions are made about monotonicity.

Table 10-7. Example of a Table of Epistatic Interaction when Exposures Have Three Levels

	$G_2 = 0$	$G_2 = 1$	$G_2 = 2$
$G_1 = 0$	0	0	0
$G_1 = 1$	0	0	0
$G_1 = 2$	0	0	1

Table 10-8. Example of a Table of Epistatic Interaction when Exposures Have Three Levels

	$G_2 = 0$	$G_2 = 1$	$G_2 = 2$
$G_1 = 0$	0	0	0
$G_1 = 1$	0	0	1
$G_1 = 2$	0	0	1

Table 10-9. Example of a Table of Epistatic Interaction when Exposures Have Three Levels

	$G_2 = 0$	$G_2 = 1$	$G_2 = 2$
$G_1 = 0$	0	0	0
$G_1 = 1$	0	0	0
$G_1 = 2$	0	1	1

Consider the response type pattern in Table 10.8. If both G_1 and G_2 have positive monotonic effects on Y, then

$$p_{21} - p_{12} - p_{20} + p_{10} > 0$$

implies the existence of individuals with response type in Table 10.9 (VanderWeele, 2010c). If only G_2 has a positive monotonic effect on Y, then $p_{21} - p_{12} - p_{02} - p_{20}$ suffices. Tests of this form cannot in general be used to test for epistatic response patterns like those in Table 10.9 if it can only be assumed that G_1 has a positive monotonic effect on Y or if no assumptions are made about monotonicity.

Finally consider the response type pattern in Table 10.10.

If both G_1 and G_2 have positive monotonic effects on Y, then

$$p_{11} - p_{20} - p_{02} + p_{00} > 0$$

implies the existence of individuals with response type in Table 10.9 (VanderWeele, 2010c). Tests of this form cannot in general be used to test for epistatic response patterns like those in Table 10.10 if it can only be assumed that just one of G_1 or

has G_2 has a positive monotonic effect on Y or if no assumptions are made about monotonicity.

Similar results hold if G_1 has two levels and G_2 has three levels or to other epistatic response patterns (VanderWeele, 2010c; Suzuki and VanderWeele, 2014). Epistasis, in the sense of masking, would be present if there were individuals for whom $Y_{12} = 1$ but $Y_{11} = Y_{10} = Y_{02} = Y_{01} = Y_{00} = 0$. If the effects of G_1 and G_2 on Y are both monotonic, then there are individuals with the epistatic response pattern above if $p_{12} - p_{11} - p_{02} + p_{01} > 0$. If only the effect G_1 on Y can be assumed to be monotonic, then there are individuals with the epistatic response pattern above if $p_{12} - p_{11} - p_{10} - p_{02} > 0$. If only the effect G_2 on Y can be assumed to be monotonic, then there are individuals with the epistatic response pattern above if $p_{12} - p_{11} - p_{02} > 0$. If neither the effect of G_1 or G_2 can be assumed to be monotonic, then there are individuals with the epistatic response pattern above if $p_{12} - p_{11} - p_{10} - p_{02} - p_{01} - p_{00} > 0$.

Epistasis, in the sense of masking, would also be said to be present if there were individuals for whom $Y_{12} = Y_{11} = 1$ but $Y_{10} = Y_{02} = Y_{01} = Y_{00} = 0$. If the effects of G_1 and G_2 on Y are both monotonic, then there are individuals with the epistatic response pattern above if $p_{11} - p_{10} - p_{02} + p_{00} > 0$. If the effects of just G_2 on Y is monotonic, then there are individuals with the epistatic response pattern above if $p_{11} - p_{10} - p_{02} > 0$. Tests of this form cannot in general be used to test for this epistatic response pattern if only the effect G_1 on Y or if neither the effect of G_1 or G_2 can be assumed to be monotonic.

In all of these settings, control for covariates could be made by logistic regression and statistical inference and testing could be done by an analogue to the relative excess risk due to interaction approach using the delta method as in Sections 9.3 and 9.4, but this would require deriving a new formula for the standard error for each of the various empirical tests above. Alternatively, bootstrapping could be used for standard errors and confidence intervals. As another alternative, a weighting approach for estimating these contrasts is possible for either cohort data (VanderWeele et al., 2010) or case–control data (VanderWeele and Vansteelandt, 2011), which will essentially calculate the standard errors for these contrasts automatically.

As noted above, similar approaches are also possible for sufficient cause interactions (rather than epistatic interaction) for a setting in which one or both of the exposures have three levels, but the interpretation of these is somewhat more complicated and the interested reader is referred to VanderWeele (2010e) for details.

10.7. EXTENSIONS TO THREE OR MORE EXPOSURES

Notions of sufficient cause and "singular" or "epistatic" interaction can likewise be extended to settings in which there are three or more exposures (VanderWeele and Richardson, 2012; Ramasahai, 2013). For example, for three binary exposures of interest, G_1, G_2, and G_3 say, let $Y_{g_1g_2g_3}$ denote be the counterfactual outcome Y for an individual if G_1, G_2, and G_3 had been set to g_1, g_2, and g_3, respectively. We say that there is a three-way sufficient cause interaction between G_1, G_2,

and G_3 if for some individual, $Y_{111} = 1$ but $Y_{110} = Y_{101} = Y_{011} = 0$. It can be shown that this implies, in the sufficient cause framework, a sufficient cause that requires G_1, G_2, and G_3 to operate (VanderWeele and Richardson, 2012). If we let $p_{g_1g_2g_3} = \mathbb{E}(Y|G_1 = g_1, G_2 = g_2, G_3 = g_3)$ and if the effects of G_1, G_2, and G_3 on Y are unconfounded, then

$$p_{111} - p_{110} - p_{101} - p_{011} > 0$$

implies a three-way sufficient cause interaction. If $Y_{g_1g_2g_3}$ is nondecreasing in g_1, g_2, and g_3 (i.e., if all three exposures have positive monotonic effects), then any of the following three conditions implies a three-way sufficient cause interaction:

$$p_{111} - p_{110} - p_{101} - p_{011} + p_{100} + p_{010} > 0$$
$$p_{111} - p_{110} - p_{101} - p_{011} + p_{100} + p_{001} > 0$$
$$p_{111} - p_{110} - p_{101} - p_{011} + p_{010} + p_{001} > 0$$

If just two of the three exposures have positive monotonic effects, $G_1$and G_2 say, then

$$p_{111} - p_{110} - p_{101} - p_{011} + p_{001} > 0$$

implies a three-way sufficient cause interaction. Analogous conditions could be formed if some other combination of two exposures had positive monotonic effects on the outcome. If only one of the exposures has a positive monotonic effect, then the same condition as in the setting of no monotonicity assumptions, namely $p_{111} - p_{110} - p_{101} - p_{011} > 0$, must be employed. As before, if the effects of $\{G_1, G_2, G_3\}$ on D are unconfounded conditional on some set of covariates C, then these conditions can be made conditional on C.

We could similarly define a singular interaction between G_1, G_2, and G_3 to be present if there were individuals for whom $Y_{111} = 1$ but $Y_{110} = Y_{101} = Y_{011} = Y_{100} = Y_{010} = Y_{001} = Y_{000} = 0$—that is, for whom the outcome occurred if and only if $G_1 = G_2 = G_3 = 1$. If all three of the exposures have positive monotonic effects on the outcomes, then the same tests for a three-way sufficient cause interaction when all three exposures have positive monotonic effects can be used. Likewise, if two of the three exposures have positive monotonic effects on the outcome, then the same tests for a three-way sufficient cause interaction when two of the three exposures have positive monotonic effects can be used. However, tests for singular interactions diverge from sufficient cause interactions if just one or if none of the exposures have a positive monotonic effect. Specifically, suppose only G_1 has a positive monotonic effect on the outcome, then the following condition suffices for a singular interaction between G_1, G_2, and G_3 (VanderWeele and Richardson, 2012):

$$p_{111} - p_{110} - p_{101} - p_{011} - p_{100} > 0$$

Without any monotonicity assumptions,

$$p_{111} - p_{110} - p_{101} - p_{011} - p_{100} - p_{010} - p_{001} - p_{000} > 0$$

would suffice for a singular interaction between G_1, G_2, and G_3.

Conditions for four-way and even n-way sufficient cause interactions and singular interactions for binary exposures can be found in VanderWeele and Richardson (2012). The basic definition for a sufficient cause interaction is that a sufficient cause interaction is said to be present if there are individuals for whom the outcome is 1 if all the exposures are 1 but the outcome is 0 whenever all but one of the exposures are one. Such a sufficient cause interaction implies the existence of a sufficient cause that requires all of the exposures to operate. The basic definition for a singular interaction between an arbitrary number of exposures is that there are individuals for whom the outcome is 1 if and only if all of the exposures are 1 (i.e., under any other combination of the exposures the outcome is 0). A singular interaction is always a form of compositional epistasis as defined in Section 10.6. If all or all but one of the exposures have monotonic effects on the outcome, then the concepts of a sufficient cause interaction and a singular interaction coincide; otherwise, a singular interaction is a stronger concept. See VanderWeele and Richardson (2012) for further details and an empirical example.

10.8. OTHER EXTENSIONS

A couple of further extensions are perhaps also worth mentioning. First, when examining interactions in settings in which at least one of the exposures is continuous, the continuous exposure is sometimes dichotomized at a particular level. The interpretation of interaction analyses can become somewhat more complicated under such dichotomization. Such dichotomization is potentially unproblematic when the dichotomized exposure occurs first, and one simply is comparing the effect of the other exposure across strata of the dichotomized exposure; this is simply a measure of effect heterogeneity as discussed in Sections 9.6 and 9.10 of the last chapter. However, when there is interest in causal interaction between both exposures, then the interpretation of interaction analyses can become more difficult when using a dichotomized exposure. Several results are now available (VanderWeele et al., 2011; Berzuini and Dawid, 2012) that describe the implications for mechanistic on the underlying continuous scale if the interaction contrasts considered above are used with a dichotomized exposure. For example, VanderWeele et al. (2011) show that with a continuous exposure and a binary exposure whose effects are unconfounded conditional on the covariates, if the continuous exposure has a positive monotonic effect on the outcome and is dichotomized at some cutoff point h and if the sufficient cause interaction condition, $p_{11} - p_{10} - p_{01} > 0$, is satisfied for the dichotomized exposure and the binary exposure, then there exist some individuals with the underlying continuous exposure at some level $v > h$, with the binary exposure present, and who had the outcome, but who would not have the outcome if the continuous exposure were reduced to level 0 and would also not have the outcome if the binary exposure were removed and the continuous exposure brought down to level h; in counterfactual notation, this is $Y_{v1} = 1$ but $Y_{01} = Y_{h0} = 0$. Various other implications also hold if it can be assumed that the distribution of the

two exposures are themselves statistically independent of one another; see VanderWeele et al. (2011) and Berzuini and Dawid (2012) for further results, examples, and discussion. Work still needs to be done in this area, however, in deriving more general conditions and tests.

Another extension that merits some discussion concerns empirical tests for sufficient cause interaction in settings in which it is assumed that the background causes (i.e., all of the components needed to complete the sufficient causes, beyond the primary exposures of interest, G and E) are independent of one another across sufficient causes. Some of the earlier literature in epidemiology (Rothman, 1974, 1978; Hogan et al., 1978; Weinberg, 1986) made this assumption. Unfortunately, such an assumption is not in general possible to verify or empirically test and in many settings will be unrealistic. However, when the assumption is made somewhat weaker, conditions are needed to test for sufficient cause interaction. For example, it can be shown (Novick and Cheng, 2004; Vansteelandt et al., 2008, Online Supplement) that if the background causes are assumed to be independent of one another and if both exposures have monotonic effects on the outcome, then a sufficient cause interaction must be present if

$$\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} < 1$$

This can also be rewritten as

$$(p_{11} - p_{10} - p_{01} + p_{00}) - p_{10}p_{01} + p_{11}p_{00} > 0$$

Note that this is somewhat different from our standard condition for sufficient cause interaction under monotonicity which was just $(p_{11} - p_{10} - p_{01} + p_{00}) > 0$. Neither is nested within the other. However, if either of these conditions are satisfied (and if monotonicity holds and control has been made for confounding), then a sufficient cause interaction must be present. It is also the case that if the background causes are independent of each other, then, without assuming monotonicity,

$$\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \neq 1$$

implies that a sufficient cause interaction must be present between G and E, or between $\overline{G}$ and E, or between G and $\overline{E}$, or between $\overline{G}$ and $\overline{E}$; that is, some form of synergism or antagonism must be present between G and E. We will consider concepts and methods for assessing antagonism in more detail in the following section. The condition $\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \neq 1$ was what was sometimes used in the earlier literature in epidemiology on assessing interaction in the sufficient cause sense. The condition is sometimes referred to as a violation of the multiplicative survival model. As we have discussed, it can be used to assess some form of sufficient cause interaction—but only under assumptions of independence of the background causes, which cannot be verified and may often be unrealistic. As we have discussed in previous sections, however, it is possible to test for sufficient cause

interaction without making the assumption of the independence of background causes, and the tests without these assumptions are thus more reliable.

Work has also been done on extending results for sufficient cause interaction to settings with stochastic counterfactuals and stochastic sufficient causes (VanderWeele and Robins, 2012; Ramasahai, 2013.)

10.9. ANTAGONISM

Throughout this chapter we have focused until this point on settings in which an outcome requires the presence of both of two exposures, at least for some individuals. In such cases we said that the exposures interacted synergistically. Of course it is also possible that in some settings, one exposure may prevent the other from operating. In such cases we might say that the exposures interact antagonistically. In this section we will provide some principles for assessing such antagonistic forms of interaction. We will focus on the setting of a dichotomous outcome and two dichotomous exposures, which we will again call G and E.

10.9.1. Notions of Antagonism

The term "antagonism" is ambiguous and is subject to different uses by different researchers. Here we will consider a number of possible uses of "antagonism" and relate these to the presence or absence of exposures within the sufficient cause framework. A classic setting that we might consider an antagonistic form of interaction within the sufficient cause framework would be the existence of a sufficient cause that required the presence of one exposure and the absence of another exposure to operate. This would essentially be synergism under the recoding of one of the exposures. However, there are also other settings that might reasonably be referred to as antagonistic interaction within the sufficient cause framework. For example, if there were a sufficient cause for the *absence* of the outcome that required the presence of both of our exposures, we might likewise call this an antagonistic interaction. This is essentially synergism under recoding of the outcome. Finally, there may be settings in which the outcome occurs if either of the exposures is present so that the exposures effectively compete to cause the outcome (e.g., if one exposure is the cause of the outcome, the other is not). We will refer to this as "competing antagonism." In what follows we will discuss the use of the term "antagonism" as it relates to the three scenarios listed above—in brief, either (i) synergism under exposure recoding, (ii) synergism under outcome recoding, or (iii) the outcome occurring if either exposure is present ("competing antagonism"). We will consider empirical tests for each of these forms of antagonism, and we will consider empirical approaches that might be used to assess whether any form of synergism or antagonism can be detected from the data.

In what follows we will, for simplicity, assume that the associations between the two exposures of interest and the outcome reflect the actual causal effects of

the exposures on the outcome (i.e., there is no confounding). If the effects of the exposures on the outcome are unconfounded conditional on some set of covariates C, then our conclusions will hold within strata of the covariates C. As before, we will make reference to "monotonicity" assumptions, meaning that for all individuals in a population, the exposure affects the outcome in the same direction. We will say that G has a positive monotonic effect on Y if the exposure is causative or neutral for all individuals (i.e., if Y_{ge} is nondecreasing in g). We will likewise say that G has a negative monotonic effect on Y (equivalently, $\overline{G}$ has a positive monotonic effect on Y) if it is preventive or neutral for all individuals (i.e., if Y_{ge} is nonincreasing in g). Similar definitions apply to E. We will offer several principles for assessing antagonism in the remarks that follow (cf. VanderWeele and Knol, 2011b).

10.9.2. Antagonism and Response Types

In this subsection we will consider the different forms of antagonism described above, and we will relate these to the response types in Table 10.4. In the following subsection we will consider empirical tests for these different forms of antagonism.

Within the sufficient cause framework, "antagonism" might be understood as "synergism under recoding," either a recoding of one of the exposures or a recoding of the outcome. We might first consider antagonism as synergism under the recoding of an exposure, so that antagonism is present if there is a sufficient cause with $\overline{G}E$ or similarly if there were a sufficient cause with $G\overline{E}$. Note that even if we restrict our attention to instances of "synergism under recoding of one of the exposures," the phrase may refer to a sufficient cause with $\overline{G}E$ or a sufficient cause with $G\overline{E}$; it would thus be important to clarify which of the two exposures is being recoded—that is, whether a sufficient cause with $\overline{G}E$ or with $G\overline{E}$ is in view. It can be shown (VanderWeele and Knol, 2011b) that response types 10, 12, and 14 in Table 10.4 necessarily imply antagonism understood as synergism under the recoding of one of the two exposures (what we will call "Class I Antagonism" or "exposure-based antagonism"). If there is an individual of response type 14, then it can be shown that there must be a sufficient cause with $G\overline{E}$ (VanderWeele and Knol, 2011b). Likewise, if there is an individual of response type 12, then there must be a sufficient cause with $\overline{G}E$. If there is an individual of response type 10, then there must be a sufficient cause with $G\overline{E}$ and another sufficient cause with $\overline{G}E$. All of these settings are instances of synergism under the recoding of the exposure, "exposure-based antagonism"; this will constitute our first class of antagonism.

Alternatively, we might conceive of antagonism as synergism under the recoding of the outcome; that is, antagonism is present if there is synergism for $\overline{Y}$. As we will see, this is not equivalent to "synergism under recoding of one of the exposures"; one may hold without the other. Response types 9 and 10 necessarily imply antagonism understood as synergism under the recoding of the outcome (what we will call "Class II Antagonism" or "outcome-based antagonism"). Suppose now that there were an individual of response type 9 or response type 10, then it can be shown that

there must be a sufficient cause GE for $\overline{Y}$ (VanderWeele and Knol, 2011b). This would imply synergism under the recoding of the outcome ("outcome-based antagonism"); this will constitute our second class of antagonism. As discussed below, causal co-action (the presence of two exposures in the same sufficient cause) is not invariant to the recoding of the outcome. Note that response type 10 is an instance of both exposure-based antagonism and outcome-based antagonism. Below we will consider how we can empirically test for these different forms of antagonism.

Finally, we might understand "antagonism" to also include response type 2—that is, individuals for whom the outcome occurs if one or the other of the exposures are present but not if both are absent so that, when both are present, the exposures effectively compete to cause the outcome (Class III Antagonism, "competing antagonism"). Response type 2 has sometimes been referred to in the literature as a "competing" or "antagonistic" response type (Greenland and Poole, 1988; Greenland et al., 2008) because if both exposures are present, they will effectively compete to cause the outcome. If response type 2 is taken as an antagonistic type, then this will be our third class of antagonism ("competing antagonism").

If we consider all three classes of antagonism together, this gives us response types 2, 9, 10, 12, and 14. Greenland et al. (2008) note that for each of these response types we have $Y_{11} - Y_{10} - Y_{01} + Y_{00} < 0$ or equivalently $(Y_{11} - Y_{00}) < (Y_{10} - Y_{00}) + (Y_{01} - Y_{00})$; that is, the effect of both exposures together is less than the sum of the effects of each considered separately. These are the only response types for which this is true. Rothman et al. (2008) thus refer to these individuals as "subadditive" types; such terminology could be used instead without making linguistic commitments as to what is and is not to be included under the category of "antagonism." Antagonism of Class I, Class II, or Class III is equivalent to the class of "subadditive" types.

10.9.3. Empirical Tests for Antagonism

In this section we will describe how it is possible to empirically test for each form of antagonism, or for any form of antagonism, both with and without monotonicity assumptions. As before, let $p_{ge} = P(Y = 1|G = g, E = e)$—that is, $p_{11}, p_{10}, p_{01}, p_{00}$—denote the overall levels of risk in the population for the outcome under each possible exposure combination. Using the results for synergism in the sections above (but recoding the exposure E), we have that there must be a sufficient cause for Y with $G\overline{E}$ if

$$p_{10} - p_{11} - p_{00} > 0 \qquad (\mathbf{10.1})$$

If (10.1) is satisfied then there must be an individual of response type 10 or 14 present and there thus must be a sufficient cause for Y with $G\overline{E}$. If G and $\overline{E}$ have positive monotonic effects on the outcome then assessing the weaker condition

$$p_{10} - p_{11} - p_{00} + p_{01} > 0 \qquad (\mathbf{10.2})$$

will suffice for this conclusion. Likewise, there must be individuals of response type 10 or 12 and consequently a sufficient cause for Y with $\overline{G}E$ if

$$p_{01} - p_{11} - p_{00} > 0 \tag{10.3}$$

If $\overline{G}$ and E have positive monotonic effects on Y, then

$$p_{01} - p_{11} - p_{00} + p_{10} > 0 \tag{10.4}$$

would suffice. Conditions (10.1)–(10.4) thus constitute empirical conditions for detecting antagonism of Class I, "exposure-based antagonism."

We can also empirically assess antagonism of Class II, "outcome-based antagonism." Again using the results for synergism under recoding we could conclude that there must be an individual of response type 9 or 10 and consequently a sufficient cause for $\overline{Y}$ with GE if

$$(1 - p_{11}) - (1 - p_{10}) - (1 - p_{01}) > 0. \tag{10.5}$$

If G and E have negative monotonic effects on Y (i.e., positive monotonic effects on $\overline{Y}$), then

$$(1 - p_{11}) - (1 - p_{10}) - (1 - p_{01}) + (1 - p_{00}) > 0 \tag{10.6}$$

would suffice.

We can also use analogues of the results for epistatic interaction, under appropriate recoding, to empirically assess not only Class I and Class II antagonism but also specific antagonistic response types. One could conclude an individual of response type 14 were present if

$$p_{10} - p_{11} - p_{00} - p_{01} > 0 \tag{10.7}$$

If one of G or $\overline{E}$ have a positive monotonic effect on Y, then (10.1) would suffice; if both do, then (10.2) suffices for this conclusion. One could conclude that an individual of response type 12 were present if

$$p_{01} - p_{11} - p_{00} - p_{10} > 0 \tag{10.8}$$

If one of $\overline{G}$ or E have a positive monotonic effect on Y, then (10.3) suffices; if both do, then (10.4) suffices. One could conclude that an individual of response type 9 were present if

$$(1 - p_{11}) - (1 - p_{10}) - (1 - p_{01}) - (1 - p_{00}) > 0 \tag{10.9}$$

If one of G or E have a negative monotonic effect on Y, then (10.5) suffices; if both do, then (10.6) suffices.

Finally one can similarly empirically assess the third class of antagonism, what we called "competing-antagonism," constituted by response type 2. One could conclude that an individual of response type 2 were present if

$$(1 - p_{00}) - (1 - p_{10}) - (1 - p_{01}) - (1 - p_{11}) > 0 \tag{10.10}$$

Table 10-10. EXAMPLE OF A TABLE OF EPISTATIC INTERACTION WHEN EXPOSURES HAVE THREE LEVELS

	$G_2 = 0$	$G_2 = 1$	$G_2 = 2$
$G_1 = 0$	0	0	0
$G_1 = 1$	0	1	1
$G_1 = 2$	0	1	1

If one of G or E have a positive monotonic effect on Y, then

$$(1 - p_{00}) - (1 - p_{10}) - (1 - p_{01}) > 0 \qquad (\mathbf{10.11})$$

will suffice for this conclusion; if both G and E have a positive monotonic effect on Y, then $(1 - p_{00}) - (1 - p_{10}) - (1 - p_{01}) + (1 - p_{11}) > 0$ will suffice, which can be rewritten as $p_{11} - p_{10} - p_{01} + p_{00} < 0$. The empirical conditions for each of the three forms of antagonism (rather than specific response types) are summarized in Table 10.11. In each case the contrast itself also serves as a lower bound for the prevalence of individuals with that form of antagonism (VanderWeele et al., 2010).

If we are simply interested in testing for any form of antagonism (Class I, II, or III)—that is, the presence of any of response types 2, 9, 10, 12, 14—then from the above discussion we have that, without any assumptions about monotonicity, it would suffice to test

$$p_{11} - p_{10} - p_{01} + p_{00} < 0 \qquad (\mathbf{10.12})$$

that is, it would suffice to test for "subadditivity." If (10.12) is satisfied, then there must be an individual of one of the antagonistic response types—that is, one of types 2, 9, 10, 12, 14. If (10.12) is satisfied and one or both of G or E have positive or negative monotonic effects on Y, then more specific conclusions can be made about which antagonistic response types must be present as described in Table 10.11.

By dividing any of the inequalities above by p_{00}, one obtains conditions on relative risks rather than risks. Such tests could also be applied to case–control data if the outcome were rare (or under incidence density sampling) so that odds ratios approximated risk ratios.

10.9.4. Subadditivity, Superadditivity, Synergism, and Antagonism

Typically, risks are said to be superadditive or to manifest positive effect-measure modification on the risk difference scale if

$$p_{11} - p_{10} - p_{01} + p_{00} > 0$$

and are said to be subadditive (negative effect-measure modification on the risk difference scale) if

$$p_{11} - p_{10} - p_{01} + p_{00} < 0$$

Table 10-11. SUMMARY OF EMPIRICAL CONDITIONS FOR DIFFERENT FORMS OF ANTAGONISM

Form of Antagonsim	**M**onotonicity Assumption	**E**mpirical Test
Class I: $G\overline{E}$ for Y	No assumption	$p_{10} - p_{11} - p_{00} > 0$
Class I: $G\overline{E}$ for Y	G and $\overline{E}$ with positive monotonicity	$p_{10} - p_{11} - p_{00} + p_{01} > 0$
Class I: $\overline{G}E$ for Y	No assumption	$p_{01} - p_{11} - p_{00} > 0$
Class I: $\overline{G}E$ for Y	$\overline{G}$ and E with positive monotonicity	$p_{01} - p_{11} - p_{00} + p_{10} > 0$
Class II: GE for $\overline{Y}$	No assumption	$(1 - p_{11}) - (1 - p_{10}) - (1 - p_{01}) > 0$
Class II: GE for $\overline{Y}$	G and E with negative monotonicity for Y	$(1 - p_{11}) - (1 - p_{10}) - (1 - p_{01}) + (1 - p_{00}) > 0$
Class III: Competing	No assumption	$(1 - p_{00}) - (1 - p_{10}) - (1 - p_{01}) - (1 - p_{11}) > 0$
Class III: Competing	Either G or E with positive monotonicity	$(1 - p_{00}) - (1 - p_{10}) - (1 - p_{01}) > 0$
Class III: Competing	Both G and E with positive monotonicity	$p_{11} - p_{10} - p_{01} + p_{00} < 0$

If risks are subadditive, then there must be an individual of response type 2, 9, 10, 12, or 14 i.e. of one of the antagonistic types and thus some form of causal co-action, either for Y or for $\overline{Y}$. By similar arguments, if risks are superadditive, then there must be an individual of response type 3, 5, 7, 8, or 15 and thus some form of causal co-action, either for Y or for $\overline{Y}$.

It thus follows that nonadditivity, that is,

$$p_{11} - p_{10} - p_{01} + p_{00} \neq 0$$

must imply one of response types 2, 3, 5, 7, 8, 9, 10, 12, 14, or 15 and causal co-action for either Y or $\overline{Y}$. Greenland and Poole (1988) call these ten types instances of "causal interdependence." This notion of "causal interdependence" given in Greenland and Poole (1988) is somewhat different from the notion of "definite interdependence," introduced by VanderWeele and Robins (2007b) and constituted by six types 7, 8, 10, 12, 14, and 15; "causal interdependence" allows us to know that there is causal co-action for either Y or $\overline{Y}$; "definite interdependence" allows us to know the specific form of causal co-action for Y.

10.9.5. A Procedure to Test for Any Form of Causal Co-Action for the Presence of an Outcome

In this section we will describe a procedure using a single simple calculation that allows one to assess, even in case–control studies, whether for the presence of a specific outcome any specific form of causal co-action between the presence or absence of two exposures can be detected from the data. Subadditivity and superadditivity

allow us to conclude that *some* form of causal co-action is present but in general do not allow us to determine what *specific* form is present. Here we will consider a simple calculation that allows one to determine whether any *specific* form of causal co-action may be present. This simple calculation, along with the one in the following subsection, will essentially automate the process of choosing the relevant empirical contrast from those above. We will consider two cases, one in which monotonicity assumptions may be plausible and a second in which no monotonicity assumptions are made; we will also consider data both on risks and on risk ratios (or odds ratios that may have been obtained from case-control data with a rare outcome).

If data are available on the risks, $p_{11}, p_{10}, p_{01}, p_{00}$, and it is thought that monotonicity assumptions may be plausible, select A as one of G or $1 - G$ and B as one of E or $1 - E$ so that p_{AB} is the largest of the four risks—that is, the largest of $p_{AB}, p_{(1-A)B}, p_{A(1-B)}, p_{(1-A)(1-B)}$ (equivalently of $p_{11}, p_{10}, p_{01}, p_{00}$). We use A and B here, rather than G and E, to notationally distinguish between the coding chosen by the investigator (i.e., A and B) and what might be the more natural coding (i.e., G and E). It is shown in the Appendix that if $p_{AB} - p_{(1-A)B} - p_{A(1-B)} + p_{(1-A)(1-B)} > 0$ and if A and B have positive monotonic effects on Y, then there is specifically causal co-action for Y between A and B. For example, if A had been selected as G and B had been selected as $1 - E$, then we would conclude that there was specifically a sufficient cause for Y with $G\overline{E}$. It is also shown in the Appendix that if $p_{AB} - p_{(1-A)B} - p_{A(1-B)} + p_{(1-A)(1-B)} \leq 0$, then no specific form causal co-action for Y can be detected from the data simply from the probabilities p_{11}, p_{10}, p_{01}, and p_{00}. Whether A and B have positive monotonic effects on Y must be evaluated on subject matter grounds though the true probabilities must at least satisfy $p_{AB} \geq \max(p_{(1-A)B}, p_{A(1-B)})$ and $p_{(1-A)(1-B)} \leq \min(p_{(1-A)B}, p_{A(1-B)})$. Without assumptions on monotonicity, the procedure is slightly modified: We again select A and B as above. If $p_{AB} - p_{(1-A)B} - p_{A(1-B)} > 0$, then there is specifically causal co-action between A and B; if $p_{AB} - p_{(1-A)B} - p_{A(1-B)} \leq 0$, then no specific form of causal co-action for Y can be detected from the data without monotonicity assumptions.

If data are instead available on risk ratios $RR_{11}, RR_{10}, RR_{01}$, and RR_{00}, where $RR_{ij} = p_{ij}/p_{00}$ and RR_{00} by default is 1 (or if data are available on odds ratios from a case–control study with rare outcome so that these approximate risk ratios), then select A as one of G or $1 - G$ and select B as one of E or $1 - E$ so that RR_{AB} is the largest (in magnitude on an absolute scale, i.e. most above 0) of the four risk ratios. The approach of the previous paragraph involves replacing the probabilities with risk ratios.

Note that the calculations above presuppose that it is known which risk is the largest; this would be the case if both exposures were assumed to have monotonic effects but it would be an additional assumption otherwise. A similar point holds also in the following section. Future work could consider settings in which the largest probability is assumed unknown and could derive statistical properties of a two-stage test procedure, first testing for the largest risk and then testing the relevant contrast. Simulations indicate that a naive two-stage testing procedure could be conservative or anti-conservative, depending on the true parameter values.

10.9.6. A Procedure to Test for Any Form of Causal Co-Action for the Absence of an Outcome

In the previous section we described a single simple calculation that allows one to assess whether, for the presence of a specific outcome, any specific form of causal co-action between the presence or absence of two exposures can be detected from the data. If we wish to detect whether any specific form of causal co-action for the absence of an outcome can be detected from the data, a different calculation can be employed.

VanderWeele and Robins (2007b) noted that whether some form of causal co-action was necessarily implied by a response type was invariant to the recoding of the exposures but not invariant to recoding of the outcome. That is to say, we might have causal co-action for the presence of the outcome but not for the absence of the outcome or vice versa. The example given in VanderWeele and Robins (2007b) was that in which G and E denote two genetic factors such that the individual will have the outcome Y if and only if both are present; there is thus causal co-action between G and E for Y as GE is a sufficient cause for Y. If instead we consider the absence of the outcome (i.e., $\overline{Y}$), then either the absence of the first factor (i.e., $\overline{G}$) or the absence of the second factor (i.e., $\overline{E}$) suffice for the absence of the outcome (i.e., for $\overline{Y}$); the sufficient causes for $\overline{Y}$ could thus be considered to be just $\overline{G}$ and $\overline{E}$, and there is no causal co-action of any form between G and E for $\overline{Y}$.

Causal co-action for the presence of an outcome is thus different from causal co-action for the absence of an outcome. In some cases, we might want to determine whether we can empirically detect any specific form of causal co-action for the absence of the outcome, that is, for $\overline{Y}$. Because causal co-action is not invariant to recoding of the outcome, we need to use simple calculations that are different from those given in the previous subsection. If data are available on the risks for Y, namely $p_{11}, p_{10}, p_{01}, p_{00}$, and it is thought that monotonicity assumptions may be plausible, select A and B so that p_{AB} is the smallest of the four risks for Y. If $p_{(1-A)B} + p_{A(1-B)} > p_{AB} + p_{(1-A)(1-B)}$ and if A and B have negative monotonic effects on Y (i.e., positive monotonic effects on $\overline{Y}$), then there is specifically causal co-action between A and B for $\overline{Y}$; if $p_{(1-A)B} + p_{A(1-B)} \leq p_{AB} + p_{(1-A)(1-B)}$, then no specific form causal co-action for $\overline{Y}$ can be detected from the data simply from the risks. Without assumptions on monotonicity, again select A and B as above. If $p_{(1-A)B} + p_{A(1-B)} > 1 + p_{AB}$, then there is specifically causal co-action between A and B for $\overline{Y}$; if $p_{(1-A)B} + p_{A(1-B)} \leq 1 + p_{AB}$, then no specific form of causal co-action for $\overline{Y}$ can be detected from the data without monotonicity assumptions.

If data are instead available on risk ratios $RR_{11}, RR_{10}, RR_{01}, RR_{00}$ for Y (or if data are available on odds ratios from a case–control study with rare outcome), then select A and B so that RR_{AB} is the smallest of the four risk ratios. If $RR_{(1-A)B} + RR_{A(1-B)} > RR_{AB} + RR_{(1-A)(1-B)}$ and if A and B have negative monotonic effects on Y (i.e., positive monotonic effects on $\overline{Y}$), then there is specifically causal co-action between A and B for $\overline{Y}$; if $RR_{(1-A)B} + RR_{A(1-B)} \leq RR_{AB} + RR_{(1-A)(1-B)}$, then no specific form causal co-action for $\overline{Y}$ can be detected from the data simply from the risk ratios. Without assumptions on monotonicity again, select A and

Table 10-12. CONCLUSIONS ABOUT THE PRESENCE OF RESPONSE TYPES UNDER VARIOUS MONOTONICITY ASSUMPTIONS AND SUPERADDITIVITY OR SUBADDITIVITY

Assumption	**S**uperadditivity	**S**ubadditivity
No monotonicity assumption	3, 5, 7, 8, or 15	2, 9, 10, 12, or 14
G positive monotonic	5 or 8	2 or 14
G negative monotonic	3 or 15	9 or 12
E positive monotonic	3 or 8	2 or 12
E negative monotonic	5 or 15	9 or 14
G positive, *E* positive	8	2
G positive, *E* negative	5	14
G negative, *E* positive	3	12
G negative, *E* negative	15	9

B as above. If $RR_{(1-A)B} + RR_{A(1-B)} > RR_{AB} + \frac{1}{p_{(1-A)(1-B)}}$, then there is specifically causal co-action between A and B for $\overline{Y}$. Note that some information on the prevalence $p_{(1-A)(1-B)}$ is needed to evaluate this inequality or alternatively one could consider the values of $p_{(1-A)(1-B)}$ for which the inequality is satisfied. If $RR_{(1-A)B} + RR_{A(1-B)} \leq RR_{AB} + \frac{1}{p_{(1-A)(1-B)}}$, then no specific form causal co-action for $\overline{Y}$ can be detected from the data simply from the risk ratios.

10.9.7. Illustrations of Antagonism

We present three examples, drawn from the existing literature on gene–gene and gene–environment interactions, illustrating the three forms of antagonism. More detailed calculations for these are given in the online supplement of VanderWeele and Knol (2011b). The confidence intervals here were obtained using methods described by Richardson and Kaufman (2009) (slightly different from the delta method procedures described in Chapter 9) but for the relative excess risk due to interaction, $RERI = RR_{11} - RR_{10} - RR_{01} + 1$, under appropriate recoding; when a contrast such as $RR_{11} - RR_{10} - RR_{01}$ is in view, a confidence interval can be obtained simply by subtracting 1 from both limits of the confidence interval for the *RERI*. We assume unconfoundedness in all examples that may not be realistic here; the examples are given only for illustrative purposes.

Consider first data summarized in Table 10.13 presented by Stern et al. (2002a), where G denotes ever smoking, E denotes the *Gln/Gln* versus the *Lys/Lys* or *Lys/Gln* genotype for the *XPY* codon 751, and Y denotes bladder cancer. If we follow the procedure in Section 10.9.5 to detect a specific form of causal co-action for Y, we select RR_{10} as the highest relative risk. Monotonicity is empirically violated in Table 10.12 (because $RR_{10} > RR_{00}$; but $RR_{11} < RR_{01}$), but if we assess the contrast for causal co-action without monotonicity, we have $RR_{10} - RR_{11} - RR_{01} = 3.6 - 2.1 - 1.0 = 0.5$ (95% CI: $-1.2, 2.1$). By the results in Section 10.9.5, we have that the point estimate of 0.5 would suggest evidence for causal co-action for

Table 10-13. Cases and Controls from Stern et al. (2002a), with Odds Ratios, for Gene-Gene Interaction; Example of "Exposure-Based Antagonism" (Class I)

	$E = 0$	$E = 1$
$G = 0$	1.0 (29 cases/68 controls)	2.6 (9 cases/8 controls)
$G = 1$	3.6 (171 cases/111 controls)	2.1 (21 cases/23 controls)

Table 10-14. Cases and Controls from Xu et al. (2007), with Odds Ratios for Gene–Environment Interaction; Example of "Outcome-Based Antagonism" (Class II)

	$E = 0$	$E = 1$
$G = 0$	1.0 (655 cases/629 controls)	1.0 (291 cases/272 controls)
$G = 1$	0.8 (50 cases/61 controls)	0.3 (15 cases/42 controls)

Y between G and $\overline{E}$ (antagonism of Class I, exposure-based antagonism) without any assumptions about monotonicity; this could also be seen from Section 10.9.3 inequality (10.1); note, however, the confidence interval here extends well below 0. Note that when monotonicity is violated, we cannot use the approach in Section 10.9.6 for causal co-action for the absence of an outcome for case–control data unless we have data on the outcome prevalence. Table 10.13 is representative of data suggesting antagonism of Form I, exposure-based antagonism.

Now consider data presented by Xu et al. (2007), considering possible interaction between CYP19A1 Rs1870050 polymorphisms and the consumption of polyphenol-rich food and beverages on endometrial cancer. Specifically, let G denote high versus low tea consumption and let E denote the CC versus the AA/AC genotype. Odds ratios (which approximate risk ratios) obtained from the number of cases and controls in Xu et al. (2007) are summarized in Table 10.14. If we follow the procedure in Section 10.9.6 to detect a specific form of causal co-action for $\overline{Y}$, then we have that RR_{11} is the lowest relative risk, and $RR_{01} + RR_{10} - RR_{00} - RR_{11} = 0.5$ (95% CI: 0.0, 0.9). By the results described in Section 10.9.6, if G and E have negative monotonic effects on Y (i.e., are preventive or neutral for all individuals), then there is causal co-action for $\overline{Y}$ between G and E (antagonism of Class II, outcome-based antagonism); this could also be seen from Section 10.9.3, inequality (10.6). Said another way, high tea consumption and the CC genotype interact synergistically to prevent Y. Table 10.14 is representative of data suggesting antagonism of Class II, outcome-based antagonism.

Finally consider the data from Stern et al. (2002b) summarized in Table 10.15 where G denotes the Arg/Arg genotype for $XRCC1$ Codon 194 and E denotes the presence of Met variants at $XRCC3$ codon 241 and Y denotes bladder cancer. If we follow the procedure in Section 10.9.5, the condition for a specific form of causal co-action for Y is not satisfied. If we follow the procedure in Section 10.9.6 to detect a specific form of causal co-action for $\overline{Y}$, then we have that RR_{00} is the lowest relative

Table 10-15. CASES AND CONTROLS FROM STERN ET AL. (2002B), WITH ODDS RATIOS FOR GENE–GENE INTERACTION; EXAMPLE OF "COMPETING ANTAGONISM" (CLASS III)

	$E = 0$	$E = 1$
$G = 0$	1.0 (6 cases/18 controls)	3.3 (83 cases/76 controls)
$G = 1$	3.2 (20 cases/19 controls)	4.0 (123 cases/93 controls)

risk and $RR_{10} + RR_{01} - RR_{11} - RR_{00} = 3.2 + 3.3 - 4.0 - 1.0 = 1.5$ (95% CI: $-1.9, 4.8$). If $1 - G$ and $1 - E$ have negative monotonic effects on Y (i.e., G and E have positive monotonic effects on Y), then by the approach of Section 10.9.6, the point estimate of 1.5 gives some evidence of causal co-action for $\overline{Y}$ between $\overline{G}$ and $\overline{E}$—that is, under monotonicity for individuals of response type 2 (antagonism of Class III, competing antagonism); this could also be seen from Section 10.9.3 inequality (10.12). Note, however, that the confidence interval here includes 0. Table 10.15 is representative of data suggesting antagonism of Class III, competing antagonism.

10.10. LIMITS OF INFERENCE CONCERNING BIOLOGY

In a recent review article on gene–gene interaction ("epistasis"), Phillips (2008) distinguished three types of epistasis when two genetic exposures were in view: (i) statistical epistasis (i.e., interaction in a statistical model), (ii) compositional epistasis (e.g., Y occurs if and only if $G_1 = G_2 = 1$ as in Section 10.5), and (iii) functional epistasis, the actual the physical interaction of proteins. In Section 10.5 we have considered new tests for "compositional epistasis." Although it was previously thought that such epistasis could not be detected using statistical tests (Cordell, 2002); we have seen that one can test for it, but this often requires a nonstandard interaction test (conditions more stringent than the presence of statistical interaction). But even compositional epistasis does not necessarily imply functional epistasis, that is, the physical interaction of proteins.

To see this, suppose that G_1 and G_2 are two genetic factors. Suppose that when $G_1 = 1$, protein 1 is not produced and that when $G_2 = 1$, protein 2 is not produced. Suppose that the outcome Y occurs if and only if neither protein 1 nor protein 2 are present. We then have an epistatic interaction because the outcome occurs if and only if $G_1 = 1$ and $G_2 = 1$ (i.e., if both genetic factors are present so that both proteins are absent), but we do not have physical interaction here. It is precisely the absence of the proteins that gives rise to the outcome; there simply is nothing to physically interact here. Here we have compositional epistasis but not "functional" epistasis. Although we can sometimes empirically draw conclusions about compositional epistasis from data, empirical tests will not in general allow us to draw conclusions about functional epistasis or the actual physical interaction between exposures, and it is important to understand the limits of the conclusions being drawn about these alternative forms of causal interaction. From our tests for

mechanistic interaction, either sufficient cause interaction or epistatic interaction, we do learn something about mechanisms—that is, about what turns the outcome "on" or "off"—but we don't necessarily learn about physical interaction between exposures.

Other examples demonstrating the limits of biologic inference concerning interaction were given by Siemiatycki and Thomas (1981). Consider, for example, a setting in which for the outcome to occur, two stages of disease development must occur. Several theories for the development of cancer follow this model. Suppose that the two exposures of interest, G_1 and G_2 say, affect different stages: G_1 acts on stage 1 and G_2 acts on stage 2. Suppose also in this example that stage 1 and stage 2 are completely independent of each other. Perhaps the baseline probability of stage 1 occurring is 1% and the baseline rate of stage 2 occurring is also 1%, so that the baseline likelihood of disease is 0.01%. Suppose that G_1 increases the probability of stage 1 occurring from 1% to 2% and G_2 increases the probability of stage 2 occurring from 1% to 5%. Suppose, however, that the presence of G_2 in no way alters the effect of G_1's increasing the probability of stage 1 occurring from 1% to 2%; that is, the probability of stage 2 is 1% if $G_1 = 0$ and 2% if $G_1 = 1$, irrespective of whether G_2 is present or absent. Suppose, similarly, that the presence of G_1 in no way alters the effect of G_2's increasing the probability of stage 2 occurring from 1% to 5%; that is, the probability of stage 2 is 1% if $G_2 = 0$ and 5% if $G_2 = 1$, irrespective of whether G_1 is present or absent. Here then we seem to have no interaction between G_1 and G_2 at the biologic level.

As noted above, if neither exposure ($G_1 = 0$ and $G_2 = 0$) is present, then the risk of stage 1 and stage 2 are both 1% and the overall likelihood of the outcome is $1\% \times 1\% = 0.01\%$. If just G_1 is present ($G_1 = 1$ and $G_2 = 0$), then the risk of stage 1 is 2% and the risk of stage 2 is 1% and the overall likelihood of the outcome is $2\% \times 1\% = 0.02\%$. If $G_1 = 0$ and $G_2 = 1$, then the risk of stage 1 is 1% and the risk of stage 2 is 5% and the overall likelihood of the outcome is $1\% \times 5\% = 0.05\%$. If $G_1 = 1$ and $G_2 = 1$, then the risk of stage 1 is 2% and the risk of stage 2 is 5% and the overall likelihood of the outcome is $2\% \times 5\% = 0.10\%$. In this example our measure of multiplicative interaction is $\frac{p_{11}p_{00}}{p_{10}p_{01}} = \frac{0.10\%(0.01\%)}{0.02\%(0.05\%)} = 1$. However, our measure of additive interaction is

$$p_{11} - p_{10} - p_{01} + p_{00} = 0.10\% - 0.02\% - 0.05\% + 0.01\% = 0.04\% > 0$$

We have positive additive interaction but no biologic interaction in this example. Here our conditions for sufficient cause interaction

$$p_{11} - p_{10} - p_{01} = 0.10\% - 0.02\% - 0.05\% = 0.03\% > 0$$

and even for "epistatic" or "singular" interaction

$$p_{11} - p_{10} - p_{01} - p_{00} = 0.10\% - 0.02\% - 0.05\% - 0.01\% = 0.02\% > 0$$

are also satisfied. But again we saw that there was no interaction between G_1 and G_2 at the biologic level. How are we to make sense of this? What we can conclude from

the condition for a singular or epistatic interaction, say, is that there are some individuals who would have the outcome if both exposures were present but who would not if just one or the other or neither exposure were present. But we see here that not even this necessarily indicates interaction at some fundamental biologic level. We have this form of "singular" or "sufficient cause" interaction because if both exposures are present, 0.10% have the outcome and this cannot be accounted by those individuals whose outcome only required the first exposure (0.02%) or only the second (0.05%) or who required neither (0.01%). Even if these three groups were mutually exclusive, they would not account for the risk of 0.10% that occurs if both exposures are present $(0.10\% - (0.02\% + 0.05\% + 0.01\%) = 0.02\% > 0)$. There thus must be some individuals for whom the outcome occurs if and only if both exposures are present. But again, this does not, as this example shows, indicate physical interaction in any fundamental biologic sense. As before, we need to be careful about drawing conclusions about the underlying biology. Not even our conclusion of sufficient cause interaction or epistatic interaction allows us to draw conclusions about exposures physically touching each other in the underlying biology.

On the basis of these and other similar examples, Thompson (1991) suggested that if an outcome required stages and one exposure affected the first stage and another exposure affected the second stage (a "multistage model") and there were no biologic interaction, we would expect a multiplicative model. Likewise, he suggested that if the occurrence of a single adverse event was sufficient for the development of the disease (a "single-hit model"), then under the absence of biologic interaction we would expect an additive model. Finally, he suggested that if the outcome occurred if an individual failed to experience any of one or more occurrence of a beneficial event (a "no-hit model", cf. Walter and Holford, 1978), then the model should again be multiplicative. While such heuristics may be of some use, if we do find that an additive model fits well, it is not necessarily that we have a "single-hit model" with no biologic interaction; it could equally be the case that we have a "multistage model" in which the factors operate antagonistically. Or if we were to find that the multiplicative model fit well, this does not necessarily indicate a "multistage model" with no biologic interaction, but could also be a "single-hit model" in which there was biologic interaction. We cannot in general draw conclusions about the type of biologic model and the presence or absence of biologic interaction simply from the statistical models we use. If we find positive multiplicative interaction, this could be a "multistage model" or a "no-hit" model with biologic interaction, or it could be a "single-hit model" with biologic interaction, or it could be a more complicated model with no biologic interaction whatsoever. We cannot tell from the data alone. Our inferences about biology from empirical data are limited. We can assess statistical interaction (on any scale we choose), we can assess additive interaction to determine how best to allocate interventions, and we can assess "sufficient cause" or "epistatic/singular" interaction to determine whether there are individuals who would have the outcome if both exposures were present but not if only one or the other were present. All of these may provide some insight into the underlying biology, but we have no way of going from any of these forms of interaction which we can assess with data directly to the underlying biology itself.

In some earlier literature, sufficient cause synergism was sometimes referred to as "biologic interaction" (e.g., Rothman and Greenland, 1998); sometimes even just additive interaction was even referred to as "biologic interaction" (e.g., Andersson et al., 2005). However, as we have seen in the examples above, neither statistical additive interaction nor even sufficient cause interaction or epistatic interaction/compositional epistasis necessarily tells us anything about physical or functional interactions. Statistical analyses can only tell us limited information about the underlying biology (Siemiatycki and Thomas, 1981; Thomas, 1991; Rothman and Greenland, 1998; Cordell, 2002). Because of this there has been a suggestion to move away from the use of "biologic interaction" for sufficient cause synergism (cf. Lawlor, 2011; VanderWeele, 2011e). It may be more appropriate to refer to these sufficient cause or epistatic interactions as "mechanistic interactions"; these are still cases in which both exposures together turn the outcome "on" and the removal of one turns the outcome "off" and thus the "mechanistic" description seems potentially appropriate. If even this is thought to be language that is too strong—if "mechanistic," rather than indicating "on" and "off," is still thought to indicate biology—then the term "counterfactual interaction" could perhaps be used instead for forms of sufficient cause or epistatic/singular interaction.

10.11. DISCUSSION

In this chapter we have considered different notions of mechanistic interaction. We have seen that it is possible to empirically test for the presence of individuals for whom an outcome will occur if both of two exposures are present but not if only one or the other is present (a "sufficient cause interaction") or alternatively for individuals for whom the outcome will occur if and only if both of two exposures are present (an "epistatic interaction"). We have seen how these counterfactual notions of interaction are related to synergism, or the presence of a sufficient cause which involves both of two exposures, within the sufficient cause framework. We have also seen that these various counterfactual forms of interaction are in general not implied by nonzero interaction terms in a statistical model, except under much stronger monotonicity assumptions. Otherwise, statistical interaction can be present even with no mechanistic interaction. We have discussed extensions of these various concepts to settings in which exposures have more than two levels, or in which there are three or more exposures, or in settings in which a continuous exposure has been dichotomized, or when antagonism, rather than synergism, is in view and of interest. We have also discussed the limits of the conclusions we can draw about biology from empirical testing. Although conclusions about "counterfactual" or "mechanistic" forms of interaction are possible, not even these necessarily imply the physical interaction between exposures, or interaction at the most fundamental biologic level. We can make progress testing for mechanistic interaction but need also to understand the limits of the types of inferences that we can draw.

11

Bias Analysis for Interactions

In Chapter 9 we discussed confounding control for interaction analyses. We noted that if interventions were being considered for both of the exposures in the interaction analysis, then control had to be made for two sets of confounding factors for each of the two exposures. If control has not been made for such confounding variables, then interaction measures will in general be biased. Likewise, we have assumed throughout our discussion that our exposures and outcomes of interest are correctly measured. Error-prone measurements might likewise bias estimates of interaction. In this chapter we will consider a number of different methods and results to address issues of bias in interaction analysis. We will provide sensitivity analysis tools for unmeasured confounding of one or both exposures in the interaction analysis. We will also see that in some settings interaction analyses may be robust to unmeasured confounding even when estimates of the effects of the individual exposures are not. We will also discuss settings in which interaction analysis may be robust to measurement and discuss some basic sensitivity analysis tools that can be employed when measurement error biases analyses.

11.1. SENSITIVITY ANALYSIS AND ROBUSTNESS FOR ADDITIVE INTERACTION

In the next several sections we describe sensitivity analysis approaches to assess the sensitivity of interaction estimates to the presence of an unmeasured confounder. We will begin in this section with results for the additive scale. In Section 11.2 we will turn to the multiplicative scale, and in Section 11.3 we will give an approach that can be used when assessing additive interaction with ratio measures using the relative excess risk due to interaction, as might be done in a case–control study and as described in Chapter 9. In each setting we will also discuss under what circumstances interaction measures are robust to unmeasured confounding, even if the main effect estimates are not.

Behavioral, biological, and chemical exposures may interact with one another in producing their effects, and the effects of such exposures may also be modified by

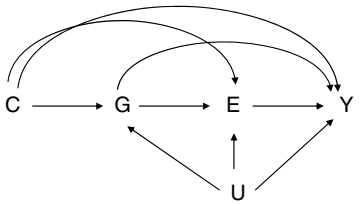

Figure 11.1 Unmeasured confounder U of the relationships between the exposures, G and E, and the outcome Y.

various genetic factors. Although genetic factors often are assumed effectively randomized, environmental factors and behavioral, biological, and chemical exposures are subject to the same confounding as they would be in any observational study. In some studies the effects of genetic factors may be confounded by population stratification if adequate control for this has not been made. Unmeasured confounding is clearly an issue in interaction analyses, and the methods described below will help address this issue.

11.1.1. Sensitivity Analysis for Additive Interaction

We will let G and E denote our two factors or exposures of interest. These might represent genetic and environmental factors respectively but could be any two behavioral, environmental, genetic, biologic, or social exposures. We will let Y denote the outcome of interest and we will let C denote the set of measured covariates. We will allow the two exposures and the outcome to be binary or continuous.

Suppose that the measured covariates C do not themselves suffice to control for confounding for the effects of the exposures G and E on the outcome but that there is some unmeasured variable U such that if it were possible to control for both C and U, then this would suffice to control for confounding. The relationships between the variables are depicted in Figure 11.1. It may also be the case that C affects U or that U affects C, and the results below will still apply.

Suppose we wish to compare two levels of the first exposure, g_1 and g_0 say, and two levels of the second exposure, call them e_1 and e_0. We will define the bias factor $B_{add}(c)$ on the additive scale as the difference between (i) an additive interaction measure on the difference scale conditional on covariates $C = c$, that is, $\mathbb{E}[Y|g_1, e_1, c] - \mathbb{E}[Y|g_0, e_1, c] - \mathbb{E}[Y|g_1, e_0, c] + \mathbb{E}[Y|g_0, e_0, c]$, and (ii) what we would have obtained as the additive interaction measure had we been able to adjust for U. We will give expressions for this bias factor in terms of sensitivity analysis parameters that relates the effects of the unmeasured confounder U to the exposures and to the outcome. Very general results are given in the Appendix but require specifying a large number of sensitivity analysis parameters. Here we will consider one relatively easy-to-use sensitivity analysis technique under simplifying assumptions. We will make the simplifying assumption that U is binary and does not interact with G on the additive scale in the sense that the effect of U within strata of G, E, C—that is, $\mathbb{E}[Y|g, e, c, U = 1] - \mathbb{E}[Y|g, e, c, U = 0]$—does not depend

on g. We will then specify four sensitivity analysis parameters. Two parameters will correspond to the effect of U on Y and two will capture how the prevalence of U varies across the strata defined by exposures. As will be seen below, the approach we take here will be very similar to that presented in Chapter 3 on sensitivity analysis for total effects. We will essentially use the same sensitivity analysis approach as in Chapter 3 and will apply it to the exposure G, but we will do so twice, once for when the environmental exposure takes value e_1 and once for when it takes value e_0.

Specifically, let $\gamma_1 = \mathbb{E}[Y|g, e_1, c, U = 1] - \mathbb{E}[Y|g, e_1, c, U = 0]$ denote the effect of U on Y when the exposure E takes value e_1; also let $\delta_1 = P(U = 1|g_1, e_1, c) - P(U = 1|g_0, e_1, c)$ denote the prevalence difference of the unmeasured confounder U when $E = e_1$ comparing $G = g_1$ and $G = g_0$. These are essentially the same types of parameters we used in Chapter 3, and here we specify them when the environmental factor takes value $E = e_1$. Now we also let $\gamma_0 = \mathbb{E}[Y|g, e_0, c, U = 1] - \mathbb{E}[Y|g, e_0, c, U = 0]$ denote the effect of U on Y when the exposure E takes value e_0; and we let $\delta_0 = P(U = 1|g_1, e_0, c) - P(U = 1|g_0, e_0, c)$ denote the prevalence difference of the unmeasured confounder U when $E = e_0$ comparing $G = g_1$ and $G = g_0$. We are again specifying sensitivity analysis parameter similar to those in Chapter 3, but now we are also specifying them for when the environmental factor takes value $E = e_0$. We thus have two sets of sensitivity analysis parameters analogous to what we considered in Chapter 3: one set for when $E = e_1$ and one set for when $E = e_0$. By allowing the sensitivity analysis parameters γ_1 and γ_0 to differ, we are allowing for potential $U \times E$ interaction: The effect of U on Y may vary across different levels of the exposure E. By allowing δ_1 and δ_0 to differ, we are essentially allowing for potential dependence between the distributions of G and E.

Under the simplifying assumption that U is binary and does not interact in its effects on Y with G on the additive scale, we have that the bias factor is given by the formula

$$B_{add}(c) = \gamma_1\delta_1 - \gamma_0\delta_0$$

Thus to calculate the bias factor, we need to specify the effect of U on Y when the exposure E takes value e_1 or e_0, respectively (i.e., γ_1 and γ_0), and we need to specify the prevalence difference of U, comparing $G = g_1$ and $G = g_0$, when the exposure E takes value e_1 or e_0, respectively (i.e., δ_1 and δ_0). We can then use the formula above to calculate the bias factor, and we can subtract this bias factor, from our estimate of additive interaction, controlling only for C, to obtain a corrected estimate. Under the simplifying assumptions above, we can also subtract this bias factor from both limits of the confidence interval to obtain a corrected confidence interval.

We may not believe any particular specification of the sensitivity parameters, but we could vary these parameters over a range of plausible values to obtain what were thought to be a plausible range of corrected estimates. The range of the values over which the sensitivity analysis parameters are varied could be determined by substantive knowledge or other prior studies that may have estimates for the covariates as well. Using this technique, we could also examine how substantial the confounding would have to be to explain away an effect (we could do this for the estimate and the confidence interval). Note that here the bias factor is the difference between the products of the two sets of sensitivity analysis parameters. Thus, for unmeasured

confounding to introduce substantial bias to interaction estimates, it is not sufficient that the effect of the unmeasured confounder on the outcome is large or that the difference in prevalence of the unmeasured confounder comparing those with $G = g_1$ versus $G = g_0$ is large. Rather, these sensitivity analysis parameter must differ when $E = e_1$ versus when $E = e_0$. If they do not differ, then the difference in the products of the two sets of sensitivity analysis parameters will largely cancel each other out and the bias will be small. Stated another way, to get substantial bias for an interaction estimate, the unmeasured confounder must affect the outcome Y and the genetic exposure G differently when the environmental factor takes values $E = e_1$ versus $E = e_0$.

Note, by symmetry, that if E does not interact with U on the additive scale (rather than U not interacting with G), then the bias factor is analogously given by $B_{add}(c) = \gamma_1^*\delta_1^* - \gamma_0^*\delta_0^*$ where $\gamma_i^* = \mathbb{E}[Y|g_i, e, c, U = 1] - \mathbb{E}[Y|g_i, e, c, U = 0]$ is the effect of U on Y in different strata of G and $\delta_i^* = P(U = 1|g_i, e_1, c) - P(U = 1|g_i, e_0, c)$ is the prevalence difference in U comparing $E = e_1$ and $E = e_0$, in different strata of G.

Example. We consider an analysis of additive interaction between the effects of smoking and arsenic exposure in wellwater on causing premalignant skin lesions (Chen et al., 2006; VanderWeele et al. 2010). Data come from a large cohort study of 11,746 individuals in Bangladesh, many of whom had been exposed to various doses of arsenic through drinking well water. Let $G = 1$ for high versus low arsenic ($>$ versus $<$ 100 μg/L) and let $E = 1$ for smoking (ever versus never) with $Y = 1$ denoting the presence of premalignant skin lesions. Adjustment is made for gender, age, education, BMI, land, and TV ownership (markers of socioeconomic status in Bangladesh), fertilizer use and pesticide use. Analyses indicate interaction on the additive risk difference scale of 3.6% (95% CI: 0.1%, 7.1%). It is possible that smoking and perhaps also arsenic exposure are subject to unmeasured confounding by, say, residual socioeconomic status (SES). If the effect of high versus low SES were a $\gamma_1 = 2.0$ percentage point increase in skin lesion risk with smoking present and a $\gamma_0 = 1.0$ percentage point increase in risk when smoking is absent and if the prevalence difference of SES for high versus low arsenic were $\delta_1 = 0.5$ for smokers, and $\delta_0 = 0.4$ for non-smokers, then we would have $B_{add}(c) = \gamma_1\delta_1 - \gamma_0\delta_0 = 2(0.5) - 1(0.4) = 0.6\%$ and the corrected estimate would be $3.6\% - 0.6\% = 3.0\%$ (95% CI: -0.5%, 6.5%). If we changed the numbers to $\gamma_1 = 4\%$ versus $\gamma_0 = 1\%$ and $\delta_1 = 0.7$ versus $\delta_0 = 0.2$, we would obtain $B_{add}(c) = \gamma_1\delta_1 - \gamma_0\delta_0 = 4(0.7) - 1(0.2) = 2.6\%$ and the corrected estimate would be $3.6\% - 2.6\% = 1.0\%$ (95% CI: -2.5%, 4.5%). In neither of these scenarios is the estimate reduced to 0; however, in both of the scenarios the confidence interval extends below 0.

11.1.2. Robustness of Additive Interaction to Unmeasured Confounding

In some cases, interaction estimates are in fact robust to unmeasured confounding even if main effect estimates are not. Suppose now that U were only a confounder

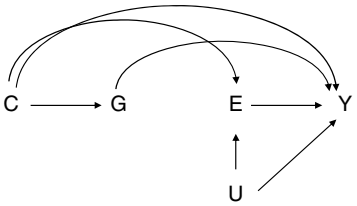

Figure 11.2 Unmeasured confounder that affects only the environmental exposure E and the outcome Y.

for E and that we had $G \times E$ independence in the sense that the distributions of E and its confounder U were statistically independent of G conditional on the measured covariates C, as in Figure 11.2.

In this case if G does not interact with U on the additive scale in the sense that $\mathbb{E}[Y|g,e,c,u] - \mathbb{E}[Y|g,e,c,u']$ is constant across strata of g, then the bias is 0, that is, $B_{add}(c) = 0$. In other words, the interaction estimates are still valid, even though we have unmeasured confounding of the environmental factor. See Appendix for proof. Thus if we have independence between the distributions of the two exposures and if the unconfounded exposure (G) does not interact with the unmeasured confounder (U), then our interaction estimates will be valid even though our estimates of the main effects of the confounded exposure (E) would not be. By symmetry, if U were only a confounder for G and we had $G \times E$ independence in the sense that $\{G, U\}$ were independent of E conditional on the measured covariates and if U does not interact with E on the additive scale, then $B_{add}(c) = 0$. Note that these robustness results do not assume that U is binary.

This robustness property also has an interesting implication that immediately follows. Suppose that U were an environmental factor that was a confounder only for E, not G. A consequence of the robustness property is that if we have G and E statistically independent in distribution and if we found that the our estimated measure of additive interaction were nonzero, then if there is no interaction between U and G on the additive scale, we have $B_{add}(c) = 0$. Thus if we found that our estimated measure of additive interaction were nonzero, then either there is an actual $G \times E$ interaction (because $B_{add}(c) = 0$ and the estimated interaction is equal to the causal interaction) or there is a $G \times U$ interaction, another form of gene–environment interaction. Essentially, under gene–environment independence, even with unmeasured confounding we have some form of gene–environment interaction either with E or with U.

A result similar to this holds also under $G \times E$ independence if there is an unmeasured genetic confounder U_1 for G and another unmeasured environmental confounder U_2 for E that are binary and independent of one another. In this case, if G doesn't interact with U_2 on the additive scale, E doesn't interact with U_1 on the additive scale, and U_1 doesn't interact with U_2 on the additive scale, then $B_{add} = 0$. Thus if the estimated additive interaction measure were nonzero, one could include either a true causal $G \times E$ interaction, a $G \times U_1$ interaction, a $E \times U_2$ interaction, or a $U_1 \times U_2$; that is, some form of gene–environment interaction would be present. A formal statement of the result is given in the Appendix.

11.2. SENSITIVITY ANALYSIS AND ROBUSTNESS FOR MULTIPLICATIVE INTERACTION

11.2.1. Sensitivity Analysis for Multiplicative Interaction

We now give a similar sensitivity analysis technique for the multiplicative scale. Suppose again that the measured covariates C do not themselves suffice to control for confounding for the effects of the exposures G and E on the outcome but that there is some unmeasured variable U such that if it were possible to control for both C and U, then this would suffice to control for confounding as in Figure 11.1 (but again allowing C to affect U or vice versa).

We will define the bias factor $B_{mult}(c)$ on the multiplicative scale as the ratio of (i) the multiplicative interaction measure on the risk ratio scale conditional on covariates $C = c$, that is, $\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_1,c]} / \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}$, and (ii) what we would have obtained as the multiplicative interaction measure had we been able to adjust for U as well. Again we will consider one relatively easy-to-use sensitivity analysis technique under simplifying assumptions. Suppose we wish to compare two levels of G, g_1, and g_0 and two levels of E, e_1, and e_0. We will make the simplifying assumption that U is binary and does not interact with G on the multiplicative scale in the sense that the effect of U within strata of G, E, C, that is, $\mathbb{E}[Y|g,e,c,U=1]/\mathbb{E}[Y|g,e,c,U=0]$, does not depend on g. We will then specify several sensitivity analysis parameters. Two parameters will correspond to the effect of U on Y on the risk ratio scale, and four will capture how the prevalence of U varies across the strata defined by exposures.

Specifically, let $\gamma_1 = \mathbb{E}[Y|g,e_1,c,U=1]/\mathbb{E}[Y|g,e_1,c,U=0]$ and $\gamma_0 = \mathbb{E}[Y|g,e_0,c,U=1]/\mathbb{E}[Y|g,e_0,c,U=0]$ denote the effect of U on Y when the exposure E takes value e_1 or e_0, respectively. Also let $P(U=1|g_i,e_j,c)$ denote the prevalence of U when $G=g_i$ and $E=e_j$, that is, the prevalence in each of the $G \times E$ strata. Under the simplifying assumption that U is binary and does not interact in its effects on Y with G on the multiplicative scale, we have that the bias factor is given by the formula

$$B_{mult}(c) = \frac{1+(\gamma_1-1)P(U=1|g_1,e_1,c)}{1+(\gamma_1-1)P(U=1|g_0,e_1,c)} \Big/ \frac{1+(\gamma_0-1)P(U=1|g_1,e_0,c)}{1+(\gamma_0-1)P(U=1|g_0,e_0,c)}$$

Thus to calculate the bias factor we need to specify the effect of U on Y when the exposure E takes value e_1 or e_0 respectively (i.e., γ_1 and γ_0), and we need to specify the prevalence difference of U in each of the four $G \times E$ strata. We can then use the formula above to calculate the bias factor and we can divide our estimate of multiplicative interaction, controlling only for C, by the bias factor $B_{mult}(c)$ to obtain a corrected estimate. Under the simplifying assumptions above, we can also divide both limits of the confidence interval by the bias factor to obtain a corrected confidence interval. As before, similar results also hold by symmetry if E does not interact with U on the additive scale (rather than U not interacting with G).

The sensitivity analysis techniques for additive interaction in the previous section would be applicable to cohort data. The multiplicative interaction results

will also be applicable to cohort designs and will moreover be applicable case–control designs when the outcome is rare so that odds ratios approximate risk ratios. The multiplicative results will likewise be applicable to so-called "case-only" designs which will be described in the next chapter, when the distribution of the genetic and environmental factors are statistically independent because under such an independence assumption, the case-only design allows one to estimate interaction on the multiplicative scale. The multiplicative results are also applicable to family-based study designs, mentioned in the next chapter which estimate the interaction on the log scale or in settings in which the outcome is rare.

11.2.2. Robustness of Multiplicative Interaction to Unmeasured Confounding

As with additive interaction, so also with multiplicative interaction, estimates are sometimes robust to unmeasured confounding. Suppose that U were only a confounder for E and that we had $G \times E$ independence in the sense that E and its confounder U were independent of G conditional on the measured covariates C, as in Figure 11.2. In this case if G does not interact with U on the multiplicative scale in the sense that $\mathbb{E}[Y|g,e,c,u]/\mathbb{E}[Y|g,e,c,u']$ is constant across strata of g, then the bias factor is 1; that is, there is no bias. The multiplicative interaction estimates in this setting are still valid, even though we have unmeasured confounding of the environmental factor. Thus if we have independence between the two exposures and the unconfounded exposure (G) does not interact on the multiplicative scale with the unmeasured confounder (U), then our interaction estimates will be valid even though our estimates of the main effects of the confounded exposure (E) would not be. Again, by symmetry, if U were only a confounder for G and we had $G \times E$ independence in the sense that $\{G,U\}$ were independent of E conditional on the measured covariates and if U does not interact with E on the multiplicative scale, then $B_{mult}(c) = 1$. Note that these robustness results do not assume that U is binary.

As with the additive case, this robustness property has interesting implications for testing for gene–environment interaction. Suppose that U were an environmental factor that was a confounder only for E, not G. A consequence of the robustness property is that we have $G \times E$ independence and we found that the our estimated measure of multiplicative interaction differed from 1 and if there is no interaction between U and G on the multiplicative scale, then $B_{mult}(c) = 1$; that is, there is no bias. Thus if we found our estimated measure of multiplicative interaction different from 1, then either there is an actual $G \times E$ multiplicative interaction (because $B_{mult}(c) = 1$ and the estimated interaction is equal to the causal interaction) or there is a $G \times U$ multiplicative interaction, another form of gene–environment interaction. Essentially, under gene–environment independence, even with unmeasured confounding we have some form gene–environment multiplicative interaction either with E or with U.

Example. Bennett et al. (1999) studied the interaction between passive smoking and glutathione S-transferase M1 (GSTM1) on lung cancer risk among nonsmokers and obtained a multiplicative interaction using the observed data of $\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_1,c]} / \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]} = 2.6$ (95% CI: 1.1, 6.1). The estimate itself and the confidence interval suggest a gene–environment interaction between passive smoking and glutathione S-transferase M1 (GSTM1) on lung cancer risk. The effect of smoking may, however, be confounded by air pollution. Poorer neighborhoods in which air pollution is high, say, may also have a higher prevalence of smoking or more extensive advertising for cigarettes. Suppose that the distribution of the genetic factor (GSTM1) is independent of both passive smoking and air pollution. By the robustness property, it would then follow from their estimate that either there is a true causal gene by passive smoking interaction or there is an interaction between the genetic factor and air pollution, either of which would constitute a gene–environment interaction.

11.3. SENSITIVITY ANALYSIS FOR THE RELATIVE EXCESS RISK DUE TO INTERACTION

As noted in Chapter 9, often in case–control studies, logistic regression is used to accommodate the case-control design. In such studies, if investigators want to assess interaction on the additive scale for public health purposes or to assess mechanistic interaction, then a measure referred to as the relative excess risk due to interaction (RERI) is sometimes used. The measure is also sometimes used when a logistic regression model is fit to the data out of convenience rather than by necessity due to a case–control design. The *RERI* conditional on c would generally be estimated by

$$\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} + 1$$

If the outcome is rare so that odds ratios approximate risk ratios, then each term $\mathbb{E}[Y|g,e,c]/\mathbb{E}[Y|g_0,e_0,c]$ can be approximated by the estimated odds ratio from the logistic regression.

Suppose that the effects of G and E on Y are not unconfounded conditional on C but would be unconfounded conditional on C and some unmeasured variable U. Define the causal *RERI* conditional on $C = c$ by the measure that would have been obtained if it had been possible to adjust for U as well.

Suppose that we make the simplifying assumption that U is binary and $\gamma = \frac{\mathbb{E}(Y|g,e,c,U=1)}{\mathbb{E}(Y|g,e,c,U=0)}$ is constant over g and e, then we have that the causal *RERI* would be given by

$$\frac{\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{1+(\gamma-1)P(U=1|g_1,e_1,c)}{1+(\gamma-1)P(U=1|g_0,e_0,c)}} - \frac{\frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{1+(\gamma-1)P(U=1|g_1,e_0,c)}{1+(\gamma-1)P(U=1|g_0,e_0,c)}} - \frac{\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{1+(\gamma-1)P(U=1|g_0,e_1,c)}{1+(\gamma-1)P(U=1|g_0,e_0,c)}} + 1$$

Here to carry out sensitivity analysis we would specify a parameter γ for the effect of U on Y and we would specify the prevalence of U, $P(U = 1|g_i, e_j, c)$, in each of the four $G \times E$ strata. We could then use the risk ratios, $\frac{\mathbb{E}[Y|g_i,e_j,c]}{\mathbb{E}[Y|g_0,e_0,c]}$, estimated from the data, our sensitivity analysis parameters, and the formula above to get a corrected estimate of *RERI* that we would have obtained had we been able to adjust for U.

Under $G \times E$ independence a simpler formula results. Suppose that we have $G \times E$ independence in the sense that the distribution of E and its confounder U were statistically independent of G conditional on the measured covariates C (as in Figure 11.2), that U is binary, and that $\gamma = \frac{\mathbb{E}(Y|g,e,c,U=1)}{\mathbb{E}(Y|g,e,c,U=0)}$ is constant over g and e, then we have that the causal *RERI* would be given by

$$RERI_c = \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} + 1$$

where

$$\kappa = \frac{1 + (\gamma - 1)P(U = 1|e_1, c)}{1 + (\gamma - 1)P(U = 1|e_0, c)}$$

The sensitivity analysis parameters now are just γ, the effect of U on Y, and the prevalence of U, $P(U = 1|e_1, c)$ and $P(U = 1|e_0, c)$, in each of the strata of the environmental factor. Once we specify the sensitivity analysis parameters, we could once again use the risk ratios, $\frac{\mathbb{E}[Y|g_i,e_j,c]}{\mathbb{E}[Y|g_0,e_0,c]}$, estimated from the observed data, our sensitivity analysis parameters, and the formula above to get a corrected estimate of *RERI* that we would have obtained had we been able to adjust for U.

Unlike the case for additive interaction using differences and for multiplicative interaction, for the sensitivity analysis for RERI, the confidence interval for the corrected RERI cannot simply be obtained by applying a formula to the confidence limits of the uncorrected RERI. Confidence limits for the corrected RERI could, however, still be obtained by bootstrapping. Finally, note that if the estimated *RERI* were found to be nonzero, then it would still follow from the robustness property for additive interactions in Section 11.1.2 that if we had $G \times E$ independence, then either there is a true causal $G \times E$ additive interaction or a $G \times U$ additive interaction.

Example. To illustrate the use of the sensitivity analysis approach, we will return to the Bangladesh data considered in Section 11.1 but we will examine a different interaction. We will apply the sensitivity analysis technique to the results of Ahsan et al. (2006), who examined the evidence for additive interaction between the effects of arsenic exposure in well water and BMI in producing premalignant skin lesions. Following their analysis, let $G = 1$ for high versus low arsenic (< 8 versus > 175 μg/L) and let $E = 1$ for low versus high BMI (< 18.1 versus > 20.4) with $Y = 1$ denoting the presence of premalignant skin lesions. Ahsan et al. (2006) adjust for gender, age, education, cigarette smoking, hukka smoking, sun exposure, and land ownership. Ahsan et al. (2006) used logistic regression to estimate the relative excess risk due to interaction in assessing potential additive interaction between BMI and arsenic exposure. Compared with the reference of

$G = 0, E = 0$, the odds ratio for $G = 1, E = 1$ was 5.25 (95% CI: 3.07, 8.99); for $G = 1, E = 0$, it was 2.96 (95% CI: 1.63, 5.37); and for $G = 0, E = 1$, it was 0.71 (95% CI: 0.38, 1.32). The overall prevalence of skin lesions is 6.3%, which is generally considered sufficiently small so that odds ratios approximate risk ratios. The estimated RERI was thus $5.25 - 2.96 - 0.71 + 1 = 2.59$ with a 95% confidence interval of (0.75, 4.24), suggesting evidence for positive additive interaction. Until the study was conducted, there was very little knowledge of which wells had high levels of arsenic; it is unlikely that the effects of arsenic are subject to substantial confounding. The correlation between arsenic exposure and other covariates is thus very weak, and the conditional association between BMI and arsenic exposure is not statistically significant in the sample. The effects of BMI on skin lesions are, however, likely confounded by, say, nutritional intake. By the robustness property for additive interaction in Section 11.1.2, because our estimated interaction is nonzero, we would have that either there is an interaction between arsenic and BMI or between arsenic and the confounders of the effect of BMI (e.g., nutritional intake). If we further wanted corrected estimates of the *RERI* between arsenic and BMI, we could use the sensitivity analysis technique presented above. Let U denote a hypothetical binary unmeasured confounder with $U = 1$ indicating high versus low nutritional intake. Suppose that high nutritional intake decreased the likelihood of skin lesions by threefold ($\gamma = 1/3$) for all strata of arsenic and BMI, with prevalence of high nutritional intake of 0.6 in those with high BMI and a prevalence of 0.2 in those with low BMI. We then have that

$$\begin{aligned}\kappa &= \frac{1 + (\gamma_1 - 1)P(U = 1|e_1, c)}{1 + (\gamma_0 - 1)P(U = 1|e_0, c)} \\ &= \frac{1 + (1/3 - 1)(0.2)}{1 + (1/3 - 1)(0.6)} = 1.44\end{aligned}$$

Under the rare outcome assumption that odds ratios approximate risk ratios, we would have a corrected *RERI* of

$$\begin{aligned}&\frac{1}{\kappa}\frac{\mathbb{E}[Y|g_1, e_1, c]}{\mathbb{E}[Y|g_0, e_0, c]} - \frac{\mathbb{E}[Y|g_1, e_0, c]}{\mathbb{E}[Y|g_0, e_0, c]} - \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_0, e_1, c]}{\mathbb{E}[Y|g_0, e_0, c]} + 1 \\ &= \frac{1}{1.44}5.25 - 2.96 - \frac{1}{1.44}0.71 + 1 = 1.19\end{aligned}$$

As an alternative scenario assuming weaker confounding, if high nutritional intake decreased the likelihood of skin lesions by twofold, with prevalence of 0.4 in those with high BMI and a prevalence of 0.2 in those with low BMI, we would have $\kappa = 1.13$ with a corrected *RERI* of 2.06.

11.4. MEASUREMENT ERROR AND ADDITIVE INTERACTION

11.4.1. Sensitivity Analysis and Robustness for Additive Interaction to Measurement Error

Misclassification and measurement error is an ubiquitous problem in studies of the effects of exposures. Earlier in this chapter we saw that often if the distributions of the two exposures were statistically independent of one another, then the interaction measures were robust to unmeasured confounding. Here we will see that somewhat similar results hold for tests for interaction. We will see that if the two exposures are independent, then misclassification will lead to conservative tests for additive interaction under measurement error. We will also see that similar results hold when testing for sufficient cause or "epistatic interaction" described in Chapters 9 and 10, and we will also provide a simple sensitivity analysis correction method that can be employed for measures of additive interaction when one or both of the exposures are subject to nondifferential misclassification.

As before, let G and E denote two binary exposures of interest, and let Y denote a binary outcome. Let G^* and E^* denote the potentially mismeasured exposures. Let $d_1 = P(G = 0|G^* = 1)$ and $d_2 = P(E = 0|E^* = 1)$ and let $u_1 = P(G = 1|G^* = 0)$ and $u_2 = P(E = 1|E^* = 0)$. These misclassification probabilities are equal to 1 minus the positive and negative predictive values, respectively, of the measured exposures for the true exposures. We say that the misclassification is nondifferential if $P(Y|G = i, E = j, G^* = l, E^* = m) = P(Y|G = i, E = j)$. Nondifferential misclassification implies that the measured exposures give no information about the outcome beyond the information given by the true exposures. We say that the misclassification is independent if the events of G and E being misclassified are independent—that is, that $P(G = i, E = j|G^* = l, E^* = m) = P(G = i|G^* = l)P(E = j|E^* = m)$. A sufficient condition for this is that $P(g^*, e^*|g, e) = P(g^*|g)P(e^*|e)$ with G and E independent. Let $p_{ge} = \mathbb{E}(Y|G = g, E = e)$ and let $p^*_{ge} = \mathbb{E}(Y|G^* = g, E^* = e)$.

The standard test for additive interaction is

$$p_{11} - p_{10} - p_{01} + p_{00} > 0$$

However, if we used the mismeasured exposures G^* and E^* and examined additive interaction using the mismeasured exposures, we would in fact be assessing $p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00} > 0$. Fortunately, under the assumption that misclassification is nondifferential and independent and if the misclassification probabilities are no larger than 1/2 (i.e., $d_i < 1/2$ and $u_i < 1/2$), then it can be shown that if the additive interaction with the mismeasured exposures is positive $p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00} > 0$, then the true measure of additive interaction is also positive: $p_{11} - p_{10} - p_{01} + p_{00} > 0$ (VanderWeele, 2012b). In a fairly early paper, Bross (1954) gave a result that implied that if a binary exposure had nondifferential misclassification, then tests of association between the outcome and the misclassified exposure constitute valid tests in that they provide conservative type I error rates of the association between the outcome and the true exposure. The result

just described for additive interaction with mismeasured exposures is essentially a generalization for interaction of the classic result of Bross (1954) for total effects.

Under independent and nondifferential misclassification, we can in fact conduct a simple sensitivity analysis as well using the following relation between the additive interaction with the true versus the misclassified exposures:

$$(p_{11} - p_{10} - p_{01} + p_{00}) = (p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00})/\{(1 - d_1 - u_1)(1 - d_2 - u_2)\}$$

For known misclassification probabilities (d_1, u_1, d_2, u_2), one can use the relation above to obtain estimates for additive interaction, corrected for exposure misclassification, by dividing additive interaction estimates using the observed data, $(p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00})$, by the factor $(1 - d_1 - u_1)(1 - d_2 - u_2)$. Confidence intervals for the true interaction could also be obtained by dividing both limits of the confidence interval for $(p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00})$ by the factor $(1 - d_1 - u_1)(1 - d_2 - u_2)$. For unknown misclassification probabilities, the parameters (d_1, u_1, d_2, u_2) could be varied in a sensitivity analysis. We will give an example of this approach below. The results above would also hold if the assumptions and probabilities were made conditional on measured covariates $C = c$.

The misclassification results above required an assumption of statistical independence of the distributions of the two exposures. This essentially required both that the exposures themselves were independent of one another and that the misclassification mechanism was independent. This might be plausible for a genetic and an environmental factor or for two genetic factors on different chromosomes. Genetic exposures generally have a low probability of misclassification; however, a genetic marker may serve as a proxy for a true causal variant and could be conceived of as a misclassified version of the causal variant with which it is associated due to linkage disequilibrium. The independence assumption may be less likely in settings with two environmental or behavioral or social exposure exposures, especially if assessed by a retrospective self-report. Whether the assumption is reasonable cannot be known definitively unless data are available on a gold standard measurement for the exposures.

For data with a dichotomous outcome, often out of convenience, logistic or log-linear models are fit; logistic models are also used to accommodate a case–control design. The results above have implications for such models. To assess additive interaction with case–control data, the relative excess risk due to interaction measure *RERI* is often used. Define $RERI^* = RR^*_{11} - RR^*_{10} - RR^*_{01} + 1$, where $RR^*_{ge} = p^*_{ge}/p^*_{00}$. Provided that the outcome is rare, we can approximate these risk ratios, RR^*_{ge}, by the odds ratios, $OR^*_{ge} = \{p^*_{ge}/(1 - p^*_{ge})\}/\{p^*_{00}/(1 - p^*_{00})\}$, estimated from the observed case–control data. The robustness of additive interaction tests also applies to *RERI* since if the mismeasured $RERI^*$ is greater than 0, then $(p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00})/p^*_{00} > 0$ and thus $(p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00}) > 0$ and so $(p_{11} - p_{10} - p_{01} + p_{00}) > 0$ and thus $RERI > 0$. It thus follows immediately from the results above that if $RERI^* > 0$ then $RERI > 0$. We can draw conclusions about additive interaction from the misclassified data using $RERI^*$. It is relatively straightforward to obtain confidence intervals for the relative excess risk

due to interaction as described in Chapter 9. Similar remarks hold for risk ratios conditional on covariates.

11.4.2. Robustness of Tests for Sufficient Cause and Epistatic Interaction to Measurement Error

In Chapters 9 and 10, it was noted that we could test for certain forms of "mechanistic" interaction such as sufficient cause interaction or epistatic interaction by testing for positive additive interaction if the monotonicity assumptions held. Without positive monotonicity assumptions we could still test for sufficient cause interaction or epistatic interaction by testing

$$p_{11} - p_{10} - p_{01} > 0$$

or

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

respectively. In fact, these tests, like that for the standard interaction contrast, are robust to nondifferential independent misclassification under very similar conditions. Suppose we use the misclassified exposures to test $p^*_{11} - p^*_{10} - p^*_{01} > 0$. Under the assumption that misclassification is nondifferential and independent, if $d_i < 1/2$ and $u_i < 1/4$, then it can be shown that if $p^*_{11} - p^*_{10} - p^*_{01} > 0$, then the true contrast is also positive, $p_{11} - p_{10} - p_{01} > 0$ (VanderWeele, 2012b). Likewise, suppose we use the misclassified exposures to test $p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00} > 0$. Under the assumption that misclassification is nondifferential and independent, if $d_i < 1/3$ and $u_i < 1/4$, then it can be shown that if $p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00} > 0$, then the true contrast is also positive $p_{11} - p_{10} - p_{01} - p_{00} > 0$ (VanderWeele, 2012b). The analogous conditions using the relative excess risk due to interaction are, as described in Chapters 9 and 10, $RERI > 1$ and $RERI > 2$. We have, under the same conditions above that $RERI^* > 1$ implies $RERI > 1$ and likewise that $RERI^* > 2$ implies $RERI > 2$. We can thus still test for sufficient cause interaction and epistatic interaction if we only have data on misclassified exposures provided that misclassification is non-differential and independent.

11.4.3. Examples of Additive Interaction Under Measurement Error

We illustrate the results with two examples. Canonico et al. (2008) tested for additive interaction between the presence of a non-O blood type and the use of oral estrogen on the risk of venous thromboembolism among postmenopausal women. They used data from a case–control study with 271 cases and 610 controls, adjusting for age, center, and admission date, and obtained a measure of $RERI = 5.4$ ($p < 0.05$). Data on blood type was self-reported and likely subject to misclassification; estrogen use is less likely to be misclassified. From the results above, under independence, and provided that the positive and negative predicted values for non-O blood group are at least 0.5 we could conclude, even from the analysis subject to misclassification, that there is true additive interaction.

VanderWeele et al. (2011) examined additive interaction between smoking and high well-water-arsenic exposure in producing premalignant skin lesions using a cohort of 11,746 persons in Bangladesh. VanderWeele et al. (2011) dichotomized arsenic at 100 μg/L and considered conclusions that could be drawn about the underlying continuous measure when dichotomizing an exposure. A measure of additive interaction for risk, using dichotomized arsenic and adjusting for covariates, was estimated to be 0.035 with 95% confidence interval $(0.0003, 0.070)$. Well water arsenic is subject to measurement error, so the dichotomized mismeasured value may thus constitute a misclassified version of the true dichotomized value. Self-reported smoking is also potentially subject to misclassification. Measured exposures have correlation close to 0, and misclassification of the two exposures is likely independent. Using the correction method in Section 11.4.1, if the positive and negative predictive values for smoking and dichotomized arsenic, conditional on covariates, were all 0.95, then we could divide our estimate and both limits of the confidence interval by $(1 - 0.05 - 0.05)(1 - 0.05 - 0.05) = 0.81$ to obtain a corrected additive interaction estimate of 0.043 with confidence interval $(0.0004, 0.086)$. If the positive predictive values for smoking and dichotomized arsenic were 0.85 and the negative predictive values were 0.93, then we could divide our initial estimate and both limits of the confidence interval by $(1 - d_1 - u_1)(1 - d_2 - u_2) = (1 - 0.07 - 0.15)(1 - 0.07 - 0.15) = 0.61$ to obtain a corrected estimate, which would be 0.058 with confidence interval $(0.0005, 0.115)$. Evidence for additive interaction remains, with somewhat higher estimates after adjusting for misclassification.

11.5. MEASUREMENT ERROR AND MULTIPLICATIVE INTERACTION

We will consider results on when multiplicative interaction is robust to measurement error. Like the results given for additive interaction, we will be essentially assuming here that the distributions of the two exposures, G and E, are statistically independent. This might be quite plausible in the context in which G and E are, respectively, genetic and environmental exposures, but it may not hold in many other contexts. However, the results given here for multiplicative interaction are somewhat different from those for additive interaction in the assumptions that are made. First, we will consider the misclassification of only one exposure, not both; second, we will allow for differential misclassification with respect to the outcome, whereas in the previous section on additive interaction we required nondifferential misclassification with respect to the outcome.

Suppose then that we are testing for multiplicative interaction on the odds ratios scale using logistic regression and that there are two exposures, G and E, and that G is correctly measured but that E is potentially misclassified. Let E^* denote the measured value of E. The logistic regression model for the actual exposures would be

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg$$

The model that is fit to the observed data would be

$$\text{logit}\{P(Y=1|G=g,E^*=e^*)\} = \gamma_0^* + \gamma_1^* g + \gamma_2^* e^* + \gamma_3^* e^* g$$

Garcia-Closas et al. (1998) show that if G and E are independent among the controls (i.e., when $Y=0$) and if the misclassification of E is nondifferential with respect to G in the sense that $P(E^*=e^*|E=e,G=g,Y=d)$ is independent of g—that is, G does not give any information about the likelihood of misclassification beyond E and Y—then if the product coefficient γ_3^* in the logistic regression using the potentially mismeasured variable is nonzero, then the measure of multiplicative interaction between the true variables, γ_3, will also be nonzero. Under these assumptions, we can thus use the potentially mismeasured E^* to test for multiplicative interaction. The result holds if E is binary or ordinal.

This result makes two assumptions. First, that G and E are independent among the controls (i.e., when $Y=0$). This assumption may not hold exactly. However, if the outcome is rare, then this will hold approximately provided that G and E are independent in the population. We could essentially replace the assumption that the distributions of G and E are statistically independent among the controls with the assumptions that G and E are independent in the population and that the outcome is rare. Second, the result makes the assumption that the misclassification of E is nondifferential with respect to G. Provided that the second exposure G does not affect whether the first exposure is misclassified, this assumption will be reasonable. Note that the result does allow (unlike the results given above for additive interaction) misclassification with respect to Y to be differential. The outcome Y can affect the likelihood that E is misclassified (as may occur in case–control studies with exposures obtained by retrospective self-report) and the result will still hold. The result, however, does require that one of the two exposures be correctly measured.

The result just described will hold if E is binary or ordinal. If E is binary, then somewhat stronger conclusions are also possible. In particular, if G and E are independent among the controls (or independent in the population with a rare outcome) and if the misclassification of E is nondifferential with respect to G, then the product coefficient γ_3^* in the logistic regression using the potentially mismeasured variable will be smaller in magnitude than the measure of multiplicative interaction between the true variables, γ_3, that is, $|\gamma_3^*| \leq |\gamma_3|$. In other words, under these assumptions of independence and nondifferential misclassification, the estimated multiplicative interaction γ_3^* will be biased toward the null: our conclusions about the magnitude of multiplicative interaction will be conservative.

Garcia-Closas et al. (1998) further show that it is to a certain extent possible to assess the assumptions being made. They show that the assumptions that G and E are independent among the controls along with the assumption that misclassification for E is nondifferential with respect to G together imply that E^* and G are independent amongst the controls. To assess the assumptions, we could thus test whether E^* and G are independent amongst the controls. They show that if we do not reject then a valid test for multiplicative interaction between E and G is testing whether E^* and G are independent amongst the cases. It turns out that these two tests are independent of one another (they use different subjects) so that we don't

need to adjust the second test for the results of the first. It is, however, possible that mismeasured E^* and G are independent amongst the controls without it being the case that true E and G are independent among the controls along and that misclassification for E is nondifferential with respect to G. We can use the procedure above to falsify the assumptions being made but not to fully verify them.

We have seen in this section and the previous section that interaction tests for either additive or multiplicative interaction are robust to various forms of exposure misclassification, provided that the exposures are independent in distribution in the population. The assumption of independence of the distributions of the exposures is quite important in ensuring the robustness of interaction tests. If the two exposures are correlated, as would be the case if one affected the other, then results from interaction analyses can be highly biased (e.g., Greenland, 1980). It is thus very important to evaluate the assumption that the exposures are independent in distribution when assessing the consequences of measurement error in interaction analyses.

11.6. DISCUSSION

In this chapter we have provided several methods for sensitivity analysis for unmeasured confounding in studies of interaction. We have considered both the additive and multiplicative scales along with additive interaction obtained from a multiplicative model (the relative excess risk due to interaction). The techniques can be used in a wide range of interaction studies and can be applied across numerous different study designs. See also Cheng and Lin (2009), Lindström et al. (2009), Tchetgen Tchetgen and Kraft (2011), and Tchetgen Tchetgen and VanderWeele (2014a) for related discussion. In many studies of gene–environment interaction, the distributions of the genetic and environmental factors are assumed to be statistically independent. We have seen above that, under this assumption, interaction findings are particularly robust to unmeasured confounding insofar as if we are concerned about unmeasured confounding of the environmental factor by another unmeasured environmental exposure and if, with the observed data, we find interaction, then either there must be a true causal interaction between the genetic and environmental factor or there is interaction between the genetic factor and the unmeasured environmental confounding variable; in either case, we have gene–environment interaction.

In this chapter we have also examined the consequences of measurement error in assessing interaction. We have seen that in a number of scenarios, measurement error does not invalidate tests for either additive or multiplicative interaction under the assumption that the distribution of the two exposures are statistically independent in the population. This assumption is again often plausible in gene–environment interaction studies, but it may not be plausible in settings in which both exposures are environmental or with social or behavioral exposures. Further methodological approaches still need to be developed to address issues of measurement error in interaction analysis when the exposures of interest are not independent of each other.

For the most part the material we have discussed in this chapter constitutes good news for interaction analysis. For both confounding and measurement error, we have seen that interaction estimates and tests are often robust to, or conservative under, these potential biases, even in settings in which main effects themselves are biased. Moreover, we can use sensitivity analysis to assess plausible ranges of interaction estimates under different assumptions about the magnitude of these biases.

12

Interaction in Genetics: Independence and Boosting Power

In this chapter we will consider a number of further issues in the analysis of interaction which are especially important in genetics. Most of the methods described here are not restricted to the genetics setting but are especially relevant to and, for the most part, developed out of, the genetics literature. Many of the methods described here are concerned with issues of power. As will be discussed further in Chapter 13, tests for interaction often require very substantial sample sizes to have much power. This issue of power is moreover further compounded in genetics studies in which hundreds or thousands of different interactions are being tested simultaneously, raising the need to adjust for multiple testing. This further decreases power. Many of the methods described in this chapter attempt to boost power. However, doing so generally requires that further assumptions be made, and these assumptions are not always plausible. The principal assumption relied on in many of these methods is that the distributions of the two exposures of interest are statistically independent in the populations conditional on the observed covariates. This assumption is plausible in many, though not all, genetic contexts, but it may be less reasonable in other settings. This independence assumption is often plausible in genetic contexts because genetic factors are essentially randomized conditional on the parents' genetic factors and will themselves only affect one particular biological mechanism or pathway; the distribution of the genetic factors themselves will thus often be independent of most, but not all, environmental factors. In any case, when the distributions of the two exposures of interested are statistically independent, many of the methods in this chapter will be applicable.

12.1. CASE-ONLY ESTIMATORS OF INTERACTION

It has become increasingly popular in genetics to use what is sometimes referred to as a case-only estimator of interaction. It turns out that under an assumption

of independence between the two exposures of interest, measures of multiplicative interaction can be obtained by using data only among the cases. This has advantages in terms of the amount of data that needs to be collected, but it has even further advantages insofar as this approach also generally boosts power by exploiting this independence assumption.

Consider the statistical interaction β_3 in the log-linear model

$$\log\{P(Y=1|G=g,E=e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg$$

Suppose now also that the distributions of the two exposures, G and E, are statistically independent in the population. This assumption may be plausible in many gene–environment interaction studies. Suppose further that data are only collected on the cases $(Y=1)$. It can be shown that under this independence assumption, the odds ratio relating G and E among the cases is equal to the interaction measure on the multiplicative scale β_3 (Yang et al. 1999; cf. Piegorsch et al., 1994):

$$\frac{P(G=1|E=1,Y=1)/P(G=0|E=1,Y=1)}{P(G=1|E=0,Y=1)/P(G=0|E=0,Y=1)} = \frac{RR_{11}}{RR_{10}RR_{01}} = \beta_3$$

Essentially to get measures of multiplicative interaction, all that is needed is data on G and E among the cases. The use of the odds ratio relating G and E among the cases is referred to as the "case-only" estimator of interaction. With the case-only estimator we can estimate the interaction parameter β_3, but we cannot estimate the main effects of the log-linear regression, β_1 and β_2. Furthermore, the case-only estimator depends critically on the assumption that the distributions of the genetic and environmental factors are statistically independent and can be quite biased if this assumption is violated (Albert et al., 2001). However, under this assumption that distributions of the genetic and environmental factors are statistically independent, the case-only estimator is in fact more efficient than using the standard estimate from a log-linear regression (Yang et al., 1997).

The same result holds for statistical interaction in logistic regression

$$\text{logit}\{P(Y=1|G=g,E=e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg$$

under the assumption that the outcome is rare (Piegorsch et al., 1994). The result for log-linear models does not require a rare outcome. Sometimes, for logistic regression, the independence assumption is articulated as one of independence of the distributions of G and E among the non-cases. For a rare outcome, this is approximately equivalent to independence in the distributions in the population.

The result also holds for log-linear or logistic regression if we control for covariates. The conditional independence assumption is then that G and E are independent conditional on C. Estimates and confidence intervals for the case-only estimator can be obtained by running a logistic regression of G on E and C among the cases:

$$\text{logit}\{P(G=1|E=e,C=c,Y=1)\} = \theta_0 + \theta_1 e + \theta_2' c$$

Table 12-1. NUMBER OF CASES BY GENOTYPE AND SMOKING STATUS (BENNETT ET AL., 1999)

	No Smoking	**Smoking**
GSTM1 present, $G = 0$	28	14
GSTM1 absent, $G = 1$	27	37

The coefficient and confidence interval for θ_1 in this regression on the cases will equal that of the product term coefficient β_3 in the log-linear model

$$\log\{P(Y = 1|G = g, E = e, C = c)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg + \beta_4' c$$

provided that the distributions of G and E are statistically independent in the population and will also approximately equal that of γ_3 in the logistic model

$$\text{logit}\{P(Y = 1|G = g, E = e, C = c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c$$

provided, in addition, that the outcome is rare.

Example. Bennett et al. (1999), for example, use data on non-smoking lung cancer cases and report exposure status for GSTM1 genotype and passive smoking as in Table 12.1.

Using data only on the cases, we have that the estimate of multiplicative interaction is

$$\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{P(G = 1|E = 1, Y = 1)/P(G = 0|E = 1, Y = 1)}{P(G = 1|E = 0, Y = 1)/P(G = 0|E = 0, Y = 1)} = \frac{37/14}{27/28} = 2.74$$

When adjusted also for age, radon exposure, saturated fat intake, and vegetable intake using logistic regression, the case-only estimate of multiplicative interaction is 2.6 (95% CI: 1.1–6.1). There is evidence here for multiplicative interaction between passive smoking and the absence of GSTM1 on lung cancer.

As discussed in Chapter 10, we can sometimes use multiplicative interaction to assess mechanistic interaction as well. The tests described in Chapter 10 apply to the case-only estimator as well. In particular, under the assumption that the distributions of the two exposures are statistically independent in the population and unconfounded conditional on the measured covariates, a sufficient cause interaction is present if $\theta_1 > 0$ and both exposures have positive monotonic effects on the outcome (i.e., the exposure are never preventive for any individual), or if $\theta_1 > \log(2)$ without any monotonicity assumptions provided that the main effects of both exposures (which cannot in fact be estimated using case-only data) are non-negative. Also, as discussed in Chapter 10, an epistatic interaction is present if $\theta_1 > 0$ and both exposures have positive monotonic effects, or, provided that the main effects of both exposures are non-negative, if $\theta_1 > \log(2)$ and at least one of the two exposures has a positive monotonic effect, or if $\theta_1 > \log(3)$ without any monotonicity assumptions.

12.2. JOINT TESTS FOR INTERACTIONS AND MAIN EFFECTS

12.2.1. Varieties of Joint Main Effect and Interaction Tests

In some genetic studies, interest lies not so much in interaction but in increasing power to detect an effect of various genetic variants on the outcome of interest. Sometimes, potential gene–environment interaction is used to attempt to boost the power of tests to detect associations of genetic variants with the outcome (Chatterjee et al., 2006; Kraft et al., 2007; Maity et al., 2009; Dai et al., 2012). This typically involves testing jointly for the presence of a genetic main effect and a gene–environment interaction or alternatively a test of marginal association combined with a test for gene–environment interaction.

These tests typically proceed by specifying a model for the association between the disease outcome and the genetic and environmental factors allowing for gene–environment interaction. For example, if logistic regression is used, this model would take the form

$$\text{logit}\{P(Y=1|G=g,E=e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

The joint test would then be a test of the joint null hypothesis that both the genetic main effect and the gene–environment interaction effect are zero—that is, that $\gamma_1 = \gamma_3 = 0$. This null hypothesis might then be tested using a likelihood ratio test. It has been shown that such a joint test has more power to detect genetic effects than does a marginal test of association between disease and the genetic variant, over a broad range of—though not all—scenarios (Kraft et al., 2007).

Alternatives to the joint test have been proposed to detect associations of a genetic factor and the outcome by instead using joint tests of marginal association with disease in one logistic regression model, along with gene–environment interaction in a second logistic regression model (Dai et al., 2012). Two logistic regression models are thus used. First, to test the marginal genetic effect $\alpha_1 = 0$, a model is used for the gene–outcome association:

$$\text{logit}\{P(Y=1|G=g)\} = \alpha_0 + \alpha_1 g$$

Second, the previous logistic model is also used for gene–environment interaction $\gamma_3 = 0$. It turns out that the Wald chi-squared test statistic for testing marginal genetic effect, $\alpha_1 = 0$, and gene–environment interaction, $\gamma_3 = 0$, are independent (Dai et al., 2012), and they can they can be combined to form chi-squared two-degree-of-freedom test of the following form:

$$T = \frac{\widehat{\alpha}_1^2}{\text{Var}(\widehat{\alpha}_1)} + \frac{\widehat{\gamma}_3^2}{\text{Var}(\widehat{\gamma}_3)}$$

where $\widehat{\alpha}_1$ and $\text{Var}(\widehat{\alpha}_1)$, and $\widehat{\gamma}_3$ and $\text{Var}(\widehat{\gamma}_3)$, are the estimators and variances for α_1 and γ_3, respectively.

Further modifications of this test were also proposed in Dai et al. (2012) by using alternative estimators for the gene–environment interaction parameter γ_3

that make use of the assumption of gene–environment independence. Several of these alternatives include (a) the case-only (CO) estimator described in the previous section and (b) an empirical Bayes (EB) method proposed by Mukherjee and Chatterjee (2008; cf. Mukherjee et al., 2012). In these cases the standard interaction estimator $\widehat{\gamma}_3$ is replaced by either (a) the case-only estimator of interaction, $\widehat{\theta}_1$ in the previous section or (b) the empirical Bayes estimator $\widehat{\gamma}_{EB}$ of interaction. The independence of the two test statistics still holds (Dai et al., 2012), and the marginal test and different gene–environment interaction tests can be combined to give rise to additional alternative to the two-degrees of freedom joint tests using either the case-only estimator (T_{CO}) or the empirical Bayes estimator (T_{EB}):

$$T_{CO} = \frac{\widehat{\alpha}_1^2}{\mathrm{Var}(\widehat{\alpha}_1)} + \frac{\widehat{\theta}_1^2}{\mathrm{Var}(\widehat{\theta}_1)}$$

$$T_{EB} = \frac{\widehat{\alpha}_1^2}{\mathrm{Var}(\widehat{\alpha}_1)} + \frac{\widehat{\gamma}_{EB}^2}{\mathrm{Var}(\widehat{\gamma}_{EB})}$$

The approach using the standard multiplicative interaction estimator $\widehat{\gamma}_3$ will be robust to correlation between G and E, but has less power than the other approaches. On the other hand, the case-only method provides substantial power gain under gene–environment independence assumption but is invalid under violations of this assumption. The empirical Bayes method is a hybrid compromise that combines the standard interaction estimator and the case-only estimator with data-adaptive weights that optimally trade off between bias and efficiency under departures from the independence assumption. The empirical Bayes approach can provide substantial power advantages compared to using the standard interaction estimator and better controls the Type I error rate than when using case-only estimator when the exposures are not independent, but is still somewhat biased in this scenario. When compared to just testing for a marginal association alone ($\alpha_1 = 0$), all of these methods can increase the power to detect associations between genetic variants and the outcome.

12.2.2. Environmental Confounding and Joint Tests

The use of the methods for joint tests do, however, suffer some problems that the tests for marginal association do not. Specifically, there are certain confounding scenarios in which the marginal test is robust but the joint tests are not. In genetic studies, effort is often made to control for population stratification so that associations between genetic variants and disease are not due to confounding by race–ethnicity. Less attention, however, is generally given to the possibility of environmental confounding in these studies of gene–environment interaction.

What happens then to the joint tests under environmental confounding? Let us first suppose that the genetic variant affects the environmental factor itself and that there is an unmeasured confounding variable of the relationship between the environmental factor and the disease outcome, as in Figure 12.1.

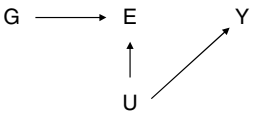

Figure 12.1 Example illustrating that joint tests of interaction and main genetic effects can lead to invalid conclusions in the presence of environmental confounding (U).

Suppose that neither the genetic variant nor the environmental factor has any effect on the disease itself. In this case, the genetic variant and the disease outcome will be unassociated marginally. A test for marginal association between the genetic factor and the disease will have valid Type I error. What happens to the joint test? Unfortunately, under environmental confounding, the joint test at significance level α will in general reject the null with far greater frequency than the nominal significance level. This can be demonstrated in simulations (VanderWeele et al., 2013b), but it also possible to see why this is so analytically.

In the setting of Figure 12.1, G and Y are unassociated marginally but both are also associated with E marginally. If two binary variables are unassociated marginally, and both are marginally associated with a third binary variable, then the two binary variables will be conditionally associated with each other within at least one stratum of the third variable; thus in Figure 12.1, G and Y will be associated conditionally within at least one stratum of E. The phenomenon is sometimes described as "conditioning on a common effect" or "collider stratification" (Hernán et al., 2004; VanderWeele and Robins, 2007a; Cole et al., 2010): Two variables, even if marginally uncorrelated, will in general be correlated conditional on the common effect. Suppose, for instance, that the mechanism for E in Figure 12.1 is that E occurs if at least one of U or G is present. Although U and G may be uncorrelated in the population, conditioning on $E = 1$, say, will induce correlation because if, for a particular subject, we had that $E = 1$ and $U = 0$, then we would know that G must be 1 for that subject since E occurs only if at least of U and G are 1. Likewise, for a subject $E = 1$ and $G = 0$ we would know that $U = 1$. The variables U and G will thus be correlated conditional on E.

The implications of this for the logistic regression model which includes both G and E as above, under Figure 12.1, is that, without controlling for U, at least one of γ_1 or γ_3 will be nonzero. In other words, in large samples, the joint test will reject the null hypothesis $\gamma_1 = \gamma_3 = 0$ even though neither G nor E has any effect on Y. This occurs because of the environmental confounder U for which control has not been made. And, unfortunately, this problem gets worse as the sample size increases. Because the value of either γ_1 or γ_3 is nonzero (having not controlled for U), the joint test will reject the null $\gamma_1 = \gamma_3 = 0$ with a probability tending toward 1 as the sample size increases. In Figure 12.1, the marginal test for G would be valid, but the joint tests using T or T_{CO} or T_{EB} will all be biased. If we could control for the environmental confounder U in the analysis, our joint test would be valid for detecting genetic effects; without such control we get an inconsistent test.

Consider now Figure 12.2 in which the genetic and environmental factors are marginally independent in the population. In this case, under the null hypothesis

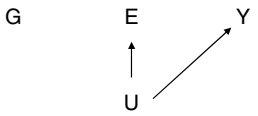

Figure 12.2 Example illustrating that the joint test is protected against environmental confounding (U) when the genetic factor G does not affect the environmental factor E.

that G has no effect on Y, the joint test is protected against unmeasured environmental confounding under the null. The variable U may induce correlation between E and Y, even if E itself has no effect on Y. However, conditional on E, if G has no effect on E and no effect on Y, then G will remain uncorrelated with Y within all strata of E. Thus under the logistic regression model, if G has no effect on E and no effect on Y, then both γ_1 and γ_3 will be zero. The joint test of a main genetic effect and a gene–environment interaction (e.g., that $\gamma_1 = \gamma_3 = 0$) will maintain valid Type I error under the null hypothesis that there is no effect of the genetic variant on the disease, even in the presence of unmeasured environmental confounding, provided that we have marginal gene–environment independence.

The joint tests can increase the power of detecting overall associations between genetic variants and various outcomes, but this increased power comes at a price: This price is robustness to environmental confounding. If it is known a priori that the distributions of the genetic and environmental factors are statistically independent, then the approach of using joint tests can increase power, but if this assumption does not clearly hold, then it may be best to avoid joint tests and use instead only marginal tests of the overall association between the genetic variant and the outcome, because of the bias that can then be introduced by environmental confounding when using joint tests. Note that if the gene–environment independence assumption is thought to hold, then the joint test that combines the marginal test with the case-only interaction estimator, T_{CO}, will generally have the most power.

12.3. MULTIPLE TESTING

In the previous two sections we discussed methods to exploit statistical independence between the distributions of two exposures in order to boost power. As was already noted, often in the genetic context, data on many different genetic markers are available and thus there is potential for testing for multiple different gene–gene or gene–environment interactions. At present, sometimes data on up to a million different genetic markers or SNPs are available. There have been proposals to examine gene–gene and gene–environment interaction throughout all of the available data and to look at interactions "genome-wide" (Kraft, 2004; Gayan et al., 2008; Khoury and Wacholder, 2009; Murcray et al., 2009; Pierce and Ahsan, 2010; Thomas, 2010). In general with testing for interaction, the Type I error rate is controlled at 5% so that the statistic used to test for interaction exceeds the critical

value only 5% of the time if there is in fact no interaction. Thus if there is no interaction, there is only a 5% chance of concluding that one is in fact present. However, if one were to test for two different interactions at the 5% significance level, then the likelihood would be rather higher than 5% that one concluded the presence of an interaction when in fact neither interaction was present. If there were twenty tests for interaction conducted, then one would expect on average to conclude by chance that one of the interactions was present, even if there were in fact no interactions. If there were four hundred tests for interaction conducted, then one would expect on average to conclude by chance that twenty of the interactions were present, even if there were in fact no true interactions. In such settings an investigator would potentially be frequently drawing false conclusions ("false positives") about the presence of interactions.

To protect against this, various multiple testing procedures have been proposed to address this issue. The basic idea is to modify the threshold or significance level for the individual tests so that the overall significance level or Type I error rate is still 5%—that is, so that there is only a 5% chance or less of concluding there is an interaction when in fact there are none. The most straightforward multiple testing method is generally referred to as the Bonferroni correction. If the desired overall significance level is α (e.g., 5%) and if there are K tests that are going to be conducted, then the Bonferroni correction is to use α/K as the significance level for the individual tests. Thus if two tests are going to be conducted, one would use $5\%/2 = 2.5\%$ as the significance level for the individual tests. If three tests are going to be conducted, one would use $5\%/3 = 1.67\%$ as the significance level for the individual tests. If 400 tests are going to be conducted, one would use $5\%/400 = 0.0125\%$ as the significance level for the individual tests; that is, one would require a p-value of less than 0.000125 to reject the null. If one million tests were conducted, one would use $5\%/1,000,000$ as the significance level for individual tests; that is, one would require a p-value of 5×10^{-8} in order to reject the null. This is the threshold that is generally required in genome-wide testing. Clearly, carrying out such multiple testing correction can lead to very extreme thresholds. However, if this is done, then the overall significance level for all tests combined will be no more than 5%. A closely related procedure is what is sometimes called the Sidak correction, which, if one is conducting K tests and wants to maintain an overall significance level of α for all tests jointly, then one takes $1 - (1 - \alpha)^{1/K}$ as the significance level of the individual tests. Once again using this as the significance level for individual tests will preserve an overall significance level of α. The Sidak correction and the Bonferroni correction will give very similar thresholds; and because the Bonferroni correction method is easier to calculate, it is more often used in practice.

These correction methods are sometimes criticized because they can lead to quite conservative thresholds. This is especially the case if the exposures that are used in the tests are highly correlated with one another. Various proposals have been put forward to attempt to come up with somewhat less stringent thresholds when exposures are correlated. There is a large literature on this topic, and we will not pursue it here. In practice, in the genetics literature at present, the Bonferroni method tends to be used most frequently.

12.4. DISCUSSION

In this chapter we have covered a number of topics relevant to testing for interactions in genetic settings. Several of these approaches concern issues of power. Generally, power is boosted with the assumption that the distributions of the genetic and environmental exposures are independent in the population. Here as elsewhere in this book we have focused on methods that are relatively easy to implement, but a number of other methods that exploit the assumption of independence in the distributions of the exposures have also been proposed which are slightly or substantially more powerful for detecting additive or multiplicative interactions (Chatterjee and Carroll, 2005; Vansteelandt et al., 2008; Han et al., 2012) but are more difficult to implement. The material in this chapter, although perhaps most relevant to genetic studies, could also be employed in other settings. Again the chief restriction of many of these methods is the assumption of independence in the distributions of the exposures which is often more plausible in examining gene–environment interaction than in other settings.

A number of other, more specialized topics related to interaction in genetics, which have not been covered here, depend even more crucially on the genetics context. For example, if data are collected on an individual and both parents, or on siblings, then often it is possible to construct tests for main genetic effects and gene–environment interactions that are much more robust to confounding than the tests that are usually carried out (Umbach and Weinberg, 2000; Lake and Laird, 2004; Hoffmann et al., 2009; Weinberg et al., 2011). These so-called family-based designs and tests, however, do crucially depend on the genetics context in which an individual's genotype is randomized conditional on that of their parents.

13

Power and Sample-Size Calculations for Interaction Analysis

In this chapter we will present power and sample-size calculations for interaction analyses. Such calculations are important in the planning of studies. If we desire to have a certain power to be able to detect a particular interaction, we may use such calculations to determine how large the study sample must be in order to do so. In many other cases, a study may have been designed to detect a main effect and the sample size fixed accordingly. With this fixed sample size, we may still be interested in the power that we have in a study to detect an interaction of a certain magnitude. The formulae and tools in this chapter allow for such calculations under a range of scenarios. We will first begin with power and sample size calculations for continuous outcomes. We will then move to the setting of binary outcomes and give power and sample size calculations for binary outcomes on a multiplicative scale using cohort, case–control, or case-only data. After this we will continue with the setting of binary outcomes but will give power and sample-size calculations for additive interaction which, as we have seen, is often more relevant for evaluating the impact of and deciding between interventions; we will give these power and sample-size calculations for additive interaction for both cohort and case–control data. Finally, we will also present power and sample-size calculations for the mechanistic interaction tests that were considered in Chapters 9 and 10. In all of these cases, we will present the general formulae for power and sample-size calculations, and then in Section 13.5 we will discuss the use of Excel spreadsheets that will carry out all of these power and sample-size calculations automatically. As will be seen throughout, power is generally much lower to detect an interaction than a main effect. Studies need to be quite large in order to have adequate power to reliably detect interaction.

13.1. POWER AND SAMPLE-SIZE CALCULATIONS FOR INTERACTION FOR CONTINUOUS OUTCOMES

In this section we will present power and sample-size calculation for continuous outcomes when one is interested in assessing interaction on the additive scale. In subsequent sections we will consider binary outcomes on both additive and multiplicative scales. We will let $\pi_{ge} = P(G = g, E = e)$ denote the proportion of subjects in each $G \times E$ category.

Suppose we fit a linear regression model:

$$\mathbb{E}[Y|G = g, E = e] = \tau_0 + \tau_1 g + \tau_2 e + \tau_3 ge$$

As discussed in Chapter 9, the coefficient τ_3 is the interaction on the additive scale. Suppose we wish to use a Wald test for the null hypothesis $\tau_3 = 0$. The sample size required to detect an interaction of magnitude $\tau_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{cts}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively (e.g. for a test at the 5% significance level with 80% power, these would be 1.96 and 0.84 respectively), of the standard normal distribution and where V_{cts} is the variance of $\widehat{\tau}_3$ under the alternative that $\tau_3 = \eta$. This is given by

$$V_{cts} = \sigma^2 \left(\frac{1}{\pi_{00}} + \frac{1}{\pi_{10}} + \frac{1}{\pi_{01}} + \frac{1}{\pi_{11}} \right)$$

where σ^2 is the variance of the error term in the regression model for Y—that is, the variance of Y conditional on G and E.

To calculate the sample size, we would need to specify (i) the significance level α, the power β, and the magnitude of the interaction $V_{cts} = \sigma^2(\frac{1}{\pi_{00}} + \frac{1}{\pi_{10}} + \frac{1}{\pi_{01}} + \frac{1}{\pi_{11}}) = \eta$ and (ii) the proportion of subjects in each exposure stratum, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$.

If instead of calculating the required sample size for a fixed power β, we wanted to calculate the power for a given sample size using the Wald test for the null hypothesis $\tau_3 = 0$ based on the linear regression model, we could proceed as follows. For a fixed sample size n the power to reject the null $\tau_3 = 0$ at significance level α under the alternative that $\tau_3 = \eta$ is given by

$$\text{Power} = \Phi^{-1}\left\{ -Z_{1-\alpha/2} + \eta\sqrt{(n/V_{cts})} \right\}$$

where Φ^{-1} is the inverse cumulative distribution function for a standard normal random variable and where V_{cts} can be calculated as above.

Finally, if the null hypothesis were rejected for extreme values of τ_3 on either side of zero (two-sided test), then the relevant power formula would be

$$\text{Power} = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{cts})}\right\} + \Phi^{-1}\left\{-Z_{1-\alpha/2} - \eta\sqrt{(n/V_{cts})}\right\}.$$

13.2. POWER AND SAMPLE-SIZE CALCULATIONS FOR BINARY OUTCOMES: MULTIPLICATIVE INTERACTION

In this section we will present power and sample-size calculation for binary outcomes when one is interested in assessing interaction on a multiplicative scale using odds ratios. We will let $p_{ge} = P(Y = 1|G = g, E = e)$ be the probability of the outcome in each of the $G \times E$ categories and we will let $\pi_{ge} = P(G = g, E = e)$ denote the proportion of subjects in each $G \times E$ category. As in Chapter 9, we define odds ratios as $OR_{ge} = \frac{P(Y=1|G=g,E=e)/P(Y=0|G=g,E=e)}{P(Y=1|G=0,E=0)/P(Y=0|G=0,E=0)} = \frac{p_{ge}/(1-p_{ge})}{p_{00}/(1-p_{00})}$ and define the measure of multiplicative interaction on the odds ratio scale as $I_{OR} = \frac{OR_{11}}{OR_{10}OR_{01}}$, respectively. This measure of multiplicative interaction corresponds to the exponentiated coefficient of the product term for the two exposures in a logistic regression models for the outcome.

13.2.1. Multiplicative Interaction with Cohort Data

Suppose we fit a logistic regression model to cohort data:

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

The coefficient γ_3 is generally referred to as a measure of interaction of the multiplicative scale. The exponentiated coefficient is equal to the odds ratio multiplicative interaction $e^{\gamma_3} = I_{OR} = \frac{OR_{11}}{OR_{10}OR_{01}}$. Suppose we wish to use a Wald test for the null hypothesis $\gamma_3 = 0$. The sample size required to detect a multiplicative interaction of magnitude $\gamma_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{mult(OR)}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where $V_{mult(OR)}$ is the variance of $\widehat{\gamma}_3$ under the alternative that $\gamma_3 = \eta$. This is given by (Demidenko, 2008)

$$V_{mult(OR)} = \frac{1}{L} + \frac{1}{F} + \frac{1}{J} + \frac{1}{R}$$

where

$$L = \frac{e^{\gamma_0}}{(1+e^{\gamma_0})^2}\pi_{00}$$
$$F = \frac{e^{\gamma_0+\gamma_1}}{(1+e^{\gamma_0+\gamma_1})^2}\pi_{10}$$
$$J = \frac{e^{\gamma_0+\gamma_2}}{(1+e^{\gamma_0+\gamma_2})^2}\pi_{01}$$
$$R = \frac{e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}{(1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3})^2}\pi_{11}$$

To calculate the sample size, we would need to specify (i) the significance level α, the power β, and the magnitude of multiplicative interaction $\gamma_3 = \eta$; (ii) the proportion of subjects in each exposure stratum, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$; and (iii) the main effect odds ratios of the two exposures on the logistic scale, γ_1 and γ_2, and the log odds of the baseline risk of the doubly unexposed group $\gamma_0 = \log\{P(Y = 1|G = 0, E = 0)/P(Y = 0|G = 0, E = 0)\}$. If instead of specifying the joint probabilities $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$, we specified the marginal probabilities of each exposure $\pi_g = P(G = 1)$ and $\pi_e = P(E = 1)$ and the odds ratio relating G and E, $\Delta = \{P(G = 1|E = 1)/P(G = 0|E = 1\}/\{P(G = 1|E = 0)/P(G = 0|E = 0\}$, then we could obtain the $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ using (Demidenko, 2008)

$$\pi_{00} = \frac{1-\pi_e}{1+C}$$
$$\pi_{10} = \frac{(1-\pi_e)C}{1+C}$$
$$\pi_{01} = \frac{\pi_e}{1+C\Delta}$$
$$\pi_{11} = \frac{C\Delta\pi_e}{1+C\Delta}$$

where

$$C = \frac{q+\sqrt{q^2+4\pi_g(1-\pi_g)\Delta}}{2(1-\pi_g)\Delta}$$
$$\text{and} \quad q = \pi_g(1+\Delta)+\pi_e(1-\Delta)-1$$

If G and E are independent, then $\Delta = 1$ and C simplifies to $C = \pi_e/(1-\pi_e)$.

If instead of calculating the required sample size for a fixed power β, we wanted to calculate the power for a given sample size using the Wald test for the null hypothesis $\gamma_3 = 0$ based on the logistic regression model, we could proceed as follows. For a fixed sample size n the power to reject the null $\gamma_3 = 0$ at significance level α under

the alternative that $\gamma_3 = \eta$ is given by

$$\text{Power} = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{mult(OR)})}\right\}$$

where Φ^{-1} is the inverse cumulative distribution function for a standard normal random variable and where $V_{mult(OR)}$ can be calculated as above. In Section 13.5 we will describe how to use a simple Excel spreadsheet to carry out such sample size and power calculations automatically.[1]

Finally, it should be noted that if the null hypothesis were rejected for extreme values of γ_3 on either side of zero (two-sided test), then the relevant power formula would be

$$\begin{aligned}\text{Power} = {} & \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{mult(OR)})}\right\} \\ & + \Phi^{-1}\left\{-Z_{1-\alpha/2} - \eta\sqrt{(n/V_{mult(OR)})}\right\}\end{aligned}$$

Note that we may sometimes want to specify the probabilities of the outcome, $p_{ge} = P(Y = 1|G = g, E = e)$, in each of the $G \times E$ categories (i.e., $p_{00}, p_{01}, p_{10}, p_{11}$), rather than the magnitude of multiplicative interaction $\gamma_3 = \eta$, the main effect odds ratios of the two exposures on the logistic scale, γ_1 and γ_2, and the log odds of the baseline risk of the doubly unexposed group $\gamma_0 = \log\{P(Y = 1|G = 0, E = 0)/P(Y = 0|G = 0, E = 0)\}$. If we wanted to carry out power or sample-size calculations while specifying the probabilities of the outcome, $p_{00}, p_{01}, p_{10}, p_{11}$, in each of the $G \times E$ categories, we could use the same formulas as above above but we replace γ_0 by $\log\{\frac{p_{00}}{1-p_{00}}\}$, γ_1 by $\log\{\frac{p_{10}}{1-p_{10}}/\frac{p_{00}}{1-p_{00}}\}$, γ_2 by $\log\{\frac{p_{01}}{1-p_{01}}/\frac{p_{00}}{1-p_{00}}\}$, and

1. Demidenko (2008) also noted that a number of previous authors (Hwang et al., 1994; Foppa and Spiegelman, 1997) who had considered sample size and power calculations for interaction in logistic regression had relied on a different formula for their sample size calculations. These other authors had assumed that for the test statistic, the variance of $\widehat{\gamma}_3$ had been calculated under the null hypothesis of no interaction. When the variance for the test statistic is calculated under the null of no interaction, then the required sample size is given by $n = \frac{(Z_{1-\alpha/2}\sqrt{V_0} + Z_\beta\sqrt{V_{mult(OR)}})}{\eta^2}$ rather than by $n = \frac{(Z_{1-\alpha/2}+Z_\beta)^2 V_{mult(OR)}}{\eta^2}$, where V_0 is the variance of $\widehat{\gamma}_3$ calculated under the null that $\gamma_3 = 0$. Demidenko (2008) points out that although the sample size calculations of Hwang et al. (1994) and Foppa and Spiegelman (1997) would be fine if, for $\widehat{\gamma}_3$, the variance were indeed calculated under the null, in practice, the variance of $\widehat{\gamma}_3$ is almost always calculated under the alternative; it is the variance under the alternative that is generally given as the default in standard logistic regression output. Thus, the sample size calculations of Hwang et al. (1994) and Foppa and Spiegelman (1997), although not technically incorrect, do not correspond to the test statistics that are generally used in practice. A similar point and criticism was made by Garcia-Closas and Lubin (1999) some years earlier. Both Garcia-Closas and Lubin (1999) and Demidenko (2008) note that when interactions are large, the sample-size calculations using the "null variance" can underestimate the required sample size if the test statistic with the variance under the alternative is in fact used. Likewise, a similar point pertains to the sample-size and power calculations of Yang et al. (1997) for multiplicative interaction in case-only studies considered in Section 13.2.3 (cf. VanderWeele, 2011g).

γ_3 by $\log\left[\frac{p_{11}}{1-p_{11}}\frac{p_{00}}{1-p_{00}}/\{\frac{p_{10}}{1-p_{10}}\frac{p_{01}}{1-p_{01}}\}\right]$. Similar remarks pertain to other power and sample-size calculations below.

We give a brief example on the use of these formulae.

Example. Suppose we wish to calculate the power of a test at significance level $\alpha = 0.05$, with $n = 5000$, with the joint prevalence of the genetic and environmental factors being $\pi_{00} = 0.35, \pi_{10} = 0.20, \pi_{01} = 0.20$, and $\pi_{11} = 0.25$ respectively, with the probability of the outcome in the reference category of $P(Y = 1|G = 0, E = 0) = 0.015$, with main effects on the odds ratio scale of $e^{\gamma_1} = 1.3$ and $e^{\gamma_2} = 1.4$ and with odds ratio multiplicative interaction $e^{\gamma_3} = 1.6$. We can calculate L, F, J, R from these values and the variance $V_{mult(OR)}$ to obtain $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V)}\right\} = 0.216$.

13.2.2. Multiplicative Interaction with Case–Control Data

Suppose instead we fit a logistic regression model to case–control data:

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

The sample size required to detect a $\gamma_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V^*_{mult(OR)}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where $V^*_{mult(OR)}$ is the variance of $\widehat{\gamma}_3$ under the alternative that $\gamma_3 = \eta$. This is given by (Demidenko, 2008)

$$V^*_{mult(OR)} = \frac{1}{L^*} + \frac{1}{F^*} + \frac{1}{J^*} + \frac{1}{R^*}$$

with

$$L^* = \frac{e^{\gamma_0}}{(1+e^{\gamma_0})^2}\pi^*_{00}$$

$$F^* = \frac{e^{\gamma_0+\gamma_1}}{(1+e^{\gamma_0+\gamma_1})^2}\pi^*_{10}$$

$$J^* = \frac{e^{\gamma_0+\gamma_2}}{(1+e^{\gamma_0+\gamma_2})^2}\pi^*_{01}$$

$$R^* = \frac{e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}{(1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3})^2}\pi^*_{11}$$

and where $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ are now the proportions of subjects in each joint exposure stratum *in the case–control sample*.

If we know the overall outcome prevalence in the underlying population, $P(Y = 1)$, we could also obtain the proportions $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ from the proportions of subjects in each joint exposure stratum in the underlying population, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$, though doing so requires solving a nonlinear equation numerically (Demidenko, 2008). Alternatively, if the outcome is rare, we can obtain $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ from $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ approximately using the following formulas (VanderWeele, 2012c):

$$\pi^*_{00} \approx \pi_{00} P^*(Y = 0) + \frac{\pi_{00}}{\pi_{00} + \pi_{10} e^{\gamma_1} + \pi_{01} e^{\gamma_2} + \pi_{11} e^{\gamma_1 + \gamma_2 + \gamma_3}} P^*(Y = 1)$$

$$\pi^*_{10} \approx \pi_{10} P^*(Y = 0) + \frac{e^{\gamma_1} \pi_{10}}{\pi_{00} + \pi_{10} e^{\gamma_1} + \pi_{01} e^{\gamma_2} + \pi_{11} e^{\gamma_1 + \gamma_2 + \gamma_3}} P^*(Y = 1)$$

$$\pi^*_{01} \approx \pi_{01} P^*(Y = 0) + \frac{e^{\gamma_2} \pi_{01}}{\pi_{00} + \pi_{10} e^{\gamma_1} + \pi_{01} e^{\gamma_2} + \pi_{11} e^{\gamma_1 + \gamma_2 + \gamma_3}} P^*(Y = 1)$$

$$\pi^*_{11} \approx \pi_{11} P^*(Y = 0) + \frac{e^{\gamma_1 + \gamma_2 + \gamma_3} \pi_{11}}{\pi_{00} + \pi_{10} e^{\gamma_1} + \pi_{01} e^{\gamma_2} + \pi_{11} e^{\gamma_1 + \gamma_2 + \gamma_3}} P^*(Y = 1)$$

where $P^*(Y = 0)$ is the proportion of controls in the case–control sample and $P^*(Y = 1)$ is the proportion of cases in the case–control sample. If we instead specify the marginal probabilities of each exposure $\pi_g = P(G = 1)$ and $\pi_e = P(E = 1)$ and the odds ratio, Δ, relating G and E, in the underlying population, then we can calculate $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ using the formulae given in Section 13.2.1.

Thus, to calculate the sample size to reject the null of no multiplicative interaction with a certain power β from case–control data, we would need to specify (i) the significance level α and the power β; (ii) the proportion of subjects in each exposure stratum, $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ in the case–control sample, or alternatively these proportions $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ or the marginal probabilities and marginal odds ratio, π_g, π_e, Δ, in the underlying population along with a rare outcome assumption and the proportions of cases $P^*(Y = 1)$ in the case–control sample, and finally (iii) the main effect odds ratios of the two exposures on the logistic scale, γ_1 and γ_2, the log odds of the baseline probability of the outcome in the doubly unexposed group $\gamma_0 = \log\{P^*(Y = 1|G = 0, E = 0)/P^*(Y = 1|G = 0, E = 0)\}$ in the case–control sample, and the magnitude of the interaction on the multiplicative scale γ_3. Note that if the joint or marginal exposure probabilities are specified separately for the cases and controls, then under an assumption of a rare outcome, the distribution of the exposures amongst the controls could be used as an approximation to $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ or π_g, π_e, Δ.

Note also that with case–control data, $\gamma_0 = \log\{P^*(Y = 1|G = 0, E = 0)/P^*(Y = 0|G = 0, E = 0)\}$ is the log odds of baseline probability of the outcome in doubly unexposed group *in the case–control sample*—that is, the log of the number of cases to controls in the study for the doubly unexposed group. Under a rare outcome assumption, γ_0 can be approximated as $\gamma_0 \approx \text{logit}[1/\{1 + (\pi_{00} + \pi_{10} e^{\gamma_1} + \pi_{01} e^{\gamma_2} + \pi_{11} e^{\gamma_1 + \gamma_2 + \gamma_3}) P^*(Y = 0)/P^*(Y = 1)\}]$ (VanderWeele, 2012c).

Example. Suppose we wish to calculate the sample size required for a test for multiplicative interaction at significance level $\alpha = 0.05$, with power $\beta = 0.80$, with the joint prevalence of the genetic and environmental factors being $\pi_g = 0.5$ and $\pi_e = 0.3$, respectively, in the underlying population with the factors being independent in the underlying population so that $\Delta = 1$. Suppose that the number of cases and controls in the study were going to be equal $P^*(Y = 1) = P^*(Y = 0) = 0.5$, with main effects on the odds ratio scale of $e^{\gamma_1} = 1.1$ and $e^{\gamma_2} = 1.1$ and with multiplicative interaction $e^{\gamma_3} = 1.5$. We can calculate that the sample size thus required to detect positive multiplicative interaction would be $n = 3447$.

If instead of calculating the required sample size for a fixed power β, we wanted to calculate the power for a given sample size using the Wald test for the null hypothesis $\gamma_3 = 0$ based on the logistic regression model, we could proceed as follows. For a fixed sample size n, the power to reject the null $\gamma_3 = 0$ at significance level α under the alternative that $\gamma_3 = \eta$ is given by

$$\text{Power} = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V^*_{mult(OR)}}\right\}$$

where Φ^{-1} is the inverse cumulative distribution function for a standard normal random variable and where $V^*_{mult(OR)}$ can be calculated as above. If the null hypothesis were rejected for extreme values of γ_3 on either side of zero (two-sided test), then the relevant power formula would be

$$\begin{aligned}\text{Power} = {} & \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V^*_{mult(OR)})}\right\} \\ & + \Phi^{-1}\left\{-Z_{1-\alpha/2} - \eta\sqrt{(n/V^*_{mult(OR)})}\right\}.\end{aligned}$$

13.2.3. Multiplicative Interaction with Case-Only Data

As noted in Chapter 12, when multiplicative interaction is of interest and the distributions of the genetic and environmental factors are independent of one another in the underlying population, a "case-only" estimator of multiplicative interaction will have greater power to detect multiplicative interaction as it exploits the independence assumption (Piegorsch et al., 1994; Yang et al., 1999).

As noted in the last chapter, under the assumption of independence of the distributions of the two exposures, along with a rare outcome assumption, the coefficient γ_3 in the logistic regression in the previous subsection can also be estimated by regressing one exposure on the other but only amongst the cases using the regression

$$\text{logit}\{P(G = 1|E = e, Y = 1)\} = \theta_0 + \theta_1 e$$

Under the assumptions of independence of the two exposures, along with a rare outcome assumption, we have $\theta_1 \approx \gamma_3$. Moreover, using θ_1 to test for multiplicative interaction will often constitute a more powerful test than using γ_3 in the logistic

regression, because it fully exploits the independence assumption, even though the test only uses data among the cases ($Y = 1$).

Power and sample-size calculations can likewise be obtained for the case-only test for interaction. The sample size required to detect an interaction of magnitude $\theta_1 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{CO}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1 - \alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where V_{CO} is the variance of $\widehat{\theta}_1$ under the alternative that $\theta_1 = \eta$. This is given by (Yang et al., 1997; VanderWeele, 2011g)

$$V_{CO} = (m_1 + m_2 + m_3 + m_4)\left(\frac{1}{m_1} + \frac{1}{m_2} + \frac{1}{m_3} + \frac{1}{m_4}\right)$$

with

$$\begin{aligned} m_1 &= (1 - \pi_g)(1 - \pi_e) \\ m_2 &= \pi_g(1 - \pi_e)\exp(\gamma_1) \\ m_3 &= (1 - \pi_g)\pi_e \exp(\gamma_2) \\ m_4 &= \pi_g \pi_e \exp(\gamma_1 + \gamma_2 + \gamma_3) \end{aligned}$$

To calculate the sample size, we would need to specify (i) the significance level α, the power β, and the magnitude of multiplicative interaction $\gamma_3 = \eta$; (ii) the proportion of subjects with each of the two exposures, π_e and π_g; and (iii) the main effect of the two exposures, γ_1 and γ_2, in the logistic regression model.

If instead of calculating the required sample size for a fixed power β, we wanted to calculate the power for a given sample size using the Wald test for the null hypothesis $\theta_1 = 0$ using the case-only estimator, we could proceed as follows. For a fixed sample size n, the power to reject the null hypothesis $\theta_1 = 0$ at the significance level α under the alternative that $\theta_1 = \eta$ is given by

$$\text{Power} = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{CO})}\right\}$$

where Φ^{-1} is the inverse cumulative distribution function for a standard normal random variable and where V_{CO} can be calculated as above.[2] For a two-sided test to detect either positive or negative multiplicative interaction, we could use Power $= \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{CO})}\right\} + \Phi^{-1}\left\{-Z_{1-\alpha/2} - \eta\sqrt{(n/V_{CO})}\right\}$. Again, in

2. Note Yang et al. (1997) give a slightly different formula for the variance V_{CO} for multiplicative interaction in case-only studies. This is because Yang et al. (1997) assumed that for the test statistic, the variance was calculated under the null hypothesis of no interaction, whereas in practice, the variance is almost always calculated under the alternative (VanderWeele, 2011g). When interactions are large, the sample size calculations using the "null-variance" can underestimate the required sample size if the test statistic with the variance under the alternative is in fact used.

Section 13.5 we will describe how to use a simple Excel spreadsheet to carry out such sample–size and power calculations automatically.

Although these case-only estimators can be quite powerful, as discussed in Chapter 12, they are also fairly sensitive to the assumption that the distributions of the two exposures are independent in the population and can result in considerable bias if this assumption does not hold (Albert et al., 2001).

13.3. POWER AND SAMPLE SIZE CALCULATIONS FOR BINARY OUTCOMES: ADDITIVE INTERACTION

As noted in Chapters 9 and 10, interaction on the additive scale is more relevant for public health purposes and is also more closely related to notions of mechanistic synergism within the sufficient cause framework. In this section we will consider measures of additive interaction based on absolute risks and also on the relative excess risk due to interaction for both cohort and case–control data and we will provide closed-form analytic expressions for power and sample size in each of these cases. We will see that when main effects of both exposures are positive, power to detect positive interaction on the additive scale will be greater than that on the multiplicative scale.

Recall that if $p_{ge} = P(Y = 1|G = g, E = e)$ and we let $\pi_{ge} = P(G = g, E = e)$, then the measure of interaction on the additive scale using risks is given by

$$p_{11} - p_{10} - p_{01} + p_{00}$$

This can be re-expressed as $(p_{11} - p_{00}) - \{(p_{10} - p_{00}) + (p_{01} - p_{00})\}$ and measures the extent to which the effect of both exposures combined exceeds (or is less than) the sum of the effects of each exposure considered separately. If $p_{11} - p_{10} - p_{01} + p_{00} > 0$, the interaction is said to be positive or "superadditive." If $p_{11} - p_{10} - p_{01} + p_{00} < 0$, the interaction is said to be negative or "subadditive." If $p_{11} - p_{10} - p_{01} + p_{00} = 0$, there is said to be no interaction on the additive scale. This measure of additive interaction corresponds to the coefficient of the product term for the two exposures in a linear risk model for the outcome.

As in Chapter 9, suppose now we were to divide our measure of additive interaction based on risks, $p_{11} - p_{10} - p_{01} + p_{00}$, by the baseline risk p_{00}. We would then obtain what is sometimes referred to as the relative excess risk due to interaction or *RERI* (Rothman, 1986):

$$RERI = RR_{11} - RR_{10} - RR_{01} + 1$$

where $RR_{ge} = \frac{P(Y=1|G=g,E=e)}{P(Y=1|G=0,E=0)} = \frac{p_{ge}}{p_{00}}$ are the relative risks. This measure *RERI* will be greater than 0 (or respectively less than 0) if and only if the measure of additive interaction using absolute risks, $p_{11} - p_{10} - p_{01} + p_{00}$, is greater than 0 (or less than 0, respectively). The relative excess risk due to interaction can thus be used to assess additive interaction using data on relative risks. When the probability of the outcome is rare in all exposure strata, then odds ratios will approximate risk ratios,

namely $\frac{p_{ge}/(1-p_{ge})}{p_{00}/(1-p_{00})} \approx \frac{p_{ge}}{p_{00}}$, and thus we can approximate *RERI* by

$$RERI_{OR} = OR_{11} - OR_{10} - OR_{01} + 1 \approx RERI$$

This final measure, $RERI_{OR} = OR_{11} - OR_{10} - OR_{01} + 1$, is advantageous because it is an approximate measure of additive interaction and yet can also be obtained directly from logistic regression analyses and from case–control data. We will, however, first begin with additive interaction on the absolute risk scale using cohort data.

13.3.1. Additive Interaction in Cohort Studies Using a Linear Risk Model

Suppose data were available from a cohort study and we were to use a linear risk model to measure additive interaction:

$$P(Y = 1|G = g, E = e) = \theta_0 + \theta_1 g + \theta_2 e + \theta_3 ge$$

In this model, $\theta_3 = p_{11} - p_{10} - p_{01} + p_{00}$ is our measure of additive interaction. Suppose we plan to fit this model to the cohort data and use a Wald test for the null hypothesis $\theta_3 = 0$. Once we have fit the model and obtained an estimate $\widehat{\theta}_3$ of θ_3 from the data, the Wald test statistic for the null hypothesis $\theta_3 = 0$ is given by $\widehat{\theta}_3/\widehat{V}$, where $\widehat{V}$ is the estimated variance of $\widehat{\theta}_3$. We would reject the null at significance level α if $|\widehat{\theta}_3/\widehat{V}| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution. Suppose we wish to calculate the sample size required to reject the null hypothesis with significance level α and power β if the magnitude of the interaction were $\theta_3 = \eta$.

The sample size required to detect an additive interaction of magnitude $\theta_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where V is the variance of $\widehat{\theta}_3$ under the alternative that $\theta_3 = \eta$. The additional computational burden lies in calculating the variance V. This variance V is given by VanderWeele (2012c):

$$V = \frac{1}{L'} + \frac{1}{F'} + \frac{1}{J'} + \frac{1}{R'}$$

where

$$L' = \frac{1}{(\theta_0)(1-\theta_0)}\pi_{00}$$
$$F' = \frac{1}{(\theta_0+\theta_1)\{1-(\theta_0+\theta_1)\}}\pi_{10}$$
$$J' = \frac{1}{(\theta_0+\theta_2)\{1-(\theta_0+\theta_2)\}}\pi_{01}$$
$$R' = \frac{1}{(\theta_0+\theta_1+\theta_2+\theta_3)\{1-(\theta_0+\theta_1+\theta_2+\theta_3)\}}\pi_{11}$$

Thus to calculate the sample size, we would need to specify (i) the significance level α, the power β, and the magnitude of additive interaction $\theta_3 = \eta$; (ii) the proportion of subjects in each exposure stratum, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$; and (iii) the main effect of the two exposures on the additive scale θ_1 and θ_2 and the baseline risk of the doubly unexposed group $\theta_0 = P(Y = 1|G = 0, E = 0)$.

Instead of specifying the proportion of subjects in each joint exposure stratum $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$, we could alternatively specify the marginal probability of each exposure $\pi_g = P(G = 1)$ and $\pi_e = P(E = 1)$ along with the odds ratio relating G and E, $\Delta = \{P(G = 1|E = 1)/P(G = 0|E = 1\}/\{P(G = 1|E = 0)/P(G = 0|E = 0\}$. The probabilities $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ are then given by (Demidenko, 2008)

$$\begin{aligned}
\pi_{00} &= \frac{1-\pi_e}{1+C} \\
\pi_{10} &= \frac{(1-\pi_e)C}{1+C} \\
\pi_{01} &= \frac{\pi_e}{1+C\Delta} \\
\pi_{11} &= \frac{C\Delta\pi_e}{1+C\Delta}
\end{aligned} \tag{2}$$

where

$$C = \frac{q + \sqrt{q^2 + 4\pi_g(1-\pi_g)\Delta}}{2(1-\pi_g)\Delta}$$
$$\text{and} \quad q = \pi_g(1+\Delta) + \pi_e(1-\Delta) - 1$$

If G and E are independent, then $\Delta = 1$ and C simplifies to $C = \pi_e/(1-\pi_e)$.

If instead of calculating the required sample size for a fixed power β, we wanted to calculate the power for a given sample size using the Wald test for the null hypothesis $\theta_3 = 0$ based on the linear risk model, we could proceed as follows. For a fixed sample size n, the power to reject the null $\theta_3 = 0$ at significance level α under the alternative that $\theta_3 = \eta$ is given by

$$\text{Power} = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V)}\right\}$$

where Φ^{-1} is the inverse cumulative distribution function for a standard normal random variable and where V can be calculated as above. Below in Section 13.5 we describe how to use a simple Excel spreadsheet to carry out such sample size and power calculations automatically. If the null hypothesis were rejected for extreme values of θ_3 on either side of zero (two-sided test), then the relevant power formula would be

$$\text{Power} = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V)}\right\} + \Phi^{-1}\left\{-Z_{1-\alpha/2} - \eta\sqrt{(n/V)}\right\}$$

Before moving on, we give a brief example of the use of these formulae for additive interaction.

Example. Suppose we wish to calculate the power of a test at significance level $\alpha = 0.05$, with $n = 4000$, with the prevalence of the genetic and environmental factors being $\pi_g = 0.5$ and $\pi_e = 0.3$ respectively and assuming these are independent so that $\Delta = 1$, with the probability of the outcome in the reference category of $\theta_0 = P(Y = 1|G = 0, E = 0) = 0.02$, with main effects on the risk difference scale of $\theta_1 = 0.01$ and $\theta_2 = 0.01$ and with additive interaction $\theta_3 = 0.02$. We can use equations (13.1) to calculate $\pi_{00} = 0.35, \pi_{10} = 0.35, \pi_{01} = 0.15, \pi_{11} = 0.15$, and from this we can calculate L', F', J', R' and the variance V and the power $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V)}\right\}$ to obtain 0.32.

13.3.2. Additive Interaction in Cohort Studies Using Logistic Regression and RERI

In this subsection we consider power and sample-size calculations for measures of interaction based on the relative excess risk for interaction using odds ratios, $RERI_{OR} = OR_{11} - OR_{10} - OR_{01} + 1$, obtained from logistic regression using cohort data. If the outcome is rare, then this will approximate the relative excess risk due to interaction for risk ratios $RERI = RR_{11} - RR_{10} - RR_{01} + 1$.

Suppose we fit a logistic regression model to cohort data:

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

The RERI from this logistic regression model is given by

$$RERI_{OR} = e^{\gamma_1 + \gamma_2 + \gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$$

Suppose we wish to use a Wald test for the null hypothesis $RERI_{OR} = 0$. The sample size required to detect a $RERI_{OR}$ of magnitude $\eta = e^{\gamma_1 + \gamma_2 + \gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{RERI(OR)}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1 - \alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where $V_{RERI(OR)}$ is the variance of $RERI_{OR} =$

$e^{\widehat{\gamma}_1+\widehat{\gamma}_2+\widehat{\gamma}_3} - e^{\widehat{\gamma}_1} - e^{\widehat{\gamma}_2} + 1$ under the alternative. This variance, $V_{RERI(OR)}$, is given by (VanderWeele, 2012c)

$$V_{RERI(OR)} = \left(\frac{1}{L}+\frac{1}{R}\right)e^{2(\gamma_1+\gamma_2+\gamma_3)} - \frac{2}{L}e^{2\gamma_1+\gamma_2+\gamma_3} - \frac{2}{L}e^{\gamma_1+2\gamma_2+\gamma_3} + \left(\frac{1}{L}+\frac{1}{F}\right)e^{2\gamma_1} + \left(\frac{1}{L}+\frac{1}{J}\right)e^{2\gamma_2} + \frac{2}{L}e^{\gamma_1+\gamma_2}$$

where L, F, J, R are given as in Section 13.2.1, that is,

$$L = \frac{e^{\gamma_0}}{(1+e^{\gamma_0})^2}\pi_{00}$$
$$F = \frac{e^{\gamma_0+\gamma_1}}{(1+e^{\gamma_0+\gamma_1})^2}\pi_{10}$$
$$J = \frac{e^{\gamma_0+\gamma_2}}{(1+e^{\gamma_0+\gamma_2})^2}\pi_{01}$$
$$R = \frac{e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}{(1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3})^2}\pi_{11}$$

To calculate the sample size to reject the null of no additive interaction using $RERI_{OR}$, we would need to specify (i) the significance level α, the power β; (ii) the proportion of subjects in each exposure stratum, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$; and (iii) the main effect odds ratios of the two exposures on the logistic scale, γ_1 and γ_2, the log odds of the baseline risk of the doubly unexposed group $\gamma_0 = \log\{P(Y=1|G=0,E=0)/P(Y=0|G=0,E=0)\}$, and the magnitude of the interaction on the multiplicative scale γ_3. Instead of specifying the magnitude of the interaction on the multiplicative scale, γ_3, one could specify the magnitude of $RERI_{OR}$ under the alternative $RERI_{OR} = \eta$ and then back-calculate the magnitude of $\gamma_3 = \log(\eta + e^{\gamma_1} - e^{\gamma_2} - 1) - \gamma_1 - \gamma_2$.

And once again, if instead of specifying the joint probabilities $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$, we specified the marginal probabilities of each exposure $\pi_g = P(G=1)$ and $\pi_e = P(E=1)$ and the odds ratio relating G and E, $\Delta = \{P(G=1|E=1)/P(G=0|E=1)\}/\{P(G=1|E=0)/P(G=0|E=0)\}$, then we could obtain $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ using the formulae in Section 13.2.1.

If instead of calculating the required sample size for a given power, we wanted to calculate the power for a given sample-size we could use Power $= \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{RERI(OR)})}\right\}$ or, for a two-sided test, to detect either positive or negative additive interaction we could use Power $= \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{RERI(OR)})}\right\} + \Phi^{-1}\left\{-Z_{1-\alpha/2} - \eta\sqrt{(n/V_{RERI(OR)})}\right\}$. Again, in Section 13.5 we will describe how to use a simple Excel spreadsheet to carry out such sample-size and power calculations automatically.

Example. Suppose again we wish to calculate the power of a test at significance level $\alpha = 0.05$, with $n = 5000$, with the joint prevalence of the genetic and environmental factors being $\pi_{00} = 0.35, \pi_{10} = 0.20, \pi_{01} = 0.20$, and $\pi_{11} = 0.25$ respectively, with the probability of the outcome in the reference category of $P(Y = 1|G = 0, E = 0) = 0.015$, with main effects on the odds ratio scale of $e^{\gamma_1} = 1.3$ and $e^{\gamma_2} = 1.4$ and with odds ratio multiplicative interaction $e^{\gamma_3} = 1.6$ as in the example for multiplicative interaction in Section 13.2.1, but that we now wish to calculate the power for testing $RERI_{OR} > 0$. Here the true $RERI_{OR}$ is $\eta = e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1 = (1.3)(1.4)(1.6) - (1.3) - (1.4) + 1 = 1.212 > 0$. From L, F, J, R we can calculate the variance $V_{RERI(OR)}$ to obtain $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{RERI(OR)})}\right\} = 0.482$. In this example, the power to detect additive interaction, 0.482, is greater than that to detect multiplicative interaction, 0.216.

The reader is reminded that the tests for additive interaction using $RERI_{OR}$ hold only approximately to the extent that the outcome is rare so that $RERI_{OR}$ approximates *RERI* on the risk ratio scale. In the Appendix we also give sample size and power formulae for the multiplicative interaction from a log-linear model and for additive interaction using *RERI* estimated from a log-linear model. However, if the measure of additive interaction is fit with cohort data, it may be preferable to fit the linear risk model directly for additive interaction using absolute risks rather than employing *RERI*.

13.3.3. Additive Interaction in Case–Control Studies Using Logistic Regression and RERI

Suppose instead we fit a logistic regression model to case–control data:

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

The sample size required to detect a $RERI_{OR}$ of magnitude $\eta = e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V^*_{RERI(OR)}}{\eta^2}$$

where (VanderWeele, 2012c)

$$V^*_{RERI(OR)} = \left(\frac{1}{L^*} + \frac{1}{R^*}\right) e^{2(\gamma_1+\gamma_2+\gamma_3)} - \frac{2}{L^*} e^{2\gamma_1+\gamma_2+\gamma_3} - \frac{2}{L^*} e^{\gamma_1+2\gamma_2+\gamma_3} + \left(\frac{1}{L^*} + \frac{1}{F^*}\right) e^{2\gamma_1} + \left(\frac{1}{L^*} + \frac{1}{J^*}\right) e^{2\gamma_2} + \frac{2}{L^*} e^{\gamma_1+\gamma_2}$$

with

$$L^* = \frac{e^{\gamma_0}}{(1+e^{\gamma_0})^2}\pi^*_{00}$$

$$F^* = \frac{e^{\gamma_0+\gamma_1}}{(1+e^{\gamma_0+\gamma_1})^2}\pi^*_{10}$$

$$J^* = \frac{e^{\gamma_0+\gamma_2}}{(1+e^{\gamma_0+\gamma_2})^2}\pi^*_{01}$$

$$R^* = \frac{e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}{(1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3})^2}\pi^*_{11}$$

and where $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ are now the proportions of subjects in each joint exposure stratum *in the case–control sample*.

If we know the overall outcome prevalence in the underlying population, $P(Y = 1)$, we could also obtain the proportions $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ from the proportions of subjects in each joint exposure stratum in the underlying population, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$, although doing so requires solving a nonlinear equation numerically (Demidenko, 2008). Alternatively, if the outcome is rare, we can obtain $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ from $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ approximately using the following formulas (VanderWeele, 2012c):

$$\pi^*_{00} \approx \pi_{00}P^*(Y=0) + \frac{\pi_{00}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$

$$\pi^*_{10} \approx \pi_{10}P^*(Y=0) + \frac{e^{\gamma_1}\pi_{10}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$

$$\pi^*_{01} \approx \pi_{01}P^*(Y=0) + \frac{e^{\gamma_2}\pi_{01}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$

$$\pi^*_{11} \approx \pi_{11}P^*(Y=0) + \frac{e^{\gamma_1+\gamma_2+\gamma_3}\pi_{11}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$

where $P^*(Y = 0)$ is the proportion of controls in the case–control sample and $P^*(Y = 1)$ is the proportion of cases in the case–control sample. If we instead specify the marginal probabilities of each exposure $\pi_g = P(G = 1)$ and $\pi_e = P(E = 1)$ and the odds ratio, Δ, relating G and E, in the underlying population, then we can calculate $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ using the formulae in Section 13.3.1.

Thus, to calculate the sample size to reject the null hypothesis of no additive interaction using $RERI_{OR}$ from case–control data, we would need to specify (i) the significance level α, the power β; (ii) the proportion of subjects in each exposure stratum, $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$ in the case–control sample, or alternatively these proportions $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ or the marginal probabilities and marginal odds ratio, π_g, π_e, Δ, in the underlying population along with a rare outcome assumption and the proportions of cases $P^*(Y = 1)$ in the case–control sample, and finally (iii) the main effect odds ratios of the two exposures on the logistic scale, γ_1 and γ_2, the log odds of the baseline probability of the outcome in the doubly unexposed

group $\gamma_0 = \log\{P^*(Y=1|G=0,E=0)/P^*(Y=1|G=0,E=0)\}$ in the case–control sample, and the magnitude of the interaction on the multiplicative scale γ_3 (or instead we could specify the magnitude of $RERI_{OR} = \eta$ and then back-calculate the magnitude of $\gamma_3 = \log(\eta + e^{\gamma_1} - e^{\gamma_2} - 1) - \gamma_1 - \gamma_2$ and use this). Note that if the joint or marginal exposure probabilities are specified separately for the cases and controls, then under an assumption of a rare outcome, the distribution of the exposures amongst the controls could be used as an approximation to $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ or π_g, π_e, Δ.

Note also that with case–control data, $\gamma_0 = \log\{P^*(Y=1|G=0,E=0)/P^*(Y=0|G=0,E=0)\}$ is the log odds of baseline probability of the outcome in doubly unexposed group *in the case–control sample*—that is, the log of the number of cases to controls in the study for the doubly unexposed group. Under a rare outcome assumption, γ_0 can be approximated as $\gamma_0 \approx \text{logit}[1/\{1 + (\pi_{00} + \pi_{10}e^{\gamma_1} + \pi_{01}e^{\gamma_2} + \pi_{11}e^{\gamma_1+\gamma_2+\gamma_3})P^*(Y=0)/P^*(Y=1)\}]$.

Example. Suppose we wish to calculate the size required for a test at significance level $\alpha = 0.05$, with power $\beta = 0.80$, with the joint prevalence of the genetic and environmental factors being $\pi_g = 0.5$, and $\pi_e = 0.3$, respectively, in the underlying population with the factors being independent in the underlying population so that $\Delta = 1$. Suppose that the number of cases and controls in the study were going to be equal, $P^*(Y=1) = P^*(Y=0) = 0.5$, with main effects on the odds ratio scale of $e^{\gamma_1} = 1.1$ and $e^{\gamma_2} = 1.1$ and with multiplicative interaction $e^{\gamma_3} = 1.5$. We can also calculate that sample size required to detect positive interaction using $RERI_{OR}$ would be $n = 2212$. In Section 13.2.2, we showed that the sample size required to detect positive multiplicative interaction, under this same setting, would be $n = 3447$.

13.3.4. Power Comparison of Additive and Multiplicative Interaction

VanderWeele (2009d) noted that in a log-linear model with non-negative main effects, whenever positive multiplicative interaction is present on the risk ratio scale, positive additive interaction on the risk difference scale will be present as well; the reverse implication does not hold. Here we present a brief comparison of power to detect such additive or multiplicative interaction and we will consider the odds ratio scale rather than the risk ratio scale. In this power comparison we will assume a case–control study with a rare outcome so that $RERI_{OR}$ approximates a measure of additive interaction. Table 13.1 reports power for a number of scenarios with varying sample sizes, main effect odds ratios, and multiplicative interaction parameters on the odds ratio scale $I_{OR} = \frac{OR_{11}}{OR_{10}OR_{01}}$.

In these examples it is assumed that the proportion of cases and controls in the case–control sample are equal and that the prevalence of the genetic and environmental factors are each $\pi_g = \pi_e = 0.5$, with the odds ratio relating these factors being $\Delta = 1.1$. Note that in all scenarios considered, there is positive interaction on both additive and multiplicative scales. Power for one-sided test (rejecting only for positive interaction) is reported.

Table 13-1. Power to Detect Additive Interaction and Multiplicative Interaction for Various Sample Sizes, Main Effects, and Interaction Parameters

I_{OR}	OR_{10}	OR_{01}	$n = 500$	$n = 1000$	$n = 3000$	$n = 5000$
1.1	1	1	.05,.05	.06,.06	.10,.09	.14,.13
1.1	1.3	1.3	.07,.04	.10,.05	.23,.09	.34,.12
1.1	1.5	1.8	.13,.04	.23,.05	.55,.08	.77,.11
1.3	1	1	.12,.11	.21,.17	.50,.42	.72,.62
1.3	1.3	1.3	.18,.10	.32,.15	.73,.37	.91,.56
1.3	1.5	1.8	.27,.09	.48,.14	.91,.33	.99,.50
1.5	1	1	.25,.19	.44,.34	.88,.77	.98,.93
1.5	1.3	1.3	.32,.17	.56,.30	.95,.70	1.00,.89
1.5	1.5	1.8	.40,.15	.68,.26	.99,.63	1.00,.84
2	1	1	.57,.44	.85,.73	1.00,.99	1.00,1.00
2	1.3	1.3	.58,.39	.86,.65	1.00,.98	1.00,1.00
2	1.5	1.8	.59,.34	.87,.59	1.00,.97	1.00,1.00
3	1	1	.81,.77	.98,.97	1.00,1.00	1.00,1.00
3	1.2	1.3	.74,.70	.96,.94	1.00,1.00	1.00,1.00
3	1.5	1.8	.68,.62	.93,.89	1.00,1.00	1.00,1.00

Note: The first number in each column is the power to detect additive interaction; the second number is the power for multiplicative interaction.)

We see that for the scenarios considered here with non-negative main effects and positive interaction, power is greater to detect additive interaction than multiplicative interaction. However, as noted in Greenland (1983), when outcome probabilities are additive or subadditive, power to detect a (negative) multiplicative interaction will often be greater.

13.4. POWER AND SAMPLE SIZE CALCULATIONS FOR BINARY OUTCOMES: MECHANISTIC INTERACTION

In Chapters 9 and 10 we discussed "mechanistic" or "sufficient cause" interactions that provide a somewhat different and often stronger notion of positive additive interaction. As discussed there, a sufficient cause interaction is present if there are individuals for whom the outcome would occur if both exposures are present but would not occur if just one or the other exposure is present. In counterfactual notation, if we let Y_{ge} denote the counterfactual outcome (or potential outcome) for each subject if, possibly contrary to fact, G had been set to g and E had been set to e, then a sufficient cause interaction is present if for some individual $Y_{11} = 1$ but $Y_{10} = Y_{01} = 0$. If the effects of the two exposures on the outcome are unconfounded, then

$$p_{11} - p_{10} - p_{01} > 0$$

implies the presence of a sufficient cause interaction (VanderWeele and Robins, 2007b, 2008). This is a stronger condition than regular positive additive interaction

which only requires $p_{11} - p_{10} - p_{01} + p_{00} > 0$. The condition $p_{11} - p_{10} - p_{01} > 0$ is more stringent because we are no longer adding back in the outcome probability p_{00} for the doubly unexposed group. The condition $p_{11} - p_{10} - p_{01} > 0$ expressed in terms of *RERI* is equivalent to $RERI > 1$.

In Chapters 9 and 10, we also discussed empirical tests for an even stronger notion of interaction. We said that there is a "singular" or "epistatic" interaction if there are individuals in the population who will have the outcome if and only if both exposures are present; in counterfactual notation, this is that there are individuals for whom $Y_{11} = 1$ but $Y_{10} = Y_{01} = Y_{00} = 0$. In the genetics literature, when gene–gene interactions are considered, such response patterns are sometimes called instances of "compositional epistasis" (Phillips, 2008; Cordell, 2009) and constitute settings in which the effect of one genetic factor is masked unless the other is present. It was noted that if the effects of the two exposures on the outcome were unconfounded, then

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

would imply the presence of such an "epistatic interaction" (VanderWeele, 2010b,c). Again this is an even stronger notion of interaction in that we are now subtracting p_{00}. The condition $p_{11} - p_{10} - p_{01} - p_{00} > 0$ expressed in terms of *RERI* is equivalent to $RERI > 2$.

It is relatively straightforward to derive sample-size and power formulae for tests for such sufficient cause or epistatic interactions. The sample size for *RERI* given in Section 13.3.2 could be used; but for sufficient cause interaction, to test $RERI > 1$, one would replace the η in the denominator of the sample-size formula by $(\eta - 1)$; and for epistatic interaction, to test $RERI > 2$, one would replace the η in the denominator of the formula by $(\eta - 2)$.

Thus, for cohort data, to detect a sufficient cause interaction $(RERI > 1)$ at significance level α with power β when the true *RERI* is $\eta = e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$, the required sample size would be

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{RERI}}{(\eta - 1)^2}$$

where V_{RERI} is the variance of *RERI*. And likewise, the power to detect a sufficient cause interaction for a given sample size is $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + (\eta - 1)\sqrt{(n/V_{RERI})}\right\}$.

Similar formulae hold for odds ratios using case–control data under a rare outcome: Once again, one simply replaces η with $(\eta - 1)$ in all relevant formulae in Section 13.3.3.

Similarly, to detect an epistatic interaction $(RERI > 2)$, with cohort data, at significance level α with power β when the true *RERI* is $\eta = e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$, the required sample size would be

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{RERI}}{(\eta - 2)^2}$$

The power to detect an epistatic interaction for a given sample size is $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + (\eta - 2)\sqrt{(n/V_{RERI})}\right\}$. Similar formulae hold for odds ratios using case–control data under a rare outcome as in Section 13.3.3: One simply replaces η with $(\eta - 2)$ in all relevant formulae.

As discussed in Chapters 9 and 10, if it can be assumed that the effects of both exposures are positive "monotonic" in the sense that the counterfactuals Y_{ge} are nondecreasing in g and e for all individuals (i.e., the exposures never have protective effects on the outcome for any individual), then the tests $p_{11} - p_{10} - p_{01} + p_{00} > 0$ and $RERI > 0$ can be used to test for sufficient cause interaction. For epistatic interactions, if the effect of at least one of the exposures is positive monotonic (Y_{ge} is nondecreasing in at least one of g and e), then $p_{11} - p_{10} - p_{01} > 0$ suffices and the tests for $RERI > 1$ could be used to test for an epistatic interaction; if the effect of both exposures are positive monotonic, then $p_{11} - p_{10} - p_{01} + p_{00} > 0$ suffices and the tests for $RERI > 0$ could be used to test for an epistatic interaction. To interpret interaction estimates causally, or to draw conclusions about sufficient cause or epistatic interaction, control must be made for confounding for both exposures.

13.5. EXCEL SPREADSHEETS FOR SAMPLE-SIZE AND POWER CALCULATIONS FOR ADDITIVE AND MULTIPLICATIVE INTERACTION FOR A BINARY OUTCOME

VanderWeele (2012c) provided two Excel spreadsheets that will automatically perform power and sample-size calculations for additive and multiplicative interaction for (i) cohort and (ii) case–control and case-only data. All of these spreadsheets return the sample size and power calculations above for the Wald test statistic for additive or multiplicative interaction with variance calculated under the alternative (cf. Demidenko, 2008; VanderWeele, 2011g).

13.5.1. Power Calculations for Cohort Studies

The first spreadsheet performs power and sample-size calculations for additive and multiplicative interaction for cohort data. For the power calculations, the user has the option of entering marginal exposure probabilities and the odds ratio relating the prevalence of both exposures (Sheet 1) or the joint exposure probabilities (Sheet 2). On Sheet 1, the user inputs the significance level of the test (alpha), the sample size (n), the probability of the outcome in the doubly unexposed reference group (p00), the main effect odds ratio for the first exposure (OR10), the main effect odds ratio for the second exposure (OR01), the odds ratio multiplicative interaction (IOR = $OR_{11}/(OR_{10} * OR_{01})$), the marginal prevalence of the first exposure ($P(G = 1)$), the marginal prevalence of the second exposure ($P(E = 1)$), and the odds ratio relating the dependence between the two exposures (OR_GE). The Excel spreadsheet returns both one-sided power (to detect positive

interaction) and two-sided power (to detect positive or negative interaction) for (i) additive interaction on the risk difference scale, (ii) multiplicative interaction on the risk ratio scale, (iii) multiplicative interaction on the odds ratio scale, (iv) additive interaction using the relative excess risk due to interaction (RERI; cf. Hosmer and Lemeshow, 1992) for risk ratios, and (v) additive interaction using the relative excess risk due to interaction for odds ratios, assuming a rare outcome. On Sheet 2, the user specifies the same inputs except that instead of the marginal probabilities and odds ratio relating the exposures ($P(G = 1)$, $P(E = 1)$, OR_GE), the user specifies the joint exposure probabilities for each of the four possible exposure combinations (in the Excel spreadsheet these are pi00, pi10, pi01, pi11). The Excel spreadsheet then again returns items (i)–(v) above.

Note that we may sometimes want to specify the probabilities of the outcome, $p_{ge} = P(Y = 1|G = g, E = e)$, in each of the $G \times E$ categories (i.e., $p_{00}, p_{01}, p_{10}, p_{11}$), rather the probability of the outcome in the doubly unexposed reference group (p00), the main effect odds ratio for the first exposure (OR10), the main effect odds ratio for the second exposure (OR01), and the odds ratio multiplicative interaction (IOR = $OR_{11}/(OR_{10} * OR_{01})$). If we do want to proceed with specifying the probabilities of the outcome, $p_{00}, p_{01}, p_{10}, p_{11}$, in each of the $G \times E$ categories instead, we could use the same spreadsheet and we enter the probability of the outcome in the doubly unexposed reference group, p_{00}, as "p00" in the spreadsheet; we calculate $\frac{p_{10}}{1-p_{10}}/\frac{p_{00}}{1-p_{00}}$ and enter this as "OR10"; we calculate $\frac{p_{01}}{1-p_{01}}/\frac{p_{00}}{1-p_{00}}$ and enter this as "OR01"; and we calculate $\frac{p_{11}}{1-p_{11}}\frac{p_{00}}{1-p_{00}}/\{\frac{p_{10}}{1-p_{10}}\frac{p_{01}}{1-p_{01}}\}$ and enter this as "IOR." The Excel spreadsheet will then once again carry out the desired power calculations.

13.5.2. Sample-Size Calculations for Cohort Studies

For sample-size calculations from cohort data, the user has the option of entering marginal exposure probabilities and the odds ratio relating the prevalence of both exposures (Sheet 3) or the joint exposure probabilities (Sheet 4). The user specifies exactly the same parameters as the spreadsheet for power calculations for cohort data except that instead of specifying the sample size, the power is specified (Power), and the Excel spreadsheet returns the required sample size for a test of the specified significance level and power to detect (i) additive interaction on the risk difference scale, (ii) multiplicative interaction on the risk ratio scale, (iii) multiplicative interaction on the odds ratio scale, (iv) additive interaction using the relative excess risk due to interaction (RERI) for risk ratios, and (v) additive interaction using the relative excess risk due to interaction for odds ratios, assuming a rare outcome.

13.5.3. Power Calculations for Case–Control and Case-Only Studies

The second spreadsheet performs power and sample-size calculations for additive and multiplicative interaction for case–control and case-only data. For power

calculations (Sheet 1), the user inputs the significance level of the test (alpha), the number of cases (n Cases), and number of controls (n Controls), the main effect odds ratio for the first exposure (OR10), the main effect odds ratio for the second exposure (OR01), the odds ratio multiplicative interaction (IOR), the marginal prevalence of the first exposure ($P(G = 1)$), the marginal prevalence of the second exposure ($P(E = 1)$), and the odds ratio relating the dependence between the two exposures (OR_GE). The Excel spreadsheet returns both one-sided power (to detect positive interaction) and two-sided power (to detect positive or negative interaction) for (i) additive interaction using the relative excess risk due to interaction (RERI) for odds ratios and (ii) multiplicative interaction on the odds ratio scale. If the two exposures are specified as independent (i.e., if OR_GE is specified as 1), then the spreadsheet will also return the power for the case-only estimator of multiplicative interaction (cf. Piegorsch et al., 1994; Yang et al., 1999) based on the number of cases. If the two exposures are not specified as independent (i.e., if OR_GE is specified as any number other than 1), the spreadsheet will return "#DIV/0!" for the power for the case-only estimator indicating that the case-only test is inapplicable in this setting because the two exposures are not independent. All power calculations for the case–control and case-only power spreadsheet make a rare outcome assumption. The power calculations are based on the variance calculated under the alternative (as in Demidenko (2008) for logistic regression multiplicative interactions and VanderWeele (2011g) for case-only multiplicative interactions) rather the variance calculated under the null, because the variance under the alternative corresponds to the test statistics that are commonly used in practice.

13.5.4. Sample-Size Calculations for Case–Control and Case-Only Studies

For sample-size calculations for additive and multiplicative interaction for case–control and case-only data (Sheet 2), the user inputs the significance level of the test (alpha), the proportion of cases in the case–control sample (Cs/(Cs+Cont)), the desired power of the test (Power), the main effect odds ratio for the first exposure (OR10), the main effect odds ratio for the second exposure (OR01), the odds ratio multiplicative interaction (IOR), the marginal prevalence of the first exposure ($P(G = 1)$), the marginal prevalence of the second exposure ($P(E = 1)$) and the odds ratio relating the dependence between the two exposures (OR_GE). The Excel spreadsheet returns the required sample size for a test of the specified significance level and power for (i) additive interaction using the relative excess risk due to interaction (RERI) for odds ratios and (ii) multiplicative interaction on the odds ratio scale. If the two exposures are specified as independent (i.e., if OR_GE is specified as 1), then the spreadsheet will also return the required sample size (i.e., number of cases), to detect multiplicative interaction for the case-only estimator of multiplicative interaction. If the two exposures are not specified as independent (i.e., if OR_GE is specified as any number other than 1), the spreadsheet will return "#DIV/0!" for the required sample size for the case-only estimator indicating that

the case-only test is inapplicable in this setting because the two exposures are not independent. All power calculations for the case–control and case-only sample-size spreadsheet make a rare outcome assumption. The sample-size calculations are based on the variance calculated under the alternative because this corresponds to the test statistics that are commonly used in practice (cf. Garcia-Closas and Lubin, 1999; Demidenko, 2008; VanderWeele, 2011g).

13.6. DISCUSSION

In this chapter we have given sample-size and power formulae for additive and multiplicative interaction in a variety of scenarios. We saw that when the main effects were both positive, then the power to detect positive interaction on the additive scale was in general greater than on the multiplicative scale. We have also discussed how the sample-size and power calculations for the relative excess risk due to interaction can be easily modified to provide sample-size and power calculations for mechanistic interaction corresponding to notions of synergism in the sufficient cause framework and to notions of compositional epistasis in genetics.

We have focused here on cohort, case–control, and case-only data; but as discussed in the Chapter 12, other study designs, such as matched case–control studies (cf. Gauderman, 2002a,b), and other methods for testing such as the joint tests of Chapter 12, are also sometimes used. Software is also available to implement power and sample-size calculations for a number of these other settings. Windows-based QUANTO, developed by Gauderman, is available at http://hydra.usc.edu/gxe and will implement sample-size calculations for likelihood ratio-based tests of interaction using various study designs, and the reader is referred there for further information.

As is often the case with analytic formulae for sample-size and power calculations, we have not considered the consequences of control for additional covariates. In settings in which these covariates are associated with the outcome but independent of the exposures (e.g., if the exposures were both randomized), adjustment for additional covariates should increase the power of tests (Robinson and Jewell, 1991) and in such cases the sample-size and power calculations here could be considered conservative estimates.

PART THREE

Synthesis and Spillover Effects

14

A Unification of Mediation and Interaction

Part I of this book was concerned with mediation, and Part II was concerned with interaction. In Part I we considered methods to decompose a total effect into a direct effect and an indirect effect. In Chapter 9 of Part II we considered how to go about assessing how much of an effect is due to an interaction. In this chapter we will consider the relation between these various methods for effect decomposition and attribution and consider also an approach that more fully encompasses and assesses mediation and interaction simultaneously. We show that the overall effect of an exposure on an outcome, in the presence of a mediator with which the exposure may interact, can be decomposed into four components: (i) the effect of the exposure in the absence of the mediator, (ii) the interactive effect when the mediator is left to what it would be in the absence of exposure, (iii) a mediated interaction, and (iv) a pure mediated effect. These four components respectively correspond to the portion of the effect that is due to neither mediation nor interaction, to just interaction (but not mediation), to both mediation and interaction, and to just mediation (but not interaction). We will see that this four-way decomposition unites methods that attribute effects to interactions and methods that assess mediation. Different combinations of these four components correspond to measures for mediation, while other combinations correspond to measures of interaction. The four-way decomposition can be carried out using standard statistical models, and software is provided to estimate each of the four components. The four-way decomposition provides the greatest insight into how much of an effect is mediated, how much is due to interaction, how much is due to both mediation and interaction together, and how much is due to neither.

14.1. NOTATION AND DEFINITIONS

As in Part I of this book, we will let A denote the exposure of interest, Y the outcome, and M a potential mediator, and let C denote a set of baseline covariates. We will suppose we want to compare two levels of the exposure, a and a^*; for a binary exposure we would have $a = 1$ and $a^* = 0$. For simplicity we will consider the setting of a binary exposure and binary mediator; however, more general results that are applicable to arbitrary exposures and mediators are given in the Appendix. SAS code is also provided at the end of the chapter to carry out the four-way decomposition in various settings. As in previous chapters when considering counterfactual notation, we let Y_a and M_a denote respectively the values of the outcome and mediator that would have been observed had the exposure A been set to level a. The total effect (TE) of the exposure A on the outcome Y is defined as $Y_1 - Y_0$; the total effect of the exposure A on the mediator M is defined as $M_1 - M_0$. As in prior sections on mediation in this book, we will also need counterfactuals of another form. Let Y_{am} denote the value of the outcome that would have been observed had A been set to level a, and M to m. Counterfactuals of the form Y_{am} consider hypothetical interventions on both the exposure and the mediator. The controlled direct effect, comparing exposure level $A = 1$ to $A = 0$ and fixing the mediator to level m, is defined by $Y_{1m} - Y_{0m}$ and captures the effect of exposure A on outcome Y, intervening to fix M to m; it may be different for different levels of m (Robins and Greenland, 1992; Pearl, 2001). It may also be different for persons. Finally we will also later consider counterfactuals of the form $Y_{aM_{a^*}}$, which is the outcome Y that would have occurred if we fixed A to a and we fixed M to the level it would have taken if A had been a^*.

14.2. FOURFOLD DECOMPOSITION: THE UNIFICATION OF MEDIATION AND INTERACTION

We show in the Appendix that we can decompose the total effect (TE) of A on Y into the following four components (VanderWeele, 2014):

$$\begin{aligned} Y_1 - Y_0 = {} & (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ & + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0) \end{aligned} \tag{14.1a}$$

We will consider the interpretation of these four components one at a time. The first component, $(Y_{10} - Y_{00})$, is the direct effect of the exposure A if the mediator were removed, that is, fixed to $M = 0$. As in Chapter 2, the first component is referred to as a "controlled direct effect" (CDE)(Robins and Greenland, 1992; Pearl, 2001). We will call the second component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, "a reference interaction" (INT_{ref}). The term $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$ is an additive interaction as in Chapter 9. It can be rewritten as $(Y_{11} - Y_{00}) - \{(Y_{10} - Y_{00}) + (Y_{01} - Y_{00})\}$ and will be nonzero for a person if the effect on the outcome of setting both the exposure and the mediator to present differs from the sum of the effects of having

only the exposure present and the effect of having only the mediator present. The second component in the decomposition in (14.1) is the product of this additive interaction and M_0. Thus this second component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, is an additive interaction that only operates if the mediator is present in the absence of exposure—that is, when $M_0 = 1$. The third component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, will be referred to as a "mediated interaction" (INT_{med}) as in Chapter 7. It is the same additive interaction contrast, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$, but now multiplied by $(M_1 - M_0)$. In other words, it is an additive interaction that only operates if the exposure has an effect on the mediator so that $M_1 - M_0 \neq 0$. The final component, $(Y_{01} - Y_{00})(M_1 - M_0)$, is the effect of the mediator in the absence of the exposure, $Y_{01} - Y_{00}$, multiplied by the effect of the exposure on the mediator itself, $M_1 - M_0$. It will be nonzero only if the mediator affects the outcome when the exposure is absent, and the exposure itself affects the mediator. We might refer to this final component as a "mediated main effect" or, as will be explained below, it turns out also to be the "pure indirect effect" (PIE) of Chapter 2 (Robins and Greenland, 1992; Pearl, 2001).

The intuition behind this decomposition is that if the exposure affects the outcome for a particular individual, then at least one of four things must be the case. The exposure might affect the outcome through pathways that do not require the mediator (i.e., the exposure affects the outcome even when the mediator is absent); in other words, the first component is nonzero. Or alternatively, the exposure effect might operate only in the presence of the mediator (i.e., there is an interaction) and it might also be the case that the exposure itself is not necessary for the mediator to be present (i.e., the mediator itself would be present in the absence of the exposure, though the mediator is itself necessary for the exposure to have an effect); in other words, the second component is nonzero. Or alternatively, the exposure effect might operate only in the presence of the mediator (i.e., there is an interaction) and it might also be the case that the exposure itself is in fact needed for the mediator to be present (i.e., the exposure causes the mediator, and the presence of the mediator is itself necessary for the exposure to have an effect); in other words, the third component is nonzero. Or finally, it might alternatively be the case that the mediator can cause the outcome in the absence of the exposure, but the exposure is necessary for the mediator itself to be present; in other words, the fourth component is nonzero. The decomposition above provides a mathematical formalization of this intuition. We could thus rewrite our decomposition as

$$TE = CDE + INT_{ref} + INT_{med} + PIE$$

As with the total effect of the exposure on the outcome, $Y_1 - Y_0$, we cannot in general hope to know the value of each of the four components for a particular individual, but below we will discuss assumptions under which we could estimate measures of these four components on average for a particular population. We will see below that under certain assumptions about confounding, assumptions identical to those in Chapter 2, the average value of each of four components is given by

the following empirical expressions:

$$\begin{aligned}
\mathbb{E}[CDE] &= (p_{10} - p_{00}) \\
\mathbb{E}[INT_{ref}] &= (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\
\mathbb{E}[INT_{med}] &= (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\
\mathbb{E}[PIE] &= (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}
\end{aligned}$$

where $p_{am} = \mathbb{E}(Y|A = a, M = m)$. If we let $p_a = \mathbb{E}(Y|A = a)$, we will have following empirical decomposition:

$$\begin{aligned}
p_{a=1} - p_{a=0} = {} & (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\
& + (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\
& + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \qquad (14.1b)
\end{aligned}$$

With such average measures, we would be able to assess how much of the total effect is due to (i) neither mediation nor interaction (the first component); (ii) how much is due to interaction but not mediation (the second component); (iii) how much is due to both mediation and interaction (the third component); and (iv) how much of the effect is due to mediation but not interaction (the fourth component). The four components of the total effect are summarized in Table 14.1.

If we let $\mathbb{E}[TE]$ denote the average total effect for the population (equal to $p_{a=1} - p_{a=0} = \mathbb{E}(Y|A = 1) - \mathbb{E}(Y|A = 0)$ in the absence of confounding), then we could also consider the proportion of the total effect that is due to each of these four components using the ratios $\frac{\mathbb{E}[CDE]}{\mathbb{E}[TE]}$, $\frac{\mathbb{E}[INT_{ref}]}{\mathbb{E}[TE]}$, $\frac{\mathbb{E}[INT_{med}]}{\mathbb{E}[TE]}$, and $\frac{\mathbb{E}[PIE]}{\mathbb{E}[TE]}$. We could also assess the overall proportion due to mediation by summing the proportions due to the mediated interaction and to the pure indirect effect, that is, $\frac{\mathbb{E}[INT_{med}]+\mathbb{E}[PIE]}{\mathbb{E}[TE]}$. We could likewise assess the overall proportion due to interaction by summing the proportions due to the reference interaction and to the mediated interaction, that is, $\frac{\mathbb{E}[INT_{ref}]+\mathbb{E}[INT_{med}]}{\mathbb{E}[TE]}$. Below, we discuss the relation of these measures to other measures considered earlier in this book.

We will first consider the no-confounding assumptions that allow us to estimate these four components on average, along with statistical methods to carry out such estimation. We will later consider the relationships between this fourfold decomposition and other concepts from the literatures on mediation and interaction that involve effect decomposition and attribution. We will see that the various other measures from the literatures on mediation and interaction essentially consist of various combinations of these four components.

14.3. IDENTIFICATION OF THE EFFECTS

Our discussion thus far has been primarily conceptual. As we have noted, the individual level effects in the four-way decomposition cannot be identified from the data, but under certain no-confounding assumptions the four components can

Table 14-1. The Four Basic Components of the Total Effect (the following four components sum to the total effect $TE = Y_1 - Y_0$)

Effect	**Counterfactual Definition**	**Empirical Analogue**	**Interpretation**
Controlled direct effect (CDE)	$(Y_{10} - Y_{00})$	$(p_{10} - p_{00})$	Due Neither to Mediation nor Interaction
Reference interaction (INT_{ref})	$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$	$(p_{11} - p_{10} - p_{01} + p_{00})P(M = 1\|A = 0)$	Due to Interaction Only
Mediated interaction (INT_{med})	$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$	$(p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1\|A = 1) - P(M = 1\|A = 0)\}$	Due to Mediation and Interaction
Pure indirect effect (PIE)	$(Y_{01} - Y_{00})(M_1 - M_0) = (Y_{0M_1} - Y_{0M_0})$	$(p_{01} - p_{00})$ $\{P(M = 1\|A = 1) - P(M = 1\|A = 0)\}$	Due to Mediation Only

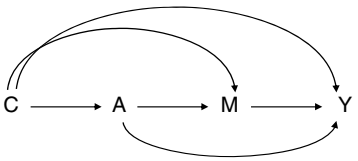

Figure 14.1 Mediation with exposure A, outcome Y, mediator M, and confounders C.

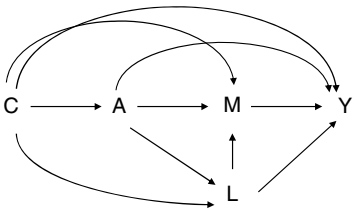

Figure 14.2 Mediation with a mediator-outcome confounder L that is affected by the exposure.

be identified from the data on average for a population. However, the same four assumptions as used in Chapter 2 also suffice to identify each of the four components from the data, namely: (A2.1) the effect the exposure A on the outcome Y is unconfounded conditional on the measured covariates C; (A2.2) the effect the mediator M on the outcome Y is unconfounded conditional on C; (A2.3) the effect the exposure A on the mediator M is unconfounded conditional on C; and (A2.4) none of the mediator–outcome confounders are themselves affected by the exposure. These are the same four assumptions we used in Chapter 2 for the two-way decomposition of a total effect into a natural direct effect and a natural indirect effect. As before, assumption (A2.4) requires that none of the mediator–outcome confounders are themselves affected by the exposure. This assumption would hold in Figure 14.1 but would be violated in Figure 14.2. If these four assumptions held without covariates, then we would have the empirical formulae given above:

$$\begin{aligned}
\mathbb{E}[CDE] &= (p_{10} - p_{00}) \\
\mathbb{E}[INT_{ref}] &= (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\
\mathbb{E}[INT_{med}] &= (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\
\mathbb{E}[PIE] &= (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}
\end{aligned}$$

More general formulae involving covariates and with arbitrary exposures and mediator (rather than binary) are given in the Appendix.

As discussed in Chapter 7, the counterfactual statement of assumption (A2.4) is somewhat more complicated and controversial because it involves what are sometimes called "cross-world" independencies. As discussed in Chapter 7, the interpretation as "none of the mediator–outcome confounders are themselves affected by the exposure" would apply in a causal diagram interpreted as a nonparametric structural equation model (Pearl, 2009) and would then hold in Figure 14.1, but

may not hold under other interpretations of causal diagrams (Robins and Richardson, 2010). See Section 3 of Chapter 7 for further discussion. We noted above that the empirical equivalent of our four-way decomposition was

$$\begin{aligned} p_{a=1} - p_{a=0} = {} & (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\ & + (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ & + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \qquad (14.1b) \end{aligned}$$

As discussed in the Appendix, this decomposition holds without any assumptions at all about confounding. However, to interpret each of the components causally does require assumptions about confounding. Assumptions (A2.1)–(A2.4) above allow for interpreting each of the components as population average causal effects of each of the four components in the four-way individual level counterfactual decomposition: *CDE*, INT_{ref}, INT_{med}, and *PIE*. In the Appendix we also discuss how a slightly weaker interpretation, analogous to the discussion in Section 7.3 of Chapter 7, is also valid essentially under just assumptions (A2.1)–(A2.3) alone, without requiring the more controversial assumption (A2.4).

Also of interest is the fact that the controlled direct effect, *CDE*, only requires assumption (A2.1) and (A2.2) to be identified (Robins and Greenland, 1992; Pearl, 2001). This thus also does not require the more controversial cross-world independence assumptions. The controlled direct effect is sometimes subtracted from the total effect to get a proportion eliminated measure $PE := TE - CDE$. Whenever we can identify the total effect and the controlled direct effect, we can calculate this portion eliminated measure. Interestingly, the four-way decomposition gives a more mechanistic interpretation of this portion-eliminated measure: The portion eliminated is the sum of the reference interaction, the mediated interaction, and the pure indirect effect ($PE = INT_{ref} + INT_{med} + PIE$); that is, it is the portion due to either mediation or interaction or both. We cannot empirically separate apart these three components without using stronger assumptions such as (A2.1)–(A2.4) above. However, whenever we can identify the total effect and the controlled direct effect (which we can do under much weaker assumptions), we can obtain also the sum of the three other components since they are simply the difference between the total effect and the controlled direct effect.

14.4. RELATION TO STATISTICAL MODELS

Suppose that assumptions (A2.1)–(A2.4) hold, that Y and M are continuous, and that the following regression models for Y and M are correctly specified:

$$\begin{aligned} \mathbb{E}[Y|a, m, c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ \mathbb{E}[M|a, c] &= \beta_0 + \beta_1 a + \beta_2' c \end{aligned}$$

For exposure levels a and a^*, and for setting the mediator to 0 in the controlled direct effect (see Appendix for derivations and for other settings of the mediator for

the *CDE*), the four components are given by

$$\begin{aligned}\mathbb{E}[CDE|c] &= \theta_1(a-a^*)\\ \mathbb{E}[INT_{ref}|c] &= \theta_3(\beta_0+\beta_1 a^*+\beta_2' c)(a-a^*)\\ \mathbb{E}[INT_{med}|c] &= \theta_3\beta_1(a-a^*)(a-a^*)\\ \mathbb{E}[PIE|c] &= (\theta_2\beta_1+\theta_3\beta_1 a^*)(a-a^*)\end{aligned}$$

If the exposure were binary, the pure direct, pure indirect, and mediated interactive effects would, respectively, simply be $\theta_1, \theta_3(\beta_0+\beta_2' c)\}, \theta_2\beta_1$, and $\theta_3\beta_1$. Standard errors for estimators of these quantities could be derived using the delta method along the lines of VanderWeele and Vansteelandt (2009) or by using bootstrapping. SAS code to implement this approach to obtain estimates and confidence intervals is given at the end of this chapter. Under our confounding assumptions (A2.1)–(A2.4), we can easily estimate these four components on average. The Appendix likewise provides a straightforward modeling approach to the four-way decomposition when the mediator is binary rather continuous, and SAS code is again given below (cf. VanderWeele, 2014).

14.5. BINARY OUTCOMES AND THE RATIO SCALE

Thus far we have been considering the definition of these four components on a difference scale. Often in epidemiology risk ratios or odds ratios are used for convenience, or ease of interpretation, or to account for study design. By dividing the decomposition in (14.1b) by $p_{a=0}$, we can rewrite this decomposition on the ratio scale as

$$\begin{aligned}RR_{a=1}-1 = {}& \kappa(RR_{10}-1)+\kappa(RR_{11}-RR_{10}-RR_{01}+1)P(M=1|A=0)\\ &+\kappa(RR_{11}-RR_{10}-RR_{01}+1)\{P(M=1|A=1)-P(M=1|A=0)\}\\ &+\kappa(RR_{01}-1)\{P(M=1|A=1)-P(M=1|A=0)\} \qquad (14.2)\end{aligned}$$

where $RR_{a=1}=\frac{p_{a=1}}{p_{a=0}}$ is the relative risk for exposure A comparing $A=1$ to the reference category $A=0$, and $RR_{am}=\frac{p_{am}}{p_{00}}$ is the relative risk for comparing categories $A=a, M=m$ to the reference category $A=0, M=0$, and where κ is a scaling factor which is given by $\kappa=\frac{p_{00}}{p_{a=0}}$. Note also here that the term, $(RR_{11}-RR_{10}-RR_{01}+1)$, is Rothman's excess relative risk due to interaction (*RERI*), described in Chapter 9, and is a measure of additive interaction using ratios (Rothman, 1986).

The decomposition in (14.2) involves decomposing the excess relative risk for the exposure A, $RR_{a=1}-1$, into four components on the excess relative risk scale involving, as before, (i) the controlled direct effect of A when $M=0$, (ii) a reference interaction, (iii) a mediated interaction, and (iv) a mediated main effect. Note that although the right-hand side of the decomposition involves a scaling factor κ, if what we are interested in is the proportion of the effect attributable to each of the components, then if we take any particular component and divide it by the

sum of all the components, then the scaling factor drops out. The proportion of the effect attributable to each of the four components is thus given by the expressions in Table 14.2.

The fourfold proportion attributable measures given in Table 14.2 allow us to estimate the proportion of the total effect attributable only to mediation (PA_{PIE}), just due to interaction (PA_{INTref}), due to both mediation and interaction (PA_{INTmed}), or due to neither mediation nor interaction (PA_{CDE}). Further technical details concerning the four-way decomposition on the ratio scale and for obtaining estimates and confidence intervals using logistic regression for the outcome along with linear regression for a continuous mediator or a second logistic regression for a binary mediator are given in the Appendix. SAS code to implement this approach is also given below (cf. VanderWeele, 2014). Once again, this can be done in a relatively straightforward manner.

14.6. ILLUSTRATION IN GENETIC EPIDEMIOLOGY

We will return to our example from genetic epidemiology which we also considered in the context of mediation in Chapter 2 to here illustrate the four-way decomposition. We will consider the extent to which the effect of chromosome 15q25.1 rs8034191 C alleles on lung cancer risk is mediated by cigarettes smoked per day and/or due to interaction with this smoking measure. rs8034191 C alleles had been found to be associated with both smoking and lung cancer (Amos et al., 2008; Hung et al., 2008; Thorgeirsson et al., 2008), but there had been debate as to whether the effects on lung cancer were direct or mediated by smoking. In Chapter 2, we assessed whether the effect was direct or indirect and found that most of the effect was not mediated by cigarettes per day (the total indirect effect was very small and the pure direct effect was large; cf. VanderWeele et al., 2012a). Here we will use the four-way decomposition to also assess how much of the pure direct effect is due to the effect of the variants in the absence of smoking and how much to the reference interaction. In large meta-analyses, Truong et al. (2010) found no association between the genetic variants amongst never smokers, suggesting strong interaction between the variants and smoking behavior; VanderWeele et al. (2012a) likewise reported statistical evidence of interaction. The analyses here will allow us to more fully assess the role of interaction in this context.

We use data on 1836 cases and 1452 controls from a lung cancer case–control study at Massachusetts General Hospital; see Miller et al. (2002) or VanderWeele et al. (2012a) for further details on the study. For the exposure we compare 2 versus 0 C alleles, and we use cigarettes per day as the mediator (the square root of this measure is used so that the measure is more normally distributed). Covariates adjusted for in the analysis include sex, age, education, and smoking duration. Analyses are restricted to Caucasians. Because the outcome, lung cancer, is rare, odds ratios approximate risk ratios. We fit a logistic regression model for lung cancer on the variants, smoking, their interaction, and the covariates; and we fit a linear regression model for smoking on the variants and covariates. Details of this modeling approach in the context of the four-way decomposition are given

Table 14-2. Proportion Attributable to the Controlled Direct Effect (PA_{CDE}), the Reference Interaction (PA_{INTref}), the Mediated Interaction (PA_{INTmed}), and the Pure Indirect Effect (PA_{PIE}) When Using a Ratio Scale

$$PA_{CDE} = \frac{(RR_{10} - 1)}{(RR_{10} - 1) + (RERI)P(M = 1|A = 1) + (RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}$$

$$PA_{INTref} = \frac{(RERI)P(M = 1|A = 0)}{(RR_{10} - 1) + (RERI)P(M = 1|A = 1) + (RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}$$

$$PA_{INTmed} = \frac{(RERI)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}{(RR_{10} - 1) + (RERI)P(M = 1|A = 1) + (RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}$$

$$PA_{PIE} = \frac{(RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}{(RR_{10} - 1) + (RERI)P(M = 1|A = 1) + (RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}.$$

in the Appendix. The overall risk ratio comparing 2 versus 0 C alleles was 1.768 (95% CI: 1.33, 2.21) for an excess relative risk of $1.768 - 1 = 0.768$ (95% CI: 0.33, 1.21). We decompose this excess relative risk into the four components. The component due to the pure indirect effect is 0.014 (95% CI: −0.008, 0.036); the component due to the mediated interaction is 0.034 (95% CI: −0.019, 0.087); the component due to the reference interaction is 0.42 (95% CI: 0.11, 0.73); and the component due to the controlled direct effect (if smoking were fixed to 0) is 0.30 (95% CI: −0.19, 0.79). The four components sum to the excess relative risk: $0.014 + 0.034 + 0.42 + 0.30 = 0.768$. Of the four components, only the reference interaction is statistically significant, highlighting the important role of interaction in this context. The overall proportion mediated (the sum of the pure indirect effect and the mediated interaction, divided by the excess relative risk) is quite small, 6.2% (95% CI: −2.7%, 15.1%), as had been indicated in the analyses of Chapter 2 (cf. VanderWeele et al., 2012a). The overall proportion attributable to interaction (the reference interaction plus the mediated interaction, divided by the excess relative risk) is relatively substantial 59.2% (95% CI: 9.2%, 109.3%). Mediation may play a role here (and probably does because the variants do affect smoking and smoking affects lung cancer); but interaction, between the variants and smoking, is clearly much more important in this context.

14.7. RELATION TO MEDIATION DECOMPOSITIONS

We will now discuss the relations between the four components above and concepts from the mediation and interaction analysis literature. Some of the discussion below is technical; the relations between the decompositions are summarized graphically in Figure 14.3 below. A reader less interested in the technical details could skip to the discussion of software implementation in Section 14.9. As above, our fourfold decomposition is

$$\begin{aligned} Y_1 - Y_0 = {} & (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ & + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0) \end{aligned}$$

The first component, $(Y_{10} - Y_{00})$, is referred to in the mediation analysis literature as a "controlled direct effect" (CDE) of the exposure when fixing the mediator to level $M = 0$. The fourth component in the four-way decomposition, $(Y_{01} - Y_{00})(M_1 - M_0)$, what we referred to above as a "mediated main effect," is in fact equivalent to what in the mediation analysis literature is referred to as a "pure indirect effect" (PIE). It is shown in the Appendix that

$$PIE := Y_{0M_1} - Y_{0M_0} = (Y_{01} - Y_{00})(M_1 - M_0)$$

The counterfactual contrast, $Y_{0M_1} - Y_{0M_0}$, in the mediation analysis literature is referred to as a "pure indirect effect" (Robins and Greenland, 1992) or as a type of "natural direct effect" (Pearl, 2001) as in Chapters 2 and 7. The contrast $Y_{0M_1} - Y_{0M_0}$ compares what would happen to the outcome if the mediator were changed from the level M_0 (the level it would be in the absence of the exposure)

to M_1 (the level it would be in the presence of exposure) while in both counterfactual scenarios fixing the exposure itself to be absent. It will be nonzero if and only if the exposure changes the mediator (so that M_0 and M_1 are different), and the mediator itself has an effect on the outcome even in the absence of the exposure. However, this is, in fact, the same quantity as what we had in our four-way decomposition above, namely $(Y_{01} - Y_{00})(M_1 - M_0)$. The third component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, was recently considered in the mediation analysis literature and called a "mediated interaction" (INT_{med})(VanderWeele, 2013b) as described in Chapter 7. The component we have not yet considered, the second component, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, what we referred to above as a "reference interaction" (INT_{ref}), has no analogue in the current literature. However, it is shown in the Appendix that the sum of the first and second component does have an analogue in the mediation analysis literature, and it is equal to what is sometimes called in the mediation analysis literature the "pure direct effect" (PDE), as in Chapters 2 and 7, defined as $Y_{1M_0} - Y_{0M_0}$. This compares what would happen to the outcome in the presence versus the absence of the exposure if in both cases the mediator were set to whatever it would be for that individual in the absence of exposure. However, this pure direct effect is in fact equal to the sum of our first components in the four-way decomposition above. In other words, we have that

$$\begin{aligned} PDE := Y_{1M_0} - Y_{0M_0} &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ &= CDE + INT_{ref} \end{aligned}$$

The pure direct effect is the sum of a controlled direct effect and, our second component, the reference interaction. If in our four-way decomposition above we replace the first two components with the pure direct effect and write the fourth component as the pure indirect effect, we obtain

$$Y_1 - Y_0 = PDE + INT_{med} + PIE \tag{14.3}$$

In other words, we can decompose the total effect into a pure direct effect, a pure indirect effect, and a mediated interaction. This decomposition in (14.3) was the three-way decomposition discussed in Chapter 7 (cf. VanderWeele, 2013b). However, even this three-way decomposition is relatively new, and in causal inference literature on mediation the two-way decomposition considered in Chapter 2 has generally been the one that has been used.

The two-way mediation decompositions in the literature also follow from the four-way decomposition above. As discussed in Chapter 7, it is also the case that the sum of the mediated interaction and the pure indirect effect is in fact equal to what in the mediation analysis literature is sometimes called a "total indirect effect" (TIE), defined as $Y_{1M_1} - Y_{1M_0}$, as in Chapters 2 and 7. Whereas the pure indirect effect, $Y_{0M_1} - Y_{0M_0}$, compares changing the mediator from M_0 to M_1 while fixing the exposure itself to be absent, the total indirect effect, $Y_{1M_1} - Y_{1M_0}$, compares changing the mediator from M_0 to M_1 fixing the exposure to present. With the total indirect so defined, we have $TIE = PIE + INT_{med}$, that is, $(Y_{1M_1} - Y_{1M_0}) = (Y_{0M_1} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$. We can then combine the

mediated interaction and the pure indirect effect in the decomposition in (14.3), into a total indirect effect to obtain the more standard two-way decomposition in the mediation analysis literature:

$$Y_1 - Y_0 = PDE + TIE \tag{14.4}$$

This is the decomposition that has been used most often when assessing direct and indirect effects and was the focus of Chapter 2; this two-way decomposition was first proposed in 1992 by Robins and Greenland; and it is the decomposition that most of the existing software packages for mediation have focused on (Imai et al., 2010a; Valeri and VanderWeele, 2013). However, as we have seen above, the pure direct effect is itself a combination of two components: a controlled direct effect and the reference interaction ($PDE = CDE + INT_{ref}$). And the total indirect effect is a combination of two components, the pure indirect effect and the mediated interaction ($TIE = PIE + INT_{med}$). These effects are not, in general, identified at the individual level; but under the confounding assumptions described above, they can be estimated on average for a population. When this is done, sometimes a proportion-mediated measure, $\frac{\mathbb{E}[TIE]}{\mathbb{E}[TE]} = \frac{\mathbb{E}[PIE]+\mathbb{E}[INT_{med}]}{\mathbb{E}[TE]}$, is used as discussed in Chapter 2.

We have also considered in Chapter 7 yet another mediation decomposition. Sometimes the mediated interaction in the decomposition in (14.3) is combined with pure direct effect, rather than with the pure indirect effect, for an alternative two-way decomposition. As discussed in Chapter 7, the sum of the mediated interaction and the pure direct effect is equal to what in the mediation analysis literature is sometimes called a "total direct effect" (TDE), defined as $Y_{1M_1} - Y_{0M_1}$. The total and the pure direct effects are sometimes also called "natural direct effects" (Pearl, 2001), and the total and the pure indirect effects are sometimes called "natural indirect effects" (Pearl, 2001). A summary of the various composite effects by combining different components of the four-way decomposition is given in Table 14.3.

Of interest here is that the total direct effect contains three components: the controlled direct effect, the reference interaction, and the mediated interaction. As we move from the first to third of these components, we see that they increasingly

Table 14-3. COMPOSITE EFFECTS

Effect	**Counterfactual Definition**	**Composite Relationship**
Total indirect effect (TIE)	$(Y_{1M_1} - Y_{1M_0})$	$TIE = PIE + INT_{med}$
Pure direct effect (PDE)	$(Y_{1M_0} - Y_{0M_0})$	$PDE = CDE + INT_{ref}$
Total direct effect (TDE)	$(Y_{1M_1} - Y_{0M_1})$	$TDE = CDE + INT_{ref} + INT_{med}$
Portion eliminated (PE)	$(Y_1 - Y_0) - (Y_{10} - Y_{00})$	$PE = PIE + INT_{ref} + INT_{med}$
Portion attributable to interaction (PAI)	$(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1)$	$PAI = INT_{ref} + INT_{med}$

involve the mediator in more substantial ways. The controlled direct effect, $(Y_{10} - Y_{00})$, operates completely independent of the mediator; for this to be nonzero, the direct effect must be present even when the mediator is absent. The reference interaction, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, requires the mediator to operate, but the effect does not come about by the exposure changing the mediator—it simply requires that the mediator itself is present even when the exposure is absent; the effect is "unmediated" in the sense that it does not operate by the exposure changing the mediator, but it requires the presence of the mediator nonetheless. The third component, the mediated interaction, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$, is a type of mediated effect; it requires that the exposure change the mediator; but it is also a direct effect insofar as an interaction must also be present (the effect of the exposure must be different for different levels of the mediator); the third component thus not only involves the mediator but it is a mediated effect, and a direct effect as well. It is for this reason that it is sometimes combined with the pure indirect effect to obtain the total indirect effect, and sometimes combined with the pure direct effect to obtain the total direct effect.

When we combine the pure direct effect and mediated interaction to get the total direct effect, $TDE := Y_{1M_1} - Y_{0M_1} = PDE + INT_{med}$, we have the alternative two-way decomposition of the total effect into the sum of the total direct effect and the pure indirect effect:

$$Y_1 - Y_0 = TDE + PIE \tag{14.5}$$

This decomposition was likewise proposed by Robins and Greenland in 1992. Relatively easy-to-use software (Imai et al., 2010a; Valeri and VanderWeele, 2013) is currently available to estimate the components of the two-way decompositions in (14.4) and (14.5) on average for a population, under the assumptions described above and considered in Chapter 2. Note that in the decomposition in (14.5), the total direct effect consists of three of the four basic components (the controlled direct effect, the reference interaction, and the mediated interaction), whereas the pure indirect effect constitutes a single component. The mediated interaction is, however, arguably part of the effect that is mediated and thus, when questions of mediation are of interest, it is arguably (14.4), rather than (14.5), that is to be preferred when assessing the extent of mediation (Suzuki et al., 2011; VanderWeele, 2011c, 2013b).

A final measure that is used in the mediation analysis literature is sometimes referred to as the "portion eliminated" (PE) (Robins and Greenland, 1992; cf. VanderWeele, 2013a) and was likewise discussed in Chapter 2. This is generally defined as the difference between the total effect and the controlled direct effect: $PE := (Y_1 - Y_0) - CDE$. It is the portion of the effect of the exposure that would remain if the mediator were fixed to 0. As discussed in Chapter 2, the portion eliminated may be of interest insofar as it allows one to assess how much of the effect of the exposure can be eliminated or prevented by intervening on the mediator; for this reason it is sometimes argued to be of policy interest (Robins and Greenland, 1992; Pearl, 2001; Hafeman and Schwartz, 2009; VanderWeele, 2013a).

Table 14-4. MEDIATION DECOMPOSITIONS

Number of Components	**Decomposition**
Two-way decomposition[a]	$TE = TIE + PDE$
Two-way decomposition[b]	$TE = TDE + PIE$
Two-way decomposition[c]	$TE = CDE + PE$
Three-way decomposition[d]	$TE = PDE + PIE + INT_{med}$
Four-way decomposition[e]	$TE = CDE + INT_{ref} + INT_{med} + PIE$

[a] $(Y_1 - Y_0) = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$
[b] $(Y_1 - Y_0) = (Y_{1M_1} - Y_{0M_1}) + (Y_{0M_1} - Y_{0M_0})$
[c] $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + [(Y_1 - Y_0) - (Y_{10} - Y_{00})]$
[d] $(Y_1 - Y_0) = (Y_{0M_1} - Y_{0M_0}) + (Y_{1M_0} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$
[e] $Y_1 - Y_0 = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0)$

The four-way decomposition above in fact shows that this portion eliminated measure is equal to the sum of the other three components: the reference interaction, the mediated interaction, and the pure indirect effect—that is, $PE = INT_{ref} + INT_{med} + PIE$—and we can write the total effect as $TE = CDE + PE$. The four-way decomposition provides a causal interpretation for the difference between the total effect and the controlled direct effect: It is the portion of the effect attributable to mediation, or interaction, or both. When the portion eliminated is estimated at the population level, sometimes a proportion-eliminated measure is also calculated as $\frac{\mathbb{E}[TE]-\mathbb{E}[CDE]}{\mathbb{E}[TE]}$, which we could also rewrite as $\frac{\mathbb{E}[INT_{ref}]+\mathbb{E}[INT_{med}]+\mathbb{E}[PIE]}{\mathbb{E}[TE]}$; note that this is different from the proportion-mediated measure considered earlier, which was $\frac{\mathbb{E}[NIE]}{\mathbb{E}[TE]} = \frac{\mathbb{E}[INT_{med}]+\mathbb{E}[PIE]}{\mathbb{E}[TE]}$. The proportion eliminated includes in the numerator the reference interaction (since this is eliminated if the mediator is removed); the proportion mediated does not include the reference interaction in the numerator (since this is not part of the mediated effect)(VanderWeele, 2013a).

Thus, we have seen a number of different decompositions. However, when we are interested in questions of mediation, we need not choose between the two-way decompositions, or even the three-way decomposition, but can in fact use the decomposition into four components above so as to assess the portion of the total effect that is attributable just to mediation, just to interaction, to both mediation and interaction, or to neither mediation nor interaction. The four-way decomposition allows us to accomplish this. The various decompositions within the context of mediation are summarized in Table 14.4, but the four-way decomposition here essentially provides a framework that encompasses them all.

14.8. RELATION TO INTERACTION DECOMPOSITIONS

In Section 9.12 we discussed methods that assess the portion of the total effect of one exposure on an outcome that is attributable to an interaction with a second

exposure. Here we will relate this to the four-way decomposition above. Our four-way decomposition above was expressed as

$$\begin{aligned} Y_1 - Y_0 &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ &\quad + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0) \end{aligned} \tag{14.1a}$$

which we also wrote as $TE = CDE + INT_{ref} + INT_{med} + PIE$. Suppose now that instead of considering how much of the total effect is mediated versus direct, as in the previous section, we were interested in the portion attributable to interaction. In our four-way decomposition, two of the four components (the second and the third) involve an interaction. We could thus define the portion due to interaction as their sum: $PAI := INT_{ref} + INT_{med} = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1)$ and we would then have the three-way decomposition:

$$\begin{aligned} TE &= CDE + PAI + PIE \\ &= (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1) + (Y_{01} - Y_{00})(M_1 - M_0) \end{aligned} \tag{14.6}$$

The total effect can be decomposed into the effect of A with M absent (CDE), a pure indirect effect (PIE), and a portion due to interaction (PAI). Consider now the empirical analogue of this decomposition using the expressions in (14.1b). We let $p_{am} = \mathbb{E}[Y|A = a, M = m]$, $p_a = \mathbb{E}[Y|A = a]$, $p_m = \mathbb{E}[Y|M = m]$ and we have the following from (14.1b):

$$\begin{aligned} (p_{a=1} - p_{a=0}) &= (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 1) \\ &\quad + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \end{aligned} \tag{14.7}$$

We again have the decomposition of the average total effect of A, into what is essentially the average controlled direct effect, the average portion attributable to interaction, and the average pure indirect effect (essentially a mediated main effect). The middle component is the component due to interaction and the proportion of the effect due to interaction could then be assessed by $(p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 1)/(p_{a=1} - p_{a=0})$.

In fact, the decomposition given above in (14.7) is that which VanderWeele and Tchetgen Tchetgen (2014) used when attributing effects to interactions. A few points are worth noting. We might consider two cases, one in which A and M were independent and one in which they are not. The decompositions in (14.6) and (14.7) are applicable even when A affects M. When A does not affect Y, we have an analogous individual counterfactual level decomposition as (14.6) then reduces to $TE = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M)$ since, when A does not affect M, $M_1 = M_0 = M$. Also when A does not affect M, and the distributions of A and M are statistically independent, the decomposition in (14.7) then reduces to $(p_{a=1} - p_{a=0}) = (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1)$ and we likewise have a similar decomposition for the total effect of M on Y: $(p_{m=1} - p_{m=0}) = (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(A = 1)$, which are the decompositions considered in Chapter 9 (cf. VanderWeele and Tchetgen Tchetgen, 2014). Likewise, on a ratio scale, when A does not affect M, the third and

Table 14-5. INTERACTION DECOMPOSITIONS

Number of Components	**Decomposition**
Two-way decomposition (No Mediation)[a]	$TE = CDE + PAI$
Three-way decomposition[b]	$TE = CDE + PAI + PIE$
Four-way decomposition[c]	$TE = CDE + INT_{ref} + INT_{med} + PIE$

[a] $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M)$
[b] $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1) + (Y_{01} - Y_{00})(M_1 - M_0)$
[c] $(Y_1 - Y_0) = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0)$

fourth components become 0 and we are left with $PA_{CDE} = \frac{(RR_{10}-1)}{(RR_{10}-1)+(RERI)P(M=1)}$ and $PA_{INTref} = \frac{(RERI)P(M=1)}{(RR_{10}-1)+(RERI)P(M=1)}$. However, once again, the four-way decomposition encompasses all of these measures. When A does affect M, we cannot use the decomposition in Chapter 9 for attributing effect to interaction but we can still use the decomposition in (14.7) to assess the portion of the total effect attributable to interaction. The four-way decomposition and the decompositions in (14.6) and (14.7) essentially generalize the approach considered in Chapter 9 to the setting in which it is also the case that A affects M. All of this follows from our four-way decomposition. These various decompositions for interaction are all summarized in Table 14.5.

Although in the more general setting when A affects M, we can estimate the portion due to interaction on the average level using the three-way decomposition in (14.6), there is no need to use only a three-way decomposition; we can instead use the four-way decomposition in (14.1a) and the empirical expressions in (14.1b) to further divide the portion due to interaction into that which is due to interaction but not mediation (the reference interaction, $\mathbb{E}[INT_{ref}]$) and the portion due to interaction and mediation (the mediated interaction $\mathbb{E}[INT_{med}]$). Such a four-way decomposition, in which the portion attributed to interaction is itself further divided, may shed additional insight.

Finally, and perhaps most importantly, this four-way decomposition, which helps better understand the portions of a total effect due to interaction, is exactly the same decomposition that was used above to shed insight into what portions of the total effect were mediated and which portions were direct. The same four-way decomposition was useful in assessing both mediation and interaction. The same four components are used in assessing mediation and interaction, but the components are combined in different ways to assess these different phenomena. However, the four-way decomposition itself essentially provides a unification of these phenomena of mediation and interaction. The fourfold decomposition underlies the various more specific decompositions in assessing both mediation and interaction.

As illustrated in Figure 14.3, the four components form the backbone of both the various mediation decompositions (Figures 14.3–14.5) and the interaction decomposition (Figure 14.3).

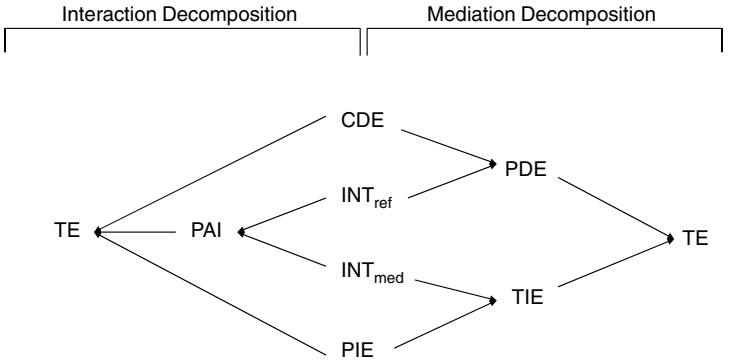

Figure 14.3 The four-fold decomposition encompasses both decompositions for mediation and interaction. For interaction, the reference interaction (INT_{ref}) and the mediated interaction (INT_{med}) combine to the portion attributable to interaction (PAI). The portion attributable to interaction (PAI) combine with the controlled direct effect (CDE) and the pure indirect effect (PIE) to give the total effect (TE). For mediation, the controlled direct effect (CDE) and the reference interaction (INT_{ref}) combine to give the pure direct effect (PDE); the pure indirect effect (PIE) combines with the mediated interaction (INT_{med}) to give the total indirect effect (TIE); and the pure direct effect (PDE) combines with total indirect effect (TIE) to give the total effect (TE).

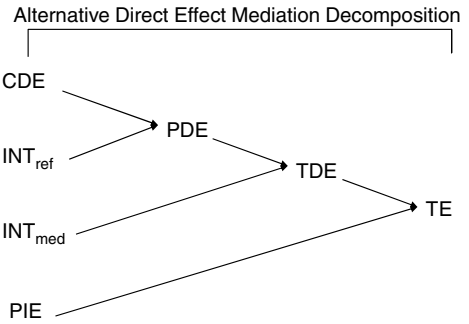

Figure 14.4 As an alternative mediation decomposition, the controlled direct effect (CDE) and the reference interaction (INT_{ref}) combine to give the pure direct effect (PDE); the pure direct effect (PDE) and the mediated interaction (INT_{med}) combine to give the total direct effect (TDE); and the total direct effect (TDE) and the pure indirect effect (PIE) combine to give the total effect (TE).

Once, again, however, the greatest insight is arguably gained when the fourfold approach is used to assess both simultaneously and the portions of the total effect that are due just to mediation, just to interaction, to both mediation and interaction, and to neither mediation nor interaction.

14.9. SAS CODE FOR THE FOUR-WAY DECOMPOSITION

In this section, we will provide SAS code to carry out the four-way decompositions with either continuous or binary outcomes and either continuous or binary

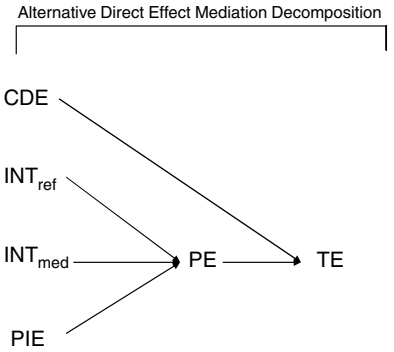

Figure 14.5 As an alternative mediation decomposition, the difference between the total effect (TE) and the controlled direct effect (CDE) is sometimes called the portion eliminated (PE) and it is equal to the sum of the reference interaction (INT_{ref}), the mediated interaction (INT_{med}), and the pure indirect effect (PIE).

mediators. In Section 14.4 we gave analytic expressions for each of the four components for linear regression models with continuous outcomes and continuous mediators. However, similar expressions can be derived under various other models and types of mediators and outcomes; several of these are derived in the Appendix. We have focused our discussion on a binary exposure and binary mediator. However, much more general results are given in the Appendix, and the approach in fact applies to arbitrary exposures and mediators. Here we provide code to implement the four-way decomposition and obtain confidence intervals for the effects in various settings.

In the discussion above, the controlled direct effect we have been considering is that in which the mediator is fixed to being absent. However, instead of focusing on a controlled direct effect that fixes the mediator to be absent, one can consider controlled direct effects that fix the mediator to some other level, m^*. Similar four-way decompositions can be carried out wherein the first component is the controlled direct effect with the mediator fixed to level m^*. When this is done, the reference interaction term changes because, with the mediator fixed to m^* (rather than 0), the controlled direct effect then picks up some of the effect of the interaction between the exposure and the mediator. With the controlled direct effect in which the mediator is fixed to m^*, the interpretation of the reference interaction is then the portion of the effect due to the interaction between the exposure and the mediator that is not mediated, and also not captured by the controlled direct effect. Again, the results in the Appendix cover very general settings. The code below can carry out the four-way decomposition with the mediator fixed to some other level m^* that may be different from 0.

14.9.1. Continuous Outcome, Continuous Mediator

To estimate the components of the four-way decomposition for the effect of exposure A on a continuous outcome Y with continuous mediator M under the regression models in Section 14.4, one can use the code below.

Suppose we have a dataset named "mydata" with outcome variable "y," exposure variables "a," mediator "m," and three covariates "c1," "c2," and "c3." If there were more or fewer covariates the user would have to modify the second, third, fourth, fifth, and tenth lines of the code below to include these covariates.

The user must input in the third line of code the two levels of A ("a1=" and "a0=") that are being compared (these are exposure levels 1 and 0 in the code below, but this could be modified for an ordinal or continuous exposure) and the level of $M = m^*$ ("mstar=") at which to compute the controlled direct effect and the remainder of the decomposition (it is assumed in the code below that the mediator is fixed to the value $M = m^* = 0$, but this could be modified). The user must also input in the third line of the code the value of the covariates C at which the effects are to be calculated ("cc1=", "cc2=," and "cc3="). Alternatively, the mean value of these covariates in the sample could be inputted on this line as a summary measure. The code below on line 3 specifies these as 10, 10, and 20, which should be altered according to the covariate values in the application of interest.

The output will include estimates and confidence intervals for the total effect as well as the four components of the total effect, that is, the controlled direct effect, the reference interaction, the mediated interaction, and the pure indirect effect; the output will also include (a) estimates and confidence intervals for the proportion of the total effect due to each of the four components and (b) estimates and confidence intervals for the overall proportion mediated, the overall proportion due to interaction, and the overall proportion of the effect that would be eliminated if the mediator M were fixed to the value m^*, specified by the user.

```
proc nlmixed data=mydata;
parms t0=0 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=0 b1=0 bc1=0 bc2=0
   bc3=0 ss_m=1 ss_y=1;
a1=1; a0=0; mstar=0; cc1=10; cc2=10; cc3=20;
mu_y=t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2 + tc3*C3;
mu_m =b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3;
ll_y= -((y-mu_y)**2)/(2*ss_y)-0.5*log(ss_y);
ll_m= -((m-mu_m)**2)/(2*ss_m)-0.5*log(ss_m);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
cde = (t1 + t3*mstar)*(a1-a0);
intref = t3*(b0 + b1*a0 + bcc - mstar)*(a1-a0);
intmed = t3*b1*(a1-a0)*(a1-a0);
pie = (t2*b1 + t3*b1*a0)*(a1-a0);
te = cde + intref + intmed + pie;
estimate 'Total Effect' te;
estimate 'CDE' cde;
estimate 'INTref' intref;
estimate 'INTmed' intmed;
estimate 'PIE' pie;
estimate 'Proportion CDE' cde/te;
estimate 'Proportion INTref' intref/te;
estimate 'Proportion INTmed' intmed/te;
estimate 'Proportion PIE' pie/te;
```

```
estimate 'Overall Proportion Mediated' (pie+intmed)/te;
estimate 'Overall Proportion Attributable to Interaction'
   (intref+intmed)/te;
estimate 'Overall Proportion Eliminated' (intref+intmed+pie)/te;
run;
```

14.9.2. Continuous Outcome, Binary Mediator

To estimate the components of the four-way decomposition for the effect of exposure A on a continuous outcome Y with binary mediator M under the regression models

$$\mathbb{E}[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M=1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

one can use the code below. The analytic expressions for each of the four components and derivations are given in the Appendix. The explanation of the code follows the one presented in Section 14.9.1.

```
proc nlmixed data=mydata;
parms t0=0 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=1 b1=0 bc1=0 bc2=0 bc3=0
   ss_y=1;
a1=1; a0=0; mstar=0; cc1=10; cc2=10; cc3=20;
mu_y=t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2 + tc3*C3;
p_m=(1+exp(-(b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3)))**-1;
ll_y= -((y-mu_y)**2)/(2*ss_y)-0.5*log(ss_y);
ll_m= m*log (p_m)+(1-m)*log(1-p_m);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
cde = (t1 + t3*mstar)*(a1-a0);
intref = t3*(a1-a0)*(exp(b0+b1*a0+bcc)/(1+exp(b0+b1*a0+bcc)) - mstar);
intmed = t3*(a1-a0)*(exp(b0+b1*a1+bcc)/(1+exp(b0+b1*a1+bcc))
   -exp(b0+b1*a0+bcc)/(1+exp(b0+b1*a0+bcc)));
pie = (t2 + t3*a0)*(exp(b0+b1*a1+bcc)/(1+exp(b0+b1*a1+bcc))
   -exp(b0+b1*a0+bcc)/(1+exp(b0+b1*a0+bcc)));
te = cde + intref + intmed + pie;
estimate 'Total Effect' te;
estimate 'CDE' cde;
estimate 'INTref' intref;
estimate 'INTmed' intmed;
estimate 'PIE' pie;
estimate 'Proportion CDE' cde/te;
estimate 'Proportion INTref' intref/te;
estimate 'Proportion INTmed' intmed/te;
estimate 'Proportion PIE' pie/te;
estimate 'Overall Proportion Mediated' (pie+intmed)/te;
estimate 'Overall Proportion Attributable to Interaction'
```

```
   (intref+intmed)/te;
estimate 'Overall Proportion Eliminated' (intref+intmed+pie)/te;
run;
```

14.9.3. Binary Outcome, Continuous Mediator

To estimate the components of the four-way decomposition on the ratio scale for the effect of exposure A on a binary outcome Y with continuous mediator M under the regression models

$$\text{logit}(P(Y=1|a,m,c)) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

with M normally distribution conditional on (A,C) with variance σ^2 and with outcome Y rare, one can use the code below. The analytic expressions for each of the four components and derivations are given in the Appendix.

Suppose we have a dataset named "mydata" with outcome variable "y," exposure variables "a," and mediator "m" and three covariates "c1," "c2," and "c3." If there were more or fewer covariates, the user would have to modify the second, third, fourth, fifth, and tenth lines of the code below to include these covariates.

The user must input in the third line of code the two levels of A ("a1=" and "a0=") that are being compared (these are exposure levels 1 and 0 in the code below, but this could be modified for an ordinal or continuous exposure) and the level of $M = m^*$ ("mstar=") at which to compute the controlled direct effect and the remainder of the decomposition (it is assumed in the code below that the mediator is fixed to the value $M = m^* = 0$, but this could be modified). The user must also input in the third line of the code the value of the covariates C at which the effects are to be calculated ("cc1=," "cc2=," and "cc3="). Alternatively, the mean value of these covariates in the sample could be inputted on this line as a summary measure. The code below on line 3 specifies these as 58.57, 1.44, and 0.34, which should be altered according to the covariate values in the application of interest.

The output will include estimates and confidence intervals for the total effect risk ratio, the excess relative risk (i.e., the relative risk minus 1), and the four components of the excess relative risk, that is, the excess relative risks due to the controlled direct effect, to the reference interaction, to the mediated interaction, and to the pure indirect effect; the output will also include (a) estimates and confidence intervals for the proportion of the excess relative risk due to each of the four components and (b) estimates and confidence intervals for the overall proportion mediated, the overall proportion due to interaction, and the overall proportion of the effect that would be eliminated if the mediator M were fixed to the value m^*, specified by the user.

```
proc nlmixed data=mydata;
parms t0=1 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=0 b1=0 bc1=0 bc2=0
  bc3=0 ss_m=1;
a1=1; a0=0; mstar=0; cc1=58.57; cc2=1.44; cc3=0.34;
```

```
p_y=(1+exp(-(t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2
   + tc3*C3)))**-1;
mu_m =b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3;
ll_m= -((m-mu_m)**2)/(2*ss_m)-0.5*log(ss_m);
ll_y= y*log (p_y)+(1-y)*log(1-p_y);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
CDE_comp = exp( t1*(a1-a0)+t2*mstar + t3*a1*mstar -
      (t2+t3*a0)*(b0+b1*a0+bcc)
   -  (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m )
   - exp(t2*mstar + t3*a0*mstar - (t2+t3*a0)*(b0+b1*a0+bcc) -
     (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m );
INTref_comp = exp((t1+t3*(b0+b1*a0+bcc+t2*ss_m))*(a1-a0) +
     (1/2)*t3*t3*ss_m*(a1*a1-a0*a0)) - (1.0)
   -exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar-(t2+t3*a0)*(b0+b1*a0+bcc)-
     (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m)
   +exp(t2*mstar+t3*a0*mstar-(t2+t3*a0)*(b0+b1*a0+bcc)-
     (1/2)*(t2+t3*a0)*(t2+t3*a0)*ss_m);
INTmed_comp = exp( (t1+t2*b1+t3*(b0+b1*a0+b1*a1+bcc+t2*ss_m))*(a1-a0)
    + (1/2)*t3*t3*ss_m*(a1*a1-a0*a0) )
   -exp( (t2*b1+t3*b1*a0)*(a1-a0) )
    -exp( (t1+t3*(b0+b1*a0+bcc+t2*ss_m))*(a1-a0)
    + (1/2)*t3*t3*ss_m*(a1*a1-a0*a0) ) + (1);
PIE_comp = exp( (t2*b1+t3*b1*a0)*(a1-a0) ) - (1);
terr=cde_comp+intref_comp+intmed_comp+pie_comp;
total = exp((t1 + t3*(b0+b1*a0+bcc + t2*ss_m))
    *(a1-a0)+(1/2)*t3*t3*ss_m*(a1*a1-a0*a0))
    *exp((t2*b1+t3*b1*a1)*(a1-a0));
estimate 'Total Effect Risk Ratio' total;
estimate 'Total Excess Relative Risk' total-1;
estimate 'Excess Relative Risk due to CDE' cde_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTref'
    intref_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTmed'
    intmed_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to PIE' pie_comp*(total-1)/terr;
estimate 'Proportion CDE' cde_comp/terr;
estimate 'Proportion INTref' intref_comp/terr;
estimate 'Proportion INTmed' intmed_comp/terr;
estimate 'Proportion PIE' pie_comp/terr;
estimate 'Overall Proportion Mediated' (pie_comp+intmed_comp)/terr;
estimate 'Overall Proportion Attributable to Interaction'
   (intref_comp+intmed_comp)/terr;
estimate 'Overall Proportion Eliminated'
   (intref_comp+intmed_comp+pie_comp)/terr;
run;
```

The code given above is applicable to cohort data. For case–control studies in which sampling is done on the outcome Y, if the outcome is rare, then the code above can be adapted by fitting the mediator regression only among

the controls. This can be done by replacing the sixth line of code by: ll_m= ((-((m-mu_m)**2)/(2*ss_m)-0.5*log(ss_m)))*(1-y);

14.9.4. Binary Outcome, Binary Mediator

To estimate the components of the four-way decomposition for the effect of exposure A on a binary outcome Y with binary mediator M under the regression models

$$\text{logit}\{P(Y=1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M=1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

one can use the code below. The analytic expressions for each of the four components and derivations are given in the Appendix. The explanation of the code follows the one presented in Section 14.9.3.

```
proc nlmixed data=mydata;
parms t0=1 t1=0 t2=0 t3=0 tc1=0 tc2=0 tc3=0 b0=0 b1=0 bc1=0 bc2=0
bc3=0; a1=1; a0=0; mstar=0; cc1=58.57; cc2=1.44; cc3=0.34;
p_y=(1+exp(-(t0 + t1*A + t2*M + t3*A*M + tc1*C1 + tc2*C2
   + tc3*C3)))**-1;
p_m =(1+exp(-(b0 + b1*A + bc1*C1 + bc2*C2 + bc3*C3)))**-1;
ll_y= y*log (p_y)+(1-y)*log(1-p_y);
ll_m= m*log (p_m)+(1-m)*log(1-p_m);
ll_o= ll_m + ll_y;
model Y ~general(ll_o);
bcc = bc1*cc1 + bc2*cc2 + bc3*cc3;
CDE_comp = exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar)*(1+exp(b0+b1*a0+bcc))/
     (1+exp(b0+b1*a0+bcc+t2+t3*a0))
   - exp(t2*mstar+t3*a0*mstar)*(1+exp(b0+b1*a0+bcc))/
   (1+exp(b0+b1*a0+bcc+t2+t3*a0));
INTref_comp = exp(t1*(a1-a0))*(1+exp(b0+b1*a0+bcc+t2+t3*a1))/
     (1+exp(b0+b1*a0+bcc+t2+t3*a0)) - (1)
   -exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar)*(1+exp(b0+b1*a0+bcc))
     *exp((t1+t3*mstar) *(a1-a0))
    /(1+exp(b0+b1*a0+bcc+t2+t3*a0))
   + exp(t2*mstar+t3*a0*mstar)*(1+exp(b0+b1*a0+bcc))/
   (1+exp(b0+b1*a0+bcc+t2+t3*a0));
INTmed_comp = exp(t1*(a1-a0))*(1+exp(b0+b1*a1+bcc+t2+t3*a1))
    *(1+exp(b0+b1*a0+bcc))
    /( (1+exp(b0+b1*a0+bcc+t2+t3*a0))*(1+exp(b0+b1*a1+bcc))  )
   - (1+exp(b0+b1*a1+bcc+t2+t3*a0))*(1+exp(b0+b1*a0+bcc)) /
     ( (1+exp(b0+b1*a0+bcc+t2+t3*a0))
    *(1+exp(b0+b1*a1+bcc))  )
   - exp(t1*(a1-a0))*(1+exp(b0+b1*a0+bcc+t2+t3*a1))/
   (1+exp(b0+b1*a0+bcc+t2+t3*a0))
    + (1);
PIE_comp = (1+exp(b0+b1*a0+bcc))*(1+exp(b0+b1*a1+bcc+t2+t3*a0)) /
     ( (1 + exp(b0+b1*a1+bcc))
```

```
    *(1+exp(b0+b1*a0+bcc+t2+t3*a0)) ) -(1);
terr=cde_comp+intref_comp+intmed_comp+pie_comp;
total = exp(t1*a1)*(1+exp(b0+b1*a0+bcc))*(1+exp(b0+b1*a1+bcc+t2+t3*a1))
    / ( exp(t1*a0)*(1 + exp(b0+b1*a1+bcc))
    *(1+exp(b0+b1*a0+bcc+t2+t3*a0)) );
estimate 'Total Effect Risk Ratio' total;
estimate 'Total Excess Relative Risk' total-1;
estimate 'Excess Relative Risk due to CDE' cde_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTref'
    intref_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to INTmed'
    intmed_comp*(total-1)/terr;
estimate 'Excess Relative Risk due to PIE' pie_comp*(total-1)/terr;
estimate 'Proportion CDE' cde_comp/terr;
estimate 'Proportion INTref' intref_comp/terr;
estimate 'Proportion INTmed' intmed_comp/terr;
estimate 'Proportion PIE' pie_comp/terr;
estimate 'Overall Proportion Mediated' (pie_comp+intmed_comp)/terr;
estimate 'Overall Proportion Attributable to Interaction'
  (intref_comp+intmed_comp)/terr;
estimate 'Overall Proportion Eliminated'
  (intref_comp+intmed_comp+pie_comp)/terr;
run;
```

The code given above is applicable to cohort data. For case–control studies in which sampling is done on the outcome Y, if the outcome is rare, then the code above can be adapted by fitting the mediator regression only among the controls. This can be done by replacing the sixth line of code by: ll_m= ((m*log (p_m)+(1-m)*log(1-p_m)))*(1-y);

14.10. DISCUSSION

The methods here allow an investigator to assess the extent to which the effect of an exposure on an outcome is completely independent of a mediator (the controlled direct effect with the mediator set to zero), as well as the extent to which the effect is due to interaction but not mediation, to which it is due to mediation but not interaction, and to which it is due to both mediation and interaction together. The four-way decomposition here encompasses and unites previous decompositions in the literature, both concerning mediation and concerning interaction, considered earlier in the book. The results here have also provided a mechanistic interpretation to the difference between a total effect and a controlled direct effect; this contrast has been used to assess policy implications, and it is more easily identified than many other causal quantities concerning mediation; the results here show that it also has a mechanistic interpretation as well. We have also shown how the four-way decomposition in this chapter can be carried out on a difference scale and on a ratio scale, and we have related the various components to standard statistical models. We have seen that in addition to reporting the four components, an investigator

can also easily report, along with these, the overall proportion attributable to interaction, the overall proportion mediated, and the proportion of the effect that would be eliminated if the mediator were removed. As seen in the empirical example in genetic epidemiology, the approach described here can shed considerable insight into the relationships between an exposure and a mediator with an outcome, as well as into the role of both mediation and interaction in these relationships.

The SAS code given here (cf. VanderWeele, 2014) likewise provides practical and relatively easy-to-use software tools to implement the approaches here in a wide range of settings. The central limitation of the approach considered in this chapter is the strong assumptions being made about confounding; these are, however, the same assumptions as in Chapter 2 and in the literature on mediation that only focuses on simpler decompositions. Future research could examine the robustness of each of the four components to confounding and measurement error. For example, in Chapter 11 we saw that interaction terms may be more robust to confounding (VanderWeele et al., 2012b), but recent work also indicates that interaction terms when the two exposures are correlated may be particularly sensitive to measurement error (Valeri et al., 2014; Valeri and VanderWeele, 2014); different components may be robust to different forms of bias. Future work could also extend existing sensitivity analysis techniques (Imai et al., 2010a; VanderWeele, 2010a; Valeri et al., 2014; Valeri and VanderWeele, 2014) for direct and indirect effects to each of the four components.

Prior work on mediation within the counterfactual framework has accommodated potential interaction. The approach here makes the role of interaction, along with its separate contribution beyond mediation, clearer and unites, within a single framework, the phenomena of mediation and interaction.

15

Social Interactions and Spillover Effects

Spillover effects refer to the phenomenon whereby the exposure of one individual can affect the outcome of another. In many settings, this sort of phenomenon does not arise: Whether one cancer patient receives surgery or chemotherapy is not likely to affect the survival outcome of a different cancer patient. However, in other contexts, spillover is quite clear. Whether a person is vaccinated might well affect whether a family member is infected. This phenomenon of spillover, also sometimes referred to as "interference," is common whenever an outcome depends upon social interactions between individuals. Causal inference in the presence of such spillover effects or social interactions is considerably more complex. For some time it was thought that formal causal inference using the counterfactual approach was intractable when such interference or spillover was involved. More recently, however, some progress has been made. In many settings, inference is indeed rendered considerably more difficult. However, in some settings, such as when there are only two individuals per household under study and interference can occur only within household, causal inference in the presence of spillover can be carried out in a straightforward manner. Indeed we will see in this setting that there are numerous connections between spillover effects, on the one hand, and many of the concepts we have already described in this book, including principal strata effects, mechanistic interaction, and even direct and indirect effects, on the other hand. As we go through this chapter, we will see that many of the methods and approaches for mediation and interaction have direct analogues in the spillover effect context (see especially Sections 15.3–15.5). We will first begin with the setting when exposures are randomized and there are only two individuals per household. We will then discuss some of the inferential challenges that arise when there are more than two individuals per household and then later discuss further issues that arise in the analysis of spillover effects and social interactions when data come from an observational study. Finally, we will provide a brief introduction to causal inference with social network data wherein an entire network of individuals may be related to and influence each other.

15.1. NOTATION AND DEFINITIONS FOR SPILLOVER EFFECTS

Suppose that in a particular study there are K households indexed by $i = 1,\ldots,K$ in which there are two people under study per household (e.g., husband and wife) indexed by $j = 1,2$. Initially we will assume that the two persons are distinguishable from one another (e.g., $j = 1$ denotes the wife and $j = 2$ denotes the husband). We let A_{ij} denote the exposure status for person j in household i. For example, for a vaccine we let $A_{ij} = 1$ denote that the person j in household i received the vaccination and let $A_{ij} = 0$ denote that the person did not. We let Y_{ij} denote the infection status of person j in household i after some suitable follow-up. We will consider the vaccine exposure and infection outcome as a running example throughout the next few sections, but none of the approaches to spillover effects described here depends on this context. The exposure could instead be a tutoring program that one individual participates in with educational outcomes assessed, or a weight loss program for individuals with obesity status assessed at the end of the study. In each of these settings it is likewise the case that the tutoring program for one individual might affect the educational outcomes for another individual in the same household, or a weight loss program for one individual might affect the obesity status for another individual in the household. Persons within households may share information with one another, and the behavior of one person may affect that of the other.

As before, we will employ counterfactual notation to define our effects of interest. Here, because of the complexities that spillover introduces, we will be using counterfactual notation throughout our discussion of the various approaches to spillover effects. Because the number of counterfactuals increases dramatically once we allow spillover and because we now also have to index the different households in the study and not just individuals, we will be changing notation a little to accommodate these additional complexities. We will let $Y_{ij}(a_{i1},a_{i2})$ denote the counterfactual outcome for person j in household i if the two people in that household i had (possibly contrary to fact) been given vaccine status of (a_{i1},a_{i2}). For example, $Y_{i2}(1,0)$ would denote what would have happened to person 2 if person 1 had received the vaccine and person 2 had not; $Y_{i1}(0,0)$ denotes what would have happened to person 1 if neither had received the vaccine, and so on. Note that under this counterfactual or "potential outcomes" notation, the potential outcome for person 1, $Y_{i1}(a_{i1},a_{i2})$, depends on the vaccine status of both persons. This allows for the possibility that the exposure status of one person affects the outcome of another, sometimes referred to as interference or a spillover effect. Most literature in causal inference makes a "no-interference" assumption (Cox, 1958) that one person's outcome does not depend on the exposure of other people. In the current context, this would imply that $Y_{i1}(a_{i1},a_{i2}) = Y_{i1}(a_{i1})$ and $Y_{i2}(a_{i1},a_{i2}) = Y_{i2}(a_{i2})$ so that each person's outcome depends only on his or her own exposure status. We will allow for such interference here, but will assume that the vaccine status of people in one household do not affect the outcomes of those in other households. This assumption is sometimes referred to as "partial interference" (Sobel, 2006; Hudgens and Halloran, 2008). This would not be an unreasonable assumption if the source population in the study were very large and

a relatively small number of households were randomly selected for inclusion in the study.

For those familiar with the more formal causal inference literature, the "no interference" assumption (Cox, 1958) was part of what Rubin in his initial work on causal inference (Rubin, 1974, 1978, 1980) called the "Stable Unit Treatment Value Assumption" or SUTVA. SUTVA was what was required for counterfactual outcomes of the form $Y_{i1}(a_{i1})$ to be well-defined. The two components of Rubin's SUTVA were (i) no interference between units so that the counterfactual outcome of one person did not depend on the exposures of another and (ii) no multiple versions of treatment, once again so that $Y_{i1}(a_{i1})$ would be well-defined; if the counterfactual outcome that resulted from assigning a particular exposure to an individual depended upon the version, then the potential outcome $Y_{i1}(a_{i1})$ would be ambiguous until the version was specified. Under SUTVA (no interference, no multiple versions of treatment) the counterfactual outcomes $Y_{i1}(a_{i1})$ would be well-defined. In Chapter 7, we touched briefly upon the issue of violations of the no multiple versions of treatment assumption. In this chapter we will discuss violations of the no-interference assumption. Rubin (1980, 1986, 1990) noted that the potential outcomes or counterfactual notation could be extended to allow for interference, but then this made analysis quite difficult. There was subsequently some informal analysis of interference and spillover effects in infectious disease epidemiology that made some reference to counterfactual-based causal inference (Halloran and Struchiner, 1991, 1995) and also in economics (Manski, 1993, 2000), but it is only in the last few years that the analysis of spillover effects has been formulated more rigorously in counterfactual-based notation and that considerable progress has been made (Sobel, 2006; Hong and Raudenbush, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008; Graham, 2008; VanderWeele and Tchetgen Tchetgen, 2011a,b; VanderWeele et al., 2012d,e; Tchetgen Tchetgen and VanderWeele, 2012; Manski, 2012; Liu and Hudgens, 2013; Graham et al., 2014). As we will see, interference is not simply a problem that gives rise to complexities in the analysis, but in fact is a phenomenon that is often of substantive interest itself and gives rise to new causal quantities and questions that can, with the methods that have been and are being developed, be addressed.

15.2. BASIC SPILLOVER AND INDIVIDUAL/DIRECT EFFECTS

If we allow for interference between individuals within a household, we can define various causal effects of interest beyond the overall effect of vaccinating versus not vaccinating entire households. For example, for husband–wife pairs, if we let person 1 denote the wife and person 2 be the husband, then we might consider an individual effect or "direct effect" (Halloran and Struchiner, 1995; Hudgens and Halloran, 2008) of having one person (e.g., the wife) vaccinated while holding the vaccine status of the other person (e.g., the husband) constant. In counterfactual notation, this would be $Y_{i1}(1, a_{i2}) - Y_{i1}(0, a_{i2})$; if we hold the husband's vaccine status to vaccinated, this is $Y_{i1}(1, 1) - Y_{i1}(0, 1)$; if we hold the husband's vaccine

status to unvaccinated, this is $Y_{i1}(1,0) - Y_{i1}(0,0)$. Alternatively, we could consider indirect or spillover effects such as the effect on the wife's outcome of having the husband vaccinated versus unvaccinated while holding the wife's vaccine status constant. In counterfactual notation, this would be $Y_{i1}(a_{i1},1) - Y_{i1}(a_{i1},0)$; if we hold the wife's vaccine status to vaccinated, this is $Y_{i1}(1,1) - Y_{i1}(1,0)$; if we hold the wife's vaccine status to unvaccinated, this is $Y_{i1}(0,1) - Y_{i1}(0,0)$. We could also define analogous effects for the husband. If any of the various spillover/indirect effects in this setting are nonzero, then we would say that interference is present. For example, if we found on average that $\mathbb{E}[Y_{i1}(0,1) - Y_{i1}(0,0)] > 0$, we would know that there was a spillover effect of the husband's vaccine on the wife's outcome.

In general we cannot hope to estimate these individual/direct and spillover/indirect effects for a particular household, but we may be able to estimate these effects on average, at least in various randomized trials in which both the husband's and the wife's exposures are randomized. We could do so by simply comparing the sample average of the wife's or husband's outcome across the various subgroups defined by the wife's and husband's vaccination status. We will discuss settings in which clusters have more than two people and with other randomization schemes in subsequent sections. When both wife's and husband's vaccine status are randomized, we can identify average individual/direct effect for individual 1 while individual 2's vaccine fixed at a_{i2} by

$$\begin{aligned}\mathbb{E}[Y_{i1}(1,a_{i2}) - Y_{i1}(0,a_{i2})] = \mathbb{E}[Y_{i1}|A_{i1}=1,A_{i2}=a_{i2}] \\ - \mathbb{E}[Y_{i1}|A_{i1}=0,A_{i2}=a_{i2}]\end{aligned}$$

The left-hand side is the causal individual/direct effect we want to identify: the effect of individual 1's receiving the exposure when individual 2's exposure is fixed at a_{i2}. This individual/direct effect may be different for different values of a_{i2}—that is, different values at which individual 2's exposure may be fixed. The right-hand side is simply an expression involving the average observed outcomes in different strata of the exposures of individuals 1 and 2. If the exposures are randomized for each person, then the causal estimand on the left-hand side will on average equal the empirical expression on the right-hand side. We could similarly identify the individual/direct effect of individual 2's exposure on individual 2's outcome with individual 1's exposure fixed at some level a_{i1} by $\mathbb{E}[Y_{i2}(a_{i1},1) - Y_{i2}(a_{i1},0)] = \mathbb{E}[Y_{i2}|A_{i1}=a_{i1},A_{i2}=1] - \mathbb{E}[Y_{i2}|A_{i1}=a_{i1},A_{i2}=0]$. We could also potentially average over these individual/direct effects for individuals 1 and 2 (Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012), an issue we will return to below.

Likewise under such randomization we can identify the spillover/indirect effect of individual 2's exposure on individual 1's outcome with individual 1's exposure fixed at a_{i1} by

$$\begin{aligned}\mathbb{E}[Y_{i1}(a_{i1},1) - Y_{i1}(a_{i1},0)] = \mathbb{E}[Y_{i1}|A_{i1}=a_{i1},A_{i2}=1] \\ - \mathbb{E}[Y_{i1}|A_{i1}=a_{i1},A_{i2}=0]\end{aligned}$$

Once again this spillover effect of individual 2's exposure on individual 1 could vary, depending on the level a_{i1} to which individual 1's exposure is fixed. And similarly the spillover/indirect effect of individual 1's exposure on individual 2's outcome with individual 2's exposure fixed at a_{i2} is identified by $\mathbb{E}[Y_{i2}(1,a_{i2}) - Y_{i2}(0,a_{i2})] = \mathbb{E}[Y_{i2}|A_{i1} = 1, A_{i2} = a_{i2}] - \mathbb{E}[Y_{i2}|A_{i1} = 0, A_{i2} = a_{i2}]$. And we could once again also potentially average over these individual/direct effects for individuals 1 and 2 (Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012).

We could also consider the effect on either individual 1 or 2 of giving both versus neither the exposure. We could identify such effects from the data by, for example,

$$\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(0,0)] = \mathbb{E}[Y_{i1}|A_{i1} = 1, A_{i2} = 1] - \mathbb{E}[Y_{i1}|A_{i1} = 0, A_{i2} = 0]$$

for individual 1 or $\mathbb{E}[Y_{i2}(1,1) - Y_{i2}(0,0)] = \mathbb{E}[Y_{i2}|A_{i1} = 1, A_{i2} = 1] - \mathbb{E}[Y_{i2}|A_{i1} = 0, A_{i2} = 0]$ for individual 2, or we could average over both individuals. We also have that this "total effect" is equal to the sum of the spillover effect and the direct/individual effect. For example, $\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(0,0)] = \mathbb{E}[Y_{i1}(1,1) - Y_{i1}(1,0)] + \mathbb{E}[Y_{i1}(1,0) - Y_{i1}(0,0)]$.

As is perhaps already clear, the terminology can easily become confused with the terminology that we have used with the mediation literature. What we have called here a "spillover effect" such as the effect of individual 2's exposure on individual 1's outcome, $\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(1,0)]$, is sometimes referred to in the infectious disease literature as an "indirect effect" (e.g., Halloran and Struchiner, 1995; Hudgens and Halloran, 2008). It is "indirect" in the sense that it is individual 2's exposure that is having an effect on individual 1's outcome. However, this is very different from the indirect effects we were considering in Part I of this book in the context of mediation in which an exposure affected an outcome through a mediator. Likewise, in the infectious disease literature the effect of individual 1's exposure on individual 1's outcome, an effect such as $\mathbb{E}[Y_{i1}(1,0) - Y_{i1}(0,0)]$, is often referred to as a "direct effect"—it is the effect of individual 1's exposure directly on individual 1's outcome. But this is again very different from the "direct effect" in the context of mediation in which there was an effect of an exposure on the outcome, not through a particular intermediate on the pathway from exposure to the outcome.[1] Because

1. In the infectious disease literature, the terminology of "direct and indirect effects" when interference is present dates at least as far back as Halloran and Struchiner (1991), although Hudgens and Halloran (2008) arguably provide the first formal counterfactual definitions in the infectious disease field. The terminology of "direct and indirect effects" in the context of mediation analysis extends at least as far back as the literature on structural equation modeling (e.g., Duncan, 1966) motivated by the method of path coefficients of Wright (1921); counterfactual notions of direct and indirect effects were described in detail by Holland (1988) and Robins and Greenland (1992). Because of the potential ambiguity in terms "direct effect" and "indirect effect," Sobel (2006) chose to use the term "spillover effect" for the effect on an individual's outcome of holding the individual's own treatment fixed but modifying the treatments received by other individuals. An early paper (Strain et al., 1976) in experimental educational psychology appears to have interchangeably used "indirect effect" and "spillover effect" to denote the effect on a child's outcome of holding the child's own treatment fixed but modifying the treatments received by other children. Complicating terminological issues yet further, the causal inference literature on mediation has itself produced alternative

of the multiple varieties of direct and indirect effects, the use of more specific terminology may be desirable. In the context of interference, "indirect effect" and "direct effect" could be replaced by "spillover effect" and "individual effect"; in the context of mediation, "indirect effect" and "direct effect" could be replaced by "mediated effect" and "unmediated effect." As will be seen below, in some contexts, both interference and mediation may be present and of interest and the terms "direct effect" and "indirect effect" become even ambiguous as they may make reference to the concepts from interference or from mediation. We will return again to these issues of terminology in Sections 15.3 and 15.4.

Nonetheless, issues of terminology aside, we can, at least in the simple setting of only two individuals per household, make progress in defining and identifying these spillover and individual/direct effects. In fact, in this simple setting, statistical inference is also quite straightforward. Because each of these effects is simply a contrast of average outcomes in different strata defined by the exposure status of the two individuals, simple t-test statistics for differences in means can be used to test for the presence of these effects and to construct confidence intervals for these spillover and individual/direct effects. As we shall see in a later section, once we have more than two individuals per cluster, statistical inference and the construction of confidence intervals can become particularly challenging in the presence of interference.

15.3. ASSESSING ''INFECTIOUSNESS'' EFFECTS

In the previous section we saw that under randomization of the exposures, it was possible to examine the spillover effect of one individual's exposure on another's outcome. For example, the vaccination of one person might prevent infection in the second person in the same household. A further distinction can be drawn between the ways such a protective effect might arise. First, vaccinating the first person may protect the second person by preventing the first from being infected and passing the infection on to the second. Alternatively, vaccinating the first person may protect the second by rendering the infection less transmissible even if the first is infected. This latter mechanism is sometimes referred to as an "infectiousness effect" of the vaccine (Datta et al., 1999; Préziosi and Halloran, 2003). In this section we will consider methods to assess such an infectiousness effect. As before, however, nothing in this setting restricts the applicability of these methods to an infectious disease context. The methods described below could also be employed in a study in which the exposure were, say, a smoking cessation program in which one of two persons in a household participated. The participation of the first person might affect the smoking behavior of the second. This might occur either (i) because smoking cessation for the first person encourages the second to stop

definitions of direct and indirect effects based on potential interventions on the mediator (Robins and Greenland, 1992; Pearl, 2001) or alternatively on the notion of principal strata (Frangakis and Rubin, 2002; Rubin, 2004).

smoking or (ii) because even if the first person does not stop smoking, the second person might nevertheless be exposed to some of the smoking cessation program materials. One could potentially assess the presence of this second type of effect by applying the methods described below concerning the "infectiousness effect." The methods could likewise be applicable to a range of health-related, social, and psychological outcomes and exposures.

15.3.1. Defining Infectiousness Effects

Here we will assume a simple randomized experiment in which one of the two persons is randomized to receive a vaccine or control and the second person is always unvaccinated/unexposed. We will let $j = 1$ denote the individual who may or may not be vaccinated and let $j = 2$ denote the individual who is always unvaccinated. The choice of which person is individual 1 and which is individual 2 could be either determined randomly or fixed in advance (e.g., in households of married couples, the husband could be the person who is never vaccinated). Because individual 2 is always unvaccinated, we are implicitly conditioning on $A_{i2} = 0$ throughout. At the end of this section, we will briefly consider other settings in which, in some households, both people are vaccinated.

Let us now consider how one might define the "infectiousness effect." Suppose we are in the setting of a vaccine trial in which in each household i individual 1 is randomized to vaccine or control and individual 2 always receives control. Thus, in half the households, one of the two people will be vaccinated, and in half of the households neither person will be vaccinated. The crude estimator for the infectiousness effect (on the risk difference scale) might then be taken as

$$\mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] - \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1] \tag{15.1}$$

This is a comparison of the infection rates for individual 2 in the subgroup in which individual 1 was vaccinated and infected versus in the subgroup in which individual 1 was unvaccinated and infected. Although this is an appealing intuitive measure for trying to capture the extent to which the vaccine may render the infection less transmissible, which may in turn prevent the second individual from being infected (i.e., the "infectiousness effect"), the measure is subject to selection bias. Although the vaccine status for individual 1, A_{i1}, is randomized, conditioning on a variable that occurs after treatment (namely, the infection status of individual 1) in effect breaks randomization. The subgroup with individual 1 vaccinated and infected may be quite different from that in which individual 1 is unvaccinated and infected. For example, those in the vaccinated group who became infected may be overall a less healthy subpopulation than the unvaccinated group who became infected. If the persons who got the disease even though they were vaccinated are less healthy, they may also be more likely to be infectious and to pass on the disease. The comparison between the two infected subgroups is not fair because, by conditioning on a variable that occurs after randomization—namely infection status of the first person—we induce selection bias. We are computing infection rates for individual 2 for subpopulations

that are quite different with respect to individual 1. Let us instead consider a second contrast (VanderWeele and Tchetgen Tchetgen, 2011a; Halloran and Hudgens, 2012a,b):

$$\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0) | Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \tag{15.2}$$

This contrast compares the infection status for individual 2 if individual 1 was vaccinated, $Y_{i2}(1,0)$, versus unvaccinated, $Y_{i2}(0,0)$, but only among the subset of households for whom individual 1 would have been infected irrespective of whether or not individual 1 was vaccinated, that is, $Y_{i1}(1,0) = Y_{i1}(0,0) = 1$. As discussed in Chapter 8, such a subgroup is sometimes referred to as a principal stratum (Frangakis and Rubin, 2002). Because we are considering only the subset of households for whom individual 1 would have been infected irrespective of whether or not individual 1 was vaccinated, individual 2 will always be exposed to the infection of individual 1 and thus any effect of the vaccine ought to occur through changing the infectiousness. We might therefore take the contrast in (15.2) as a formal causal contrast more closely corresponding to the "infectiousness effect." Moreover, unlike with the crude comparison in (15.1), we are now comparing, in contrast (15.2), the infection rates for individual 2 for the same subpopulation, namely the subpopulation for which individual 1 would have been infected irrespective of whether individual 1 was vaccinated. We are no longer considering a more healthy or less healthy subgroup for individual 1. Note that our framework and definitions do not assume that the second person cannot also be exposed outside the household.

Unfortunately, however, with regard to this causal infectiousness effect, we do not know which households fall into this subpopulation in which individual 1 would have been infected irrespective of whether individual 1 was vaccinated. This is because in each household we can only observe the outcome of individual 1 either with the vaccine or without—but not under both scenarios. Because we do not know which households fall into this subpopulation, we cannot compute the contrast in (15.2) in any straightforward manner from the data. The contrast is, in general, unidentified. In the next section, however, we will show that, under some fairly reasonable assumptions, the crude estimator in (15.1) is in fact conservative for the causal "infectiousness effect" in contrast (15.2).

15.3.2. Bounding the Infectiousness Effect

To show that the crude contrast in (15.1), which can be estimated from data in a trial, is conservative for the causal "infectiousness effect" contrast in (15.2), we will need two assumptions. The first assumption states that the vaccine will never be the cause of the infection; that is, there may be persons who would be infected irrespective of vaccination status or who would not be infected irrespective of vaccination status or who would be infected if unvaccinated and not infected if vaccinated, but there is no one who would be infected if vaccinated and uninfected if unvaccinated. More formally, we will assume that (A15.1) for all i, $Y_{i1}(1,0) \leq Y_{i1}(0,0)$. Again,

assumption (A15.1) states the vaccine will never be the cause of the infection; that is, there is no one who would be infected if vaccinated but uninfected if unvaccinated. Assumption (A15.1) is sometimes referred to as a monotonicity assumption.

To show that the crude estimator in (15.1) is conservative for the causal effect in (15.2), we will impose one further assumption. We will state the assumption formally and then provide some explanation and intuition. Our second assumption is that (A15.2) $\mathbb{E}[Y_{i2}(0,0)|A_{i1}=0,Y_{i1}=1] \leq \mathbb{E}[Y_{i2}(0,0)|A_{i1}=1,Y_{i1}=1]$. Assumption (A15.2) states that the average infection rate for individual 2 if both individuals 1 and 2 would have been unvaccinated would be lower in the subgroup of households for which individual 1 is actually unvaccinated and infected than in the subgroup of households for which individual 1 is actually vaccinated and infected. Note that assumption (A15.2) compares two subgroups: (i) the subgroup of households for which individual 1 was actually unvaccinated and infected and (ii) the subgroup of households for which individual 1 was actually vaccinated and infected. It then states that if instead we had, contrary to fact, vaccinated no one, then infection rates for individual 2 would be at least as high in the second subgroup as in the first. The assumption is arguably reasonable insofar as the subgroup for which individual 1 was vaccinated and infected is likely less healthy (or the infection more virulent) than the subgroup for which individual 1 was unvaccinated and infected; thus, under the scenario in which both people are unvaccinated, individual 2 is more likely to be infected in the second subgroup than in the first. If this is indeed the case, then Assumption (A15.2) will hold.

If assumptions (A15.1) and (A15.2) hold, then the crude estimator is conservative for the causal infectiousness effect in that

$$\begin{aligned}&\mathbb{E}[Y_{i2}(1,0)-Y_{i2}(0,0)|Y_{i1}(1,0)=Y_{i1}(0,0)=1]\\&\quad\leq \mathbb{E}[Y_{i2}|A_{i1}=1,Y_{i1}=1]-\mathbb{E}[Y_{i2}|A_{i1}=0,Y_{i1}=1]\end{aligned}$$

If in the crude comparison in (15.1) we find a negative (i.e., protective) effect, then the true causal infectiousness effect is even larger in magnitude. If an investigator uses the crude contrast and finds the vaccine of the first person protective for the second person in the subset for whom the first person is infected, then this gives evidence for a true causal infectiousness effective. If the crude estimate is not found protective, this may indicate the absence of a true causal infectiousness effect, or it may be the case that there is a true causal infectiousness effect but the crude estimator, being conservative, is unable to detect it.

Of course, the result that our crude estimator is conservative depends upon the assumptions being made. These may not hold in all contexts. In such cases we can use a sensitivity analysis technique. Let $\theta = \mathbb{E}[Y_{i2}(0,0)|A_{i1}=1,Y_{i1}=1]-\mathbb{E}[Y_{i2}(0,0)|A_{i1}=0,Y_{i1}=1]$ denote the sensitivity parameter that contrasts the average counterfactual infection rates for individual 2 if both individuals 1 and 2 were, possibly contrary to fact, unvaccinated in the subgroup of households for which individual 1 is actually vaccinated and infected versus the subgroup of households for which individual 1 is actually unvaccinated and infected. Assumption (A15.2) is then simply that $\theta \geq 0$. This assumption could then be relaxed by

specifying possibly negative values of the sensitivity parameter. Under the monotonicity assumption (A15.1) alone, we have that (VanderWeele and Tchetgen Tchetgen, 2011a)

$$\begin{aligned}&\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0) | Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ &\quad = \mathbb{E}[Y_{i2} | A_{i1} = 1, Y_{i1} = 1] - \mathbb{E}[Y_{i2} | A_{i1} = 0, Y_{i1} = 1] - \theta\end{aligned}$$

In other words, to obtain the infectiousness effect under monotonicity, we can calculate the crude infectiousness effect in (15.1), specify the sensitivity parameter θ, and subtract the sensitivity parameter θ from the crude estimate to obtain the infectiousness effect. We can vary θ over a range of plausible values in a sensitivity analysis to produce a range of plausible values for the infectiousness effect. Because of the simple relationship above, a corrected confidence interval under sensitivity parameter θ can be obtained simply by subtracting θ from both limits of the confidence interval for the crude estimate in (15.1). Halloran and Hudgens (2012a,b) also derive upper and lower bounds for causal effects on infectiousness under just assumption (A15.1) alone without specifying any sensitivity analysis parameter. See also Hudgens and Halloran (2006, 2012a) and VanderWeele et al. (2014b) for alternative sensitivity analysis techniques.

15.3.3. Other Measures of Effect

We defined the causal infectiousness effect above on a risk difference scale. Other measures of effect might also be of interest. For notational convenience in this section we define the following:

$$\begin{aligned}p_v &= \mathbb{E}[Y_{i2}(1,0) | Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ p_u &= \mathbb{E}[Y_{i2}(0,0) | Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ p_1 &= \mathbb{E}[Y_{i2} | A_{i1} = 1, Y_{i1} = 1] \\ p_0 &= \mathbb{E}[Y_{i2} | A_{i1} = 0, Y_{i1} = 1]\end{aligned}$$

The causal infectiousness effect on the risk difference scale is then just $p_v - p_u$, and then the result above on the naive estimator being conservative under assumptions (A15.1) and (A15.2) can be simply stated as $p_v - p_u \leq p_1 - p_0$. We might similarly be interested in the causal infectiousness effect on the risk ratio scale, p_v/p_u, or on the odds ratio scale, $p_v(1-p_u)/\{p_u(1-p_v)\}$. We might further be interested in what one might refer to as the causal vaccine efficacy infectiousness effect, $1 - p_v/p_u$. Under assumptions (A15.1) and (A15.2), for each of these additional measures of effect, the crude estimator is conservative for the true causal infectiousness effect (VanderWeele and Tchetgen Tchetgen, 2011a); that is, we have

$$\begin{gathered}p_v/p_u \leq p_1/p_0 \\ p_v(1-p_u)/\{p_u(1-p_v)\} \leq p_1(1-p_0)/\{p_0(1-p_1)\} \\ \text{and } 1 - p_v/p_u \geq 1 - p_1/p_0\end{gathered}$$

Once again, if an investigator found, using the crude estimator, a protective effect on the risk ratio, odds ratio, or vaccine efficacy scales, this would be conservative for the true causal effect; the true causal effective would be even more protective than indicated by the crude estimator.

15.3.4. Illustration

Millar et al. (2008) analyzed results from a group randomized trial in which 7-valent pneumococcal conjugate vaccine (PCV7) was compared with meningococcal conjugate vaccine against serogroup C (MCC) among southwestern American Indian communities. In each household, a child in the household was vaccinated with either the PCV7 vaccine or the MCC vaccine. The primary purpose of the study was to examine whether pneumococcal colonization rates were lower for unvaccinated adults and children in households in which the vaccinated child was vaccinated with PCV7 versus MCC. The study found a protective effect, with odds ratios of 0.57 (95% CI: 0.33–0.99) for unvaccinated adults and 0.57 (95% CI: 0.26–0.98) for unvaccinated children, respectively.

In addition, the investigators conducted a secondary analysis in which the outcome Y was vaccine type (VT) pneumococcal colonization. They compared the odds of VT colonization for unvaccinated adults for PCV7 versus MCC among households in which the vaccinated child was colonized with a VT strain and obtained an odds ratio of 0.34 (95% CI: 0.11–0.99). This estimate is likely biased for the causal infectiousness effect due to selection and stratifying on a post-randomization variable, colonization status of the vaccinated child. However, under the assumption that the PCV7 vaccine never causes VT pneumococcal colonization [assumption (A15.1)] and that VT colonization rates for unvaccinated individuals in households in which the child received the PCV7 vaccine and was colonized would be higher than those in households in which the child received the MCC vaccine and was colonized if for both subgroups the child had been given the MCC vaccine (assumption (A15.2), i.e., $\mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1] \leq \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1} = 1]$), then we could conclude that the odds ratio of 0.34 (95% CI: 0.11–0.99) was in fact conservative for the causal "infectiousness" effect odds ratio. We would have evidence for a true infectiousness effect. Note that here the 7-valent vaccine and the colonization status are for the collection of the seven pneumococcal vaccine types; the assumptions would have to hold for colonization rates of the seven types taken as a collection.

We note also that the actual design of the Millar et al. (2008) study was a group randomized trial with two to four American Indian chapters constituting a randomization unit. Because of this, the "partial interference" assumption that the vaccination status of one household in the study does not affect members of other households will likely be partially violated. The example here is given for illustrative purposes.

15.3.5. Further Extensions

Thus far, for simplicity, we have considered a randomized design in which one of two persons is randomized to receive a vaccine or control and the other is always unvaccinated. All of the above results would also hold if within each household, one person were randomized to vaccination or control and all others were unvaccinated; in this case Y_{i2} would simply be a vector. This was in fact the design used in the study of Millar et al. (2008).

In the study design we have been considering, all households have either one person or no one vaccinated. We might instead consider a design in which each of the two persons is randomized to receive vaccine or control, so that in some households neither individual is vaccinated, in some just one, and in some both. In this more general design, in addition to the causal infectiousness defined above in (A15.2), we could also define a further causal infectiousness effect,

$$\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(1,0) | Y_{i2}(1,1) = Y_{i2}(1,0) = 1] \tag{15.3}$$

that is, the effect on individual 1 of individual 2's receiving the vaccine when individual 1 is also vaccinated within the subpopulation for whom individual 2 would be infected irrespective of whether individual 2 received the vaccine. Once again under this more general design the crude estimator, $\mathbb{E}[Y_{i1}|A_{i1} = 1, A_{i2} = 1, Y_{i2} = 1] - \mathbb{E}[Y_{i1}|A_{i1} = 1, A_{i2} = 0, Y_{i2} = 1]$, will be conservative for this causal infectiousness effect under assumptions similar to assumptions (A15.1) and (A15.2) above.

Other subtleties can arise when the infectiousness effect is of interest. VanderWeele and Tchetgen Tchetgen (2011a) consider different levels of virulence of the pathogen that may be present if individual 1 is infected with versus without the vaccine and develop conservative estimators for the infectiousness effect in this setting. Halloran and Hudgens (2012a,b) further consider additional inferences that can be drawn if something is known about the timing concerning the infections of individuals 1 and 2. We have focused here on the setting of a randomized trial, but the approach is potentially applicable to observational studies as well if, conditional on some set of covariates C, the treatment was jointly independent of the counterfactual outcomes (i.e., effectively randomized within strata of C). The sensitivity analysis parameters would then also have to be conditional on C.

15.4. CONTAGION VERSUS INFECTIOUSNESS EFFECTS

As discussed in the last section, when the vaccine of one person prevents infection in a second person in the same household, a further distinction can be drawn between the ways such a protective effect might arise. Vaccinating the first person may protect the second person by preventing the first from being infected and passing the infection on to the second. Alternatively, vaccinating the first person may protect the second by rendering the infection less transmissible even if the first person is

infected. We referred to this latter mechanism as an "infectiousness effect" of the vaccine. We might refer to the former as a "contagion effect," following terminology in the social network literature (Christakis and Fowler, 2007). Although the terms "infectiousness" and "contagion" are sometimes used interchangeably in the infectious disease literature, there are clearly two distinct mechanisms or pathways. In this section we will consider decomposing the spillover/indirect effect of one person's vaccine on another's outcome into two components corresponding to such "infectiousness" and "contagion" effects. The technical development is somewhat lengthy and the reader may at any point skip the remainder of this section and move ahead to Section 15.5 on other aspects of spillover effects. None of remainder of the chapter is dependent on the following material in the current section.

To illustrate the distinctions and methods, we will consider a hypothetical vaccine trial setting in which one-year-olds at a day-care center are randomized to receive pneumococcal conjugate vaccine. The colonization status of the one-year-old and one of its parents (the mother, say) is also monitored. We assume that pneumococcus is highly prevalent in young children who attend day care, and thus the probability that the mother is acquires pneumococcus from transmission routes other than the child is negligible. In settings in which it can effectively be assumed that, at least for the study period, the second person (e.g., the mother) can be infected only from the first (e.g., the one-year-old), we will see that the average spillover effect defined in Section 15.2 can itself can be decomposed into a contagion effect and what will be defined below as an unconditional infectiousness effect. Understanding what proportion of an indirect effect is due to decreasing infectiousness can give insight into the mechanism by which the vaccine protects others.

The methods in this section essentially arise from methods for mediation analysis considered in Part I of this book in assessing the extent to which the effect of an exposure on an outcome is mediated by a particular intermediate variable and the extent to which it is "direct" or through other pathways. Within the context of a vaccine trial, we will essentially take the vaccine status of one person as the exposure variable, the infection status of that person as the intermediate variable, and the infection status of a second person in the same household as the outcome variable. We consider effect measures and assumptions to interpret estimates causally within this context (cf. VanderWeele et al., 2012d).

15.4.1. Definition of Contagion and Infectiousness Effects

As in Section 15.3, we assume a simple randomized experiment in which one of the two persons is randomized to receive a vaccine or control and the second person is always unvaccinated. This could correspond to the hypothetical pneumococcal vaccine trial described above where we are interested in the effect on the mother of vaccinating the one-year-old. We will let $j = 1$ denote the individual who may or may not be vaccinated and let $j = 2$ denote the individual who is always unvaccinated. Unlike Section 15.3, we will assume that only person 1, not person 2, can be infected from outside the household; person 2 can be infected only by person 1.

This might be somewhat plausible in the day care example above if, because pneumococcus is highly prevalent in young children who attend day care, the mother is much more likely to acquire the pneumococcus from the child than through other transmission routes. It might also be plausible in settings in which the second person is homebound (e.g., an elderly person who does not leave the household). If this assumption holds, then we have that $Y_{i1}(a_{i1}, a_{i2}) = 0$ implies $Y_{i2}(a_{i1}, a_{i2}) = 0$, since person 2 cannot be infected if person 1 is not. Note that the assumption that individual 2 is always unvaccinated allows a simplified notation. Counterfactuals $Y_{i1}(a_{i1}, a_{i2})$ and $Y_{i2}(a_{i1}, a_{i2})$ can be written as $Y_{i1}(a_{i1}) := Y_{i1}(a_{i1}, 0)$ and $Y_{i2}(a_{i1}) := Y_{i2}(a_{i1}, 0)$. We are still, however, allowing interference/spillover in that the vaccine of person 1 affects the outcome of person 2.

As in Section 15.2, the average indirect effect of the vaccine for person 1 on the outcome of person 2 is

$$\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0)]$$

that is, the difference in infection status for person 2 if person 1 is vaccinated versus unvaccinated. If vaccine status for person 1 is randomized, this can be estimated by

$$\mathbb{E}[Y_{i2}|A_{i1} = 1, A_{i2} = 0] - \mathbb{E}[Y_{i2}|A_{i1} = 0, A_{i2} = 0]$$

To proceed with decomposing this indirect effect into the two effects, we need to consider counterfactuals of a different form. Suppose that in addition to potentially intervening to give person 1 the vaccine we could also, at least hypothetically, intervene to infect or not infect person 1. Then $Y_{i2}(a_{i1}, a_{i2}, y_{i1})$ would denote the infection status of person 2 if we would set the vaccine status of person 1 and person 2 to a_{i1} and a_{i2} and the infection status of person 1 to y_{i1}. The simple setting in which person 2 is always unvaccinated also allows us to rewrite the counterfactual $Y_{i2}(a_{i1}, a_{i2}, y_{i1})$ as $Y_{i2}(a_{i1}, y_{i1}) := Y_{i2}(a_{i1}, 0, y_{i1})$. We thus consider counterfactuals of the form $Y_{i1}(a_{i1})$, $Y_{i2}(a_{i1})$ and $Y_{i2}(a_{i1}, y_{i1})$. The direct effect of person 1's vaccine on person 1's outcome is $\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)]$; the indirect effect of person 1's vaccine on person 2's outcome is simply $\mathbb{E}[Y_{i2}(1) - Y_{i2}(0)]$. In the next section we will use these counterfactuals to define contagion and unconditional infectiousness effects.

Consider now the counterfactual contrast

$$\mathbb{E}[Y_{i2}(0, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(0))]$$

The term $Y_{i2}(0, Y_{i1}(1))$ considers what the potential infection outcome of person 2 would be if person 1 is left unvaccinated but we set the infection status of person 1 to the level it would have been if person 1 was vaccinated. The contrast compares this counterfactual to $Y_{i2}(0, Y_{i1}(0))$, the potential infection outcome of person 2 if person 1 is not vaccinated, and we also set the infection status of person 1 to the level it would be if person 1 was unvaccinated. For this contrast to be nonzero, $Y_{i1}(1)$ and $Y_{i1}(0)$ have to differ—that is, vaccination of person 1 would have to affect the infection status of person 1—and this change in infection for person 1 would have to change the infection status for person 2, even if person 1 had been left

unvaccinated. Essentially, the contrast is nonzero if the vaccine prevents infection in person 1, and this in turn prevents person 2 from being infected. We refer to this counterfactual contrast as a contagion effect.

Consider now the contrast

$$\mathbb{E}[Y_{i2}(1, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(1))]$$

This compares the potential infection outcome of person 2 if person 1 had been vaccinated versus unvaccinated and person 1 had the infection status that would occur if vaccinated. This contrast will be nonzero only if person 1 is infected when vaccinated (because person 1's vaccination status will not affect person 2's outcome unless person 1 is infected). If the contrast is nonzero, this will be because even when person 1 is vaccinated and infected, the vaccine itself affects whether person 2 is infected by person 1. This new measure is in some ways analogous to what in infectious disease epidemiology is called an infectiousness effect (Datta et al., 1999; Préziosi and Halloran, 2003) and to the infectiousness effect in the previous section. However, this new measure differs in essential ways from the ordinary or "conditional infectiousness effect" considered in the previous section in that it does not condition on person 1 actually being infected. The standard (conditional) infectiousness effect would, in contrast, compare outcomes for person 2 when person 1 is, versus is not, vaccinated but would only make this comparison for subgroups in which 1 is actually infected; and in the previous section we considered counterfactual-based formalizations of this conditional effect. We will return again below to the relation of the conditional and unconditional infectiousness effects. Until then, however, our discussion will focus on this new "unconditional infectiousness effect" and we will thus omit the word "unconditional" before "infectiousness effect" unless otherwise needed for clarity.

These counterfactual definitions of the contagion effect and infectiousness effects have the desirable feature that we can decompose an indirect/spillover into a contagion and an infectiousness effect by taking the indirect effect and adding and subtracting the term $\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]$:

$$\begin{aligned}\mathbb{E}[Y_{i2}(1) - Y_{i2}(0)] &= \mathbb{E}[Y_{i2}(1, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(0))] \\ &= \mathbb{E}[Y_{i2}(1, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(1))] \\ &\quad + \mathbb{E}[Y_{i2}(0, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(0))]\end{aligned}$$

where the first term in the sum is the infectiousness effect and the second term in the sum is the contagion effect. This decomposition is analogous to what we referred to in Part I of this book as "natural direct and indirect effects." We exploit this analogy here in our discussion of identification, estimation, and sensitivity analysis. As already noted above, the term "indirect effect" is used differently in mediation analysis than in causal inference with interference. Further attention to these issues of terminology are given below.

Thus far we have been considering measures of effect on a risk-difference scale. However, risk-ratio, odds-ratio, or vaccine-efficacy measures are more commonly

employed in the vaccine literature. The effects and their decomposition described above have analogues for ratio and vaccine-efficacy measures. For example, the indirect effect on the risk ratio and odds-ratio scale could be defined as $\frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]}$ or $\frac{\mathbb{E}[Y_{i2}(1)]/\{1-\mathbb{E}[Y_{i2}(1)]\}}{\mathbb{E}[Y_{i2}(0)]/\{1-\mathbb{E}[Y_{i2}(0)]\}}$. The decomposition for the risk ratio is

$$\frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]} = \frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]} \times \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}$$

Here the first term in the product is the infectiousness effect on the risk-ratio scale and the second term is the contagion effect on the risk-ratio scale; the indirect effect is the product of the contagion and infectiousness effects on the risk-ratio scale, rather than their sum. A similar decomposition holds for odds-ratio measures.

Similar definitions and a somewhat analogous decomposition holds with a vaccine efficacy measure. The vaccine efficacy measure for the indirect effect would be defined as

$$VE_{indirect} = 1 - \frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]}$$

We might likewise define vaccine efficacy for the contagion effect and infectiousness effect as

$$VE_{cont} = 1 - \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}$$

$$VE_{\text{inf}} = 1 - \frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}$$

Some algebra gives

$$1 - \frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]} = \left(1 - \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}\right) + \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]} \left(1 - \frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}\right)$$

and we thus have

$$VE_{indirect} = VE_{cont} + \left(\frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}\right) VE_{\text{inf}}$$

In words, the vaccine efficacy measure for the indirect effect is the sum of the vaccine efficacy for the contagion effect and that of the infectiousness effect, where the vaccine efficacy of the infectiousness effect is adjusted by the factor $\left(\frac{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(0))]}\right)$ to account for the fact that when the infectiousness effect operates, the contagion effect has essentially already occurred (the infectiousness effect makes the infection less infectious, but this infectiousness effect will not operate if the vaccine in fact prevents person 1 from being infected).

Each of these effect measures could also be defined conditional on covariates C_i. For example, the contagion and infectiousness effects on the risk ratio scale conditional on covariates $C_i = c$ would be $\frac{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))|C_i=c]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(0))|C_i=c]}$ and $\frac{\mathbb{E}[Y_{i2}(1,Y_{i1}(1))|C_i=c]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))|C_i=c]}$, respectively.

15.4.2. More on Terminology

In the mediation analysis, "indirect effect" describes the effect of an exposure on an outcome for one person that operates through some intermediate or mediator in that same person, also called a mediated effect. In causal inference in the presence of interference, the indirect effect (also called a "spillover effect" in the social sciences) of, say, vaccinating some persons in a population is a contrast of potential outcomes comparing the outcomes in those other persons who did not receive the vaccine to what their outcomes would have been if the vaccinated persons were not vaccinated. Clearly there are the close relations between what we have defined as the "contagion and unconditional infectiousness effects" on the one hand and "natural direct and indirect effects" on the other.

In mediation analysis, "indirect effect" is used to describe situations in which the effect of an exposure on an outcome for one person operates through some intermediate or mediator for that individual. The "contagion effect" and "infectiousness effect" in this section are, analytically somewhat analogous to the "natural indirect effect" and "natural direct effect," respectively, in mediation analysis. The "contagion effect" is essentially the effect of person 1's vaccine on person 2's infection outcome mediated by person 1's infection outcome. The "unconditional infectiousness effect" is essentially the effect of person 1's vaccine on person 2's infection outcome not mediated by person 1's infection outcome.

In the infectious disease and vaccine literature, the "indirect effect of vaccination" is used to describe settings in which vaccination of one person affects the outcome of another individual. In other statistical and causal inference literature effects due to interference are sometimes called "spillover effects." Here, we are decomposing the "indirect effect of a vaccination" in the literature on causal inference in the presence of interference into the "natural indirect effect" and "natural direct effect" of mediation analysis. Because these two literatures—causal inference in the presence of interference on the one hand and causal inference mediation analysis on the other hand—use the same terms for different concepts, and moreover because, as we have seen here, these concepts are not entirely unrelated, it is important to clarify in each instance specifically the various terms being employed. Because of the terminological overlap, the language employed can be somewhat confusing. We give a glossary in Table 15.1 to help guide the reader through the various terms used in these literatures.

15.4.3. Identification of Contagion and Infectiousness Effects

We have defined the contagion and unconditional infectiousness effects in terms of counterfactuals that are not immediately estimable from the data. Although these

effects may be of substantive interest, we cannot estimate them without further assumptions. We continue to assume that person 2 cannot be infected from outside the household, only by transmission from person 1. Suppose that data are available on some set of baseline covariates C_i that may be attributes of person 1 or of person 2 or of their household. Conditional on the set of covariates C_i we make the following assumptions:

(A15.3) The effect of A_{i1} on Y_{i2} is unconfounded conditional on C_i
(A15.4) The effect of Y_{i1} on Y_{i2} is unconfounded conditional on (C_i, A_{i1})
(A15.5) The effect of A_{i1} on Y_{i1} is unconfounded conditional on C_i
(A15.6) Given that (A15.4) holds, there is no confounder of the relationship between Y_{i1} and Y_{i2} that is itself affected by A_{i1}

Under these four assumptions, the contagion and infectiousness effects are identified from the data. In the appendix we give formal counterfactual statements of these assumptions. The assumptions can also be illustrated using the diagram in the Figure 15.1, which assumes the vaccine status of person 1 is randomized.

Table 15-1. GLOSSARY OF MAIN TERMS USED FOR SPILLOVER EFFECTS IN THIS SECTION

Indirect Effect/Spillover Effect: The effect the vaccine status of other people on a particular individual's outcome
Contagion Effect: The effect of one person's vaccine on another's outcome by preventing the first from being infected; analytically somewhat analogous to the natural indirect effect of mediation analysis.
(Conditional) Infectiousness Effect: The effect of a vaccine on rendering the infection of an infected person less infectious
(Unconditional) Infectiousness Effect: An infectiousness effect that averages over also those households for whom person 1 is uninfected; analytically somewhat analogous to the natural direct effect of mediation analysis.
Direct Effect: The effect of a person's vaccination status on his or her outcome
Interference: The phenomenon whereby the exposure (vaccination) of one person can affect the outcome of another.

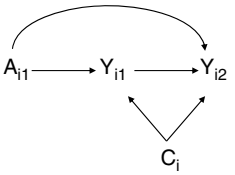

Figure 15.1 Vaccine trial in which person 1 is randomized to vaccine and person 2 does not receive the vaccine. A_{i1} denotes the vaccine status of person 1; Y_{i1} denotes the infection status of person 1; Y_{i2} denotes the infection status of person 2; C_i denotes individual and household covariates for household *i*.

The four assumptions are analogous to assumptions (A2.1)–(A2.4), which we considered in Chapter 2 to identify natural direct and indirect effects. We now describe the four assumptions in a bit more detail. If, as assumed, vaccine status of person 1 is randomized, then assumptions (A15.3) and (A15.5) will hold by randomization. In an observational setting, assumptions (A15.3) and (A15.5) would hold only if a sufficiently rich set of covariates C_i were available such that vaccination was effectively randomized within strata of covariates C_i.

Assumption (A15.4) effectively requires that within the set of available covariates C_i we have all variables that are common causes of person 1's infection status and person 2's infection status (see Figure 15.1). Such common causes might include, for example, environmental factors related to the sanitary, spatial, and nutritional characteristics of the household. Assumption (A15.4) is a strong assumption. It can perhaps be made more plausible by attempting to control for such variables, but in general it will not be possible to verify assumption (A15.4). Assumption (A15.6), by contrast, is arguably somewhat weaker: It requires that of all the common causes of person 1's and person 2's infection status, none is affected by the vaccine itself; that is, there is no arrow from A_{i1} to C_i in the figure. Because most of these common causes are likely to be characteristics of the household environment, it seems reasonably plausible that such characteristics would not be changed by the vaccine.

The key to identifying the contagion and infectiousness effects thus arguably lies with trying to ensure the validity of assumption (A15.4): trying to adjust for covariates that may be common causes of person 1's and person 2's infection status. We will consider sensitivity analysis for violations of this assumption below. First, however, we will discuss a statistical modeling approach to the estimation of these effects when they are identified.

15.4.4. Statistical Models to Estimate Contagion and Infectiousness Effects

We now consider using two logistic regression models to estimate the contagion and infectiousness effects when they are in fact identified. Suppose that the following two logistic regression models are fit to the observed data: (i) one model for the probability of infection for person 1 conditional on person 1's vaccine status a_1 and the covariates c and (ii) a second model for the probability of infection for person 2, conditional on person 1's vaccine status a_1, person 1's infection outcome and the covariates c:

$$\text{logit}\{P(Y_1 = 1|a_1, c)\} = \beta_0 + \beta_1 a_1 + \beta_2' c$$
$$\text{logit}\{P(Y_2 = 1|a_1, Y_1 = 1, c)\} = \theta_0 + \theta_1 a_1 + \theta_4' c$$

Note that under the assumption that person 2 cannot be infected from outside the household, we have $\mathbb{E}[Y_{i2}|a_{i1}, Y_{i1} = 0, c] = 0$ and thus we do not need to model $\mathbb{E}[Y_{i2}|a_{i1}, Y_{i1} = 0, c]$; a model for $\text{logit}\{P(Y_2 = 1|a_1, Y_1 = 1, c)\}$ suffices and thus we can ignore terms such as $\theta_2 y_1$ and $\theta_3 a_1 y_1$, as we had in Chapter 2, in this model (cf. Ogburn and VanderWeele, 2014a).

The following results suppose that the infection outcome for person 2 is rare enough for the odds ratios to approximate risk ratios and the logistic link to approximate a log link. If the infection outcome for person 2 is not rare, then the results given below will hold if the logistic regression model for Y_2 is replaced by a log-linear model while the model for Y_1 is kept as a logistic model. No rare-outcome assumption or log-linear model is needed for Y_1.

If covariates C_i satisfy assumptions (A15.3)–(A15.6), and the models above are correctly specified, then, as shown in the Appendix, the contagion effect on the risk-ratio scale conditional on the covariates $C_i = c$ is given by

$$\frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))|c]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))|c]} = \frac{(1 + e^{\beta_0 + \beta_2' c})(e^{\beta_0 + \beta_1 + \beta_2' c} + 1)}{(1 + e^{\beta_0 + \beta_1 + \beta_2' c})(e^{\beta_0 + \beta_2' c} + 1)} \tag{15.4}$$

provided $\theta_1 \neq -\infty$ (i.e., provided that the probability of $Y_2 = 1$ is not zero; if it were zero, then person 2 would never be infected; there would be no effect of person 1's infection on person 2 and thus no mediation or contagion). The infectiousness effect on the risk-ratio scale conditional on the covariates is given by

$$\frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))|c]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))|c]} = e^{\theta_1} \tag{15.5}$$

These expressions can be obtained directly from the estimates of the logistic regression parameters. Standard errors could be obtained by bootstrapping or the delta method. See Appendix for further details.

15.4.5. Sensitivity Analysis for Contagion and Infectiousness Effects

Identification and estimation of the contagion and infectiousness effects depend critically on assumptions (A15.3)–(A15.6). Unfortunately, these are fairly strong assumptions, especially assumption (A15.4). In this section we give a relatively straightforward sensitivity analysis technique, adapted from those in Chapter 3 for this context, that can be employed to assess how vulnerable one's estimates and conclusions are to violations of assumption (A15.4). The technique assumes that there is an unmeasured binary confounding variable U that is a common cause of the infection status of person 1 and person 2, as well as that assumptions (A15.3)–(A15.6) would hold conditional on (C_i, U) but not on the measured covariates C_i alone. The investigator specifies sensitivity parameters corresponding to (i) the effect of the unmeasured confounding U on the infection status of person 2 conditional on the vaccine status of person 1, the infection status of person 1, and the observed covariates C_i and (ii) the prevalence of U within strata defined by the vaccine status of person 1, conditional on the observed covariates C_i and person 1's being infected. The technique then uses the estimates obtained by controlling only for observed covariates C_i, along with these sensitivity parameters, to calculate the corrected estimates that would have been obtained had it been possible to control for the unmeasured confounding variable U as well. The sensitivity-analysis parameters can be varied across a range of plausible values to assess how sensitive

the conclusions and estimates are to a potential unmeasured common cause of the infection status of person 1 and person 2.

The technique assumes that the effect of U on the infection status of person 2, when person 1 is infected, is constant across the vaccine status of person 1 and is given by

$$\gamma = \frac{P(Y_2 = 1|a_1, Y_1 = 1, c, U = 1)}{P(Y_2 = 1|a_1, Y_1 = 1, c, U = 0)}$$

The sensitivity analysis parameter γ thus captures the effect of U on the infection status of person 2. The investigator also specifies the prevalence of U in each stratum defined by the vaccine status of person 1 and the infection status of person 1 conditional on the observed covariates C_i:

$$\pi_{01} = P(U = 1|A_1 = 0, Y_1 = 1, c)$$
$$\pi_{11} = P(U = 1|A_1 = 1, Y_1 = 1, c)$$

From these sensitivity-analysis parameters the following bias factor can be calculated:

$$B = \frac{1 + (\gamma - 1)\pi_{11}}{1 + (\gamma - 1)\pi_{01}}$$

It follows from our discussion of sensitivity analysis in Chapter 3 that if we replace θ_1 with θ_1/B in formula (15.5), this gives a corrected infectiousness effect estimate corresponding to what would have been obtained had we been able to adjust for U and C_i rather than only the observed covariates C_i alone. The contagion effect, as it turns out, is not biased by such unmeasured confounding in this context. In general we will not know the true values of the sensitivity-analysis parameters; however, varying the parameters γ and π_{01}, π_{11} will give some sense as to how sensitive the results are to potential unmeasured common causes of the infection status of person 1 and person 2. The sensitivity analysis technique is of course also limited by the assumptions made, which are (i) a single unmeasured binary confounder and (ii) the effect of U on the infection status of person 2 is constant across vaccine status of person 1.

15.4.6. Illustration

Consider data from a hypothetical vaccine trial in Table 15.2 in which one-year-olds ($j = 1$) are randomized to pneumococcal conjugate vaccine with follow-up for both the one-year-olds and their mothers ($j = 2$).

In this example, pneumococcus is assumed to be highly prevalent in children in day care, so that the mother is much more likely to acquire the pneumococcus from the child than through other transmission routes and thus that during the study period the mother is infected only from the one-year-old. Suppose we fit a logistic model for the probability of infection for person 1 conditional on person

Table 15-2. NUMBERS INFECTED, (Y_{i1}, Y_{i2}), FROM A HYPOTHETICAL RANDOMIZED TRIAL OF PNEUMOCOCCAL CONJUGATE VACCINE WITH 2000 HOUSEHOLDS, WHERE PERSON 1 (ONE-YEAR-OLD) WAS RANDOMIZED 1:1 TO VACCINE OR CONTROL, AND PERSON 2 (THE MOTHER) WAS NOT VACCINATED (VACCINATION STATUS (A_{i1}, A_{i2})) AND HALF THE HOUSEHOLDS HAVE EITHER LOW OR HIGH SOCIOECONOMIC STATUS (SES)

	$Y_{i1}=0$, $Y_{i2}=0$ ($n=1200$)	$Y_{i1}=1$, $Y_{i2}=0$ ($n=416$)	$Y_{i1}=1$, $Y_{i2}=1$ ($n=384$)	Total ($n=2000$)
Low SES: $A_{i1}=0, A_{i2}=0$	200	120	180	500
$A_{i1}=1, A_{i2}=0$	350	96	54	500
High SES: $A_{i1}=0, A_{i2}=0$	250	125	125	500
$A_{i1}=1, A_{i2}=0$	400	75	25	500

1's vaccine status a_1 and the covariates c and a log-linear model for the probability of infection for person 2, conditional on person 1's vaccine status a_1, person 1's infection outcome and the covariates c. Using expressions (15.4) and (15.5) above for the contagion and infectiousness effects, and setting the covariate to its mean value, we obtain, on the risk-ratio scale, under assumptions (A15.3)–(A15.6), an overall estimate of the spillover/indirect effect of 0.26 (95% CI: 0.20, 0.32), an estimate of the contagion effect of 0.45 (95% CI: 0.40, 0.51), and an estimate of the infectiousness effect of 0.57 (95% CI: 0.47, 0.69). The indirect effect on the risk-ratio scale decomposes into the product of the contagion and infectiousness effects: $0.26 = 0.45 \times 0.57$. On the vaccine-efficacy scale, we would have an overall indirect effect of $1 - 0.26 = 74\%$, a contagion effect of $1 - 0.45 = 55\%$, an infectiousness effect of $1 - 0.57 = 43\%$, and vaccine-efficacy component due to the infectiousness effect of $(0.45)(43\%) = 19\%$ (essentially taking into account the fact that the infectiousness effect will operate only if the contagion effect has not). We can then decompose the indirect effect on the vaccine efficacy scale into the sum of the contagion effect and the component due to infectiousness: $74\% = 55\% + 19\%$. In this hypothetical example, the contagion effect seems somewhat more important than the infectiousness effect.

15.4.7. Deciding on Effects of Interest and Further Extensions

In the previous subsections we have considered how a spillover/indirect effect of vaccination of one person on the outcome of another can be decomposed into two components: one corresponding to the vaccine preventing the infection in person 1, which then protects person 2 (the contagion effect), and another corresponding to the fact that even if person 1 is infected the vaccine may render the infection less infectious (the unconditional infectiousness effect). In Section 15.3 we discussed the conditional infectiousness effect examining the effect of the vaccine of person 1 on the infection status of person 2 in the principal stratum in which person 1 would

be infected irrespective of vaccine status. This infectiousness effect based on principal strata is different from that considered here: Essentially the "principal stratum" infectiousness effect is a conditional effect (it conditions on the subgroup for which person 1 would be infected irrespective of vaccine status), and the traditional naive infectiousness effect conditions on person 1 being infected regardless of principal stratum. However, the infectiousness effect considered here is an unconditional infectiousness effect: It averages over also those clusters for whom person 1 is uninfected (for which any potential infectiousness effect of the vaccine would not have the opportunity to operate).

These issues are important in the interpretation of these effects; both types of infectiousness effects (conditional and unconditional) could potentially be reported. Using the methodology and assumptions of Section 15.3, an upper bound on the conditional infectiousness effect on the risk-ratio scale from the data in Table 15.2 would be 0.57 (a lower risk ratio implies a stronger protective effect). This in fact coincides with the unconditional infectiousness effect risk ratio of 0.57 reported in Section 15.3. In fact it can be shown that under the assumption that individual 2 cannot be infected from outside of the household and also assumption (A15.1) that for all i, $Y_{i1}(1,0) \leq Y_{i1}(0,0)$ the conditional and unconditional infectiousness effect will coincide on the ratio scale and, on the difference scale, the unconditional infectiousness effect will be equal to the infectiousness effect multiplied by a scaling factor $P(Y_1 = 1|A_1 = 1)$ (Chiba and Taguri, 2013). The advantage of the infectiousness effect given in this section (the unconditional version) is that it can be used to decompose the overall effect into the contagion and infectiousness components.

We have considered the setting in which there are two persons per cluster and only one person is randomized to vaccination. However, in settings in which both are randomized to vaccination, the analysis could be pursued separately for households in which person 2 is or is not vaccinated. Another simple extension of the approach here might involve settings in which only one person in each household is randomized to vaccine but outcome data are collected on numerous additional persons per household. In such cases the outcome Y_{i2} above could be replaced with the proportion in the household who are infected (other than the person randomized); the logistic regression would then have to be replaced with a linear or log-linear regression, but similar methods from the mediation-analysis literature could potentially be adapted and applied. If the numbers in each household vary across households, this number could also be controlled for in the analysis.

15.5. TESTS FOR SPECIFIC FORMS OF INTERFERENCE USING CAUSAL INTERACTIONS

In this section we will return to the general setting in which both individuals in a household might be exposed/vaccinated (cf. VanderWeele et al., 2012f), rather than assuming that only one individual is randomized to receive the vaccine. We

will also allow all individuals to be infected (or more generally have the outcome occur) from either within or outside of the household. We will see that not only can we test for spillover effects as described above, but we can also sometimes detect specific forms of interference. We can do so essentially by employing the tests for mechanistic interaction described in Chapter 9 and 10 to the interference context. We will very briefly review these tests and then discuss how they can be used to draw conclusions about the specific form of interference that may be present.

15.5.1. Review of Tests for Mechanistic Interaction

Let us suppose we are no longer in the interference context and that we have two binary exposures of interest, A_1 and A_2, and a binary outcome D; that is, we assume a person's outcome depends on the values of the two exposures that the person receives, but not on the exposures received by other people. Let $D_{a_1a_2}$ denote the counterfactual outcome (or potential outcome) for D for a person if (possibly contrary to fact) A_1 had been set to a_1 and A_2 had been set to a_2. In Chapters 9 and 10, we said that a sufficient-cause interaction was present between A_1 and A_2 if there was some person such that $D_{11} = 1$ but $D_{10} = D_{01} = 0$. With a sufficient cause interaction, D_{00} is allowed to be 0 or 1. For such a person the outcome occurs if both exposures are present, but not if just one or the other is present, and for this reason this was referred to as a form of "mechanistic interaction."

We will assume for simplicity that the effects of the exposures A_1 and A_2 on D are unconfounded (e.g., randomized). Let $p_{a_1a_2} = P(D = 1|A_1 = a_1, A_2 = a_2)$ be the observed probability of the outcome among those who actually received $A_1 = a_1, A_2 = a_2$. As discussed in Chapters 9 and 10, the standard condition for positive additive interaction is

$$p_{11} - p_{10} - p_{01} + p_{00} > 0 \tag{15.6}$$

If the effects of A_1 and A_2 on D were unconfounded, then if a slightly different condition holds, namely,

$$p_{11} - p_{10} - p_{01} > 0 \tag{15.7}$$

then this would imply the presence of a sufficient-cause interaction (VanderWeele and Robins, 2007b, 2008). The test for a sufficient-cause interaction in condition (15.7) is a more stringent condition than the standard additive interaction being positive in condition (15.6), in that we are no longer adding p_{00} in the probability contrast. In fact the magnitude of the contrast $p_{11} - p_{10} - p_{01}$ gives a lower bound on the prevalence of persons with a sufficient cause interaction.

In Chapters 9 and 10 it was also noted that the test for a standard additive interaction $p_{11} - p_{10} - p_{01} + p_{00} > 0$ would suffice to conclude the presence of a sufficient-cause interaction under an assumption called "monotonicity." The effects of A_1 and A_2 are said to be positive monotonic if $D_{a_1a_2}$ is nondecreasing in a_1 and a_2 for all persons—that is, if an increase in the exposure A_1 or A_2 would increase or leave unchanged the outcome, not just on average, but for all people in the population. Under this assumption that A_1 and A_2 have positive monotonic effects on the outcome, the standard additive interaction contrast $p_{11} - p_{10} - p_{01} + p_{00} > 0$

would suffice to draw the conclusion of the presence of a sufficient-cause interaction, and the contrast $p_{11} - p_{10} - p_{01} + p_{00}$ then gives a lower bound on the prevalence of individuals with a sufficient-cause interaction. Testing condition $p_{11} - p_{10} - p_{01} > 0$ would be necessary without such monotonicity assumptions. It should be noted that these conditions are sufficient but not necessary conditions for a sufficient-cause interaction. A sufficient-cause interaction may be present even if these conditions are not satisfied.

Chapters 9 and 10 also discussed empirical tests for an even stronger notion of interaction. We might say that there is an "epistatic" or "singular" interaction if there is some person such that $D_{11} = 1$ and $D_{10} = D_{01} = D_{00} = 0$. This means that for this person the outcome occurs if and only if both exposures are present. Note that this is an even stronger notion of mechanistic interaction than that of a sufficient-cause interaction, in that we are now requiring that $D_{00} = 0$. Although it is in general a stronger notion of interaction, if at least one of the two exposures has a positive monotonic effect on the outcome, then the notions of a "sufficient cause" interaction and a "singular/epistatic" interaction coincide.

If the effects of A_1 and A_2 on D were unconfounded, then

$$p_{11} - p_{10} - p_{01} - p_{00} > 0 \tag{15.8}$$

would imply an "epistatic interaction"; the contrast $p_{11} - p_{10} - p_{01} - p_{00}$ in fact gives a lower bound on the prevalence of individuals that manifest such a singular/epistatic interaction. In contrast with the condition of the standard interaction contrast in (15.6), in the condition in (15.8) we now subtract rather than add p_{00}. If at least one of the two exposures has a positive monotonic effect on the outcome, then we can test $p_{11} - p_{10} - p_{01} > 0$ to conclude the presence of a singular/epistatic interaction. If both exposures have positive monotonic effects on the outcome, then we can use the standard interaction contrast $p_{11} - p_{10} - p_{01} + p_{00} > 0$ to test for the presence of a singular/epistatic interaction. Statistical tests for these conditions were discussed in Chapters 9 and 10.

15.5.2. Tests for Specific Forms of Interference

Let us now return to the setting of interference in which we have two individual's per household but only one exposure for each individual: A_{i1} for individual 1 and A_{i2} for individual 2. Suppose now that we are interested in trying to detect patterns of interference of a particular form. For example, we might be interested in whether there are any households such that the wife is not infected if and only if both the husband and the wife are vaccinated. Expressed in terms of counterfactuals, we would be asking whether there is some household i such that $Y_{i1}(1,1) = 0$ but $Y_{i1}(1,0) = Y_{i1}(0,1) = Y_{i1}(0,0) = 1$. This pattern is somewhat analogous to the epistatic/singular interaction considered above. In fact, by redefining our outcome, we can test for it empirically. For each household i, define $D_i(a_{i1}, a_{i2}) = 1 - Y_{i1}(a_{i1}, a_{i2})$; that is, $D_i(a_{i1}, a_{i2})$ is an indicator that the wife is not infected if the wife and husband receive vaccines corresponding to a_{i1} and a_{i2}, respectively. Suppose that both the wife's and husband's vaccine status were randomized. If we

then let $p_{a_1a_2} = P(D = 1|A_1 = a_1, A_2 = a_2) = P(Y_1 = 0|A_1 = a_1, A_2 = a_2)$, then by using the tests for causal interaction we would have that if

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

then there must be some households such that the wife is not infected if and only if both the husband and the wife are vaccinated. In fact, the contrast $p_{11} - p_{10} - p_{01} - p_{00}$ will be a lower bound on the prevalence of such households. This is now a conclusion concerning a much more specific form of interference than simply the presence of some spillover effect, as in the previous sections. By redefining $D_i(a_{i1}, a_{i2}) = 1 - Y_{i2}(a_{i1}, a_{i2})$, we could test for similar patterns of interference for the husband. Similarly, by yet other alternative definitions for D_i, we could attempt to detect yet other forms of potential interference. If, for example, we define $D_i(a_{i1}, a_{i2}) = [1 - Y_{i1}(a_{i1}, a_{i2})] \times [1 - Y_{i2}(a_{i1}, a_{i2})]$ and find, for $p_{a_1a_2} = P(D = 1|A_1 = a_1, A_2 = a_2)$ with D so defined, that $p_{11} - p_{10} - p_{01} - p_{00} > 0$, then we could conclude that there were households such that the husband and the wife both remain uninfected if and only if both receive the vaccine. By defining D_i as other combinations of the wife's and husband's outcomes, other forms of interference or response patterns could potentially be tested for. As with causal interactions, however, so also here with tests for specific forms of interference, the conditions tested are sufficient for the specific form of interference in question, but they are not necessary. Such forms of interference might be present even if the condition on the probabilities is not satisfied.

However, as was also the case for "causal interactions," so too here, when testing for various forms of interference, monotonicity assumptions will allow us to test weaker conditions. Consider again trying to test for whether there are any households such that the wife is not infected if and only if both the husband and the wife are vaccinated. Suppose that we thought the wife's vaccine would never cause the wife to be infected, so that monotonicity held for A_{i1} (i.e. if $D_i(a_{i1}, a_{i2}) = 1 - Y_{i1}(a_{i1}, a_{i2})$ denotes the absence of the wife being infected, then A_{i1} will have a positive monotonic effect on D_i). By the results on causal interaction, it follows that for $p_{a_1a_2}$, defined as $p_{a_1a_2} = P(D = 1|A_1 = a_1, A_2 = a_2) = P(Y_1 = 0|A_1 = a_1, A_2 = a_2)$, if we found that

$$p_{11} - p_{10} - p_{01} > 0$$

then at least some households would have to be such that the wife is not infected if and only if both the husband and the wife are vaccinated. Similarly, if we also thought that the vaccination of the husband would never, for any household, cause the wife to be infected, then to test for households with this specific interference pattern for the wife we could test

$$p_{11} - p_{10} - p_{01} + p_{00} > 0$$

Note that in this case the monotonicity assumptions for the wife's and the husband's vaccine are somewhat different insofar as both pertain to the outcome status of the wife. These monotonicity assumptions would be violated if the vaccine might itself

cause the infection. But if, in a specific context, the monotonicity assumptions are thought reasonable, then these weaker conditions, $p_{11} - p_{10} - p_{01} > 0$ or $p_{11} - p_{10} - p_{01} + p_{00} > 0$, could be tested; without these monotonicity assumptions, the more stringent condition $p_{11} - p_{10} - p_{01} - p_{00} > 0$ would be tested. Other forms of interference corresponding to the outcome occurring for one person but not the other could be formed by simply redefining D_i as the relevant function of Y_{i1} and Y_{i2}. It should be noted that when one is testing for other forms of interference or other response patterns—such as whether both remain uninfected if and only if both receive the vaccine or whether at least one remains uninfected if and only if both receive the vaccine—the monotonicity assumptions that are considered will change because the definition of the outcome $D_i(a_{i1}, a_{i2})$ changes.

We have thus far been considering forms of interference analogous to those of epistatic/singular interactions, but similar results pertain to patterns of interference analogous to sufficient-cause interactions. For example, if we were interested in whether there are households such that the wife would be uninfected if both received the vaccine but would be infected if only one or the other of the spouses received the vaccine, then this would be analogous to a sufficient-cause interaction. We then could, for example, test for such a form of interference, without making monotonicity assumptions, using the condition for sufficient-cause interaction without monotonicity, namely, $p_{11} - p_{10} - p_{01} > 0$. Note that for these forms of interference corresponding to sufficient-cause interactions, conclusions are not being drawn regarding what occurs if both persons are unvaccinated as they were with the analogue of the epistatic/singular interactions. Further variations on these ideas are also possible insofar as we could draw conclusions about what sorts of outcomes might occur if and only if one person is vaccinated and the other unvaccinated by recoding exposure status and not simply outcomes. Such forms of interference would then be somewhat analogous to antagonism (VanderWeele and Knol, 2011b) in the context of causal interactions as described in Section 10.9.

The discussion to this point has assumed that we can distinguish between people in the household (e.g., the husband and the wife)—that is, that the subscript labelings $j = 1$ and $j = 2$ are meaningful. This may not always be the case; for example, we may have data only on the number vaccinated in each household and the number who have the outcome in each household. Alternatively, we may be considering siblings such that there is no clear classification of $j = 1$ and $j = 2$. Suppose once again that both exposures are randomized. We could then define D_i as, say, that both people have the outcome (or, alternatively, don't have the outcome; or that at least one has the outcome; etc.). If in each cluster we arbitrary select one person as $j = 1$ and the other as $j = 2$ (as will be seen below it will not matter which is which), then we could define $D_i(a_{i1}, a_{i2})$ as, for example, both persons having the outcome if we set $A_{i1} = a_{i1}$ and $A_{i2} = a_{i2}$ and let $p_{a_1 a_2} = P(D = 1|A_1 = a_1, A_2 = a_2)$. Now let A_i denote the number who received the exposure in cluster or household i (i.e. $A_i = 0$, 1 or 2) and let $p_a = P(D = 1|A = a)$; furthermore, note that if the exposures received by both people are randomized with the same probability of receiving the exposure, then $p_1 = P(D = 1|A = 1) = \frac{1}{2}P(D = 1|A_1 = 1, A_2 = 0) + \frac{1}{2}P(D = 1|A_1 = 0, A_2 = 1) = \frac{1}{2}(p_{10} + p_{01})$. By the arguments above, we

could test whether there are clusters where both persons would have the outcome if and only if both were exposed by testing $p_{11} - p_{10} - p_{01} - p_{00} > 0$, which, since $p_{11} = p_2, \frac{1}{2}(p_{10} + p_{01}) = p_1, p_{00} = p_0$, is equivalent to

$$p_2 - 2p_1 - p_0 > 0$$

Under the assumption that the exposure for at least one person in each cluster had a positive monotonic effect on the outcome D_i, we could, by similar arguments, instead test

$$p_2 - 2p_1 > 0$$

to conclude that this form of interference was present; and if the exposure for both persons in each cluster had a positive monotonic effect on the outcome D_i, we could test

$$p_2 - 2p_1 + p_0 > 0$$

As before, by recoding the outcome D_i or the exposures, we could also form tests for other forms of interference.

We note that these tests apply to persons 1 and 2 across households even in settings in which there are in fact more than two people per cluster (e.g., there may be varying number of children within each household), provided that the randomization of the exposure for persons $j = 1$ and $j = 2$ does not depend on the exposure status of the other people in the household. In fact, a similar approach can also be employed in settings in which more than two people are potentially randomized to the exposure or in which the exposure may have more than two levels (VanderWeele et al., 2012f). In fact, the entire theory of causal interaction for n-way interactions between exposures and also for multivalued exposures discussed in Chapters 9 and 10 maps onto tests for specific forms of interference (cf. VanderWeele et al., 2012f).

15.5.3. Illustration

Consider a hypothetical vaccine trial with two persons per household as described above, in which each of the two persons is randomized to receive the vaccine or a placebo with probability 1/2. Let $q_{uv}^{rs} = P(Y_{i1} = r, Y_{i2} = s | A_{i1} = u, A_{i2} = v)$ and suppose that the results from the vaccine trial, consisting of the infection status probabilities, q_{uv}^{rs}, are as given in Table 15.3 (cf. VanderWeele et al., 2012f). We will assume a very large trial and, for the purposes of the illustration, ignore sampling variability—that is, assume that the table represents the true infection-status probabilities.

Suppose we are interested in whether there are households such that both persons become infected if and only if neither were vaccinated. Without making any assumptions about monotonicity, by the arguments above, we could evaluate

$$q_{00}^{11} - q_{01}^{11} - q_{10}^{11} - q_{11}^{11} > 0$$

Table 15-3. PROBABILITIES OF INFECTION (Y_{i1}, Y_{i2}) BY VACCINATION STATUS (A_{i1}, A_{i2})

	$Y_{i1}=0,$ $Y_{i2}=0$	$Y_{i1}=0,$ $Y_{i2}=1$	$Y_{i1}=1,$ $Y_{i2}=0$	$Y_{i1}=1,$ $Y_{i2}=1$
$A_{i1}=0, A_{i2}=0$	0.69	0.10	0.09	0.12
$A_{i1}=0, A_{i2}=1$	0.81	0.05	0.09	0.05
$A_{i1}=1, A_{i2}=0$	0.83	0.10	0.03	0.04
$A_{i1}=1, A_{i2}=1$	0.96	0.02	0.01	0.01

In this case, we would have $q_{00}^{11} - q_{01}^{11} - q_{10}^{11} - q_{11}^{11} = 0.12 - 0.05 - 0.04 - 0.01 = 0.02 > 0$, which would indicate that there were households such that both persons became infected if and only if neither were vaccinated.

Suppose now instead we were interested in examining whether there are households such that neither person would be infected if and only if both were vaccinated. By the arguments above, we would be able to draw this conclusion without monotonicity assumptions if $q_{11}^{00} - q_{01}^{00} - q_{10}^{00} - q_{00}^{00} > 0$; or, under the assumption that for at least one person, administering the vaccine was never causative of infection for either person, if $q_{11}^{00} - q_{01}^{00} - q_{10}^{00} > 0$; or, under the assumption that for both people, administering the vaccine was never causative of infection for either, if $q_{11}^{00} - q_{01}^{00} - q_{10}^{00} + q_{00}^{00} > 0$. In this case, neither of the first two empirical conditions hold; however, for the third we have that $q_{11}^{00} - q_{01}^{00} - q_{10}^{00} + q_{00}^{00} = 0.96 - 0.81 - 0.83 + 0.69 = 0.01 > 0$. Thus if we were willing to assume that for both people, receiving the vaccine was never causative of infection for either, we could also draw the conclusion that there were households such that neither person would be infected if and only if both were vaccinated.

15.5.4. Interactions and Interference

The results here have generally been cast within the context of randomized exposures. However, all of the results are still applicable if randomization is conditional on strata of cluster covariates C_i and analysis is done conditional on C_i (or, somewhat more generally, if the effects of the exposure status for a cluster is unconfounded conditional on cluster covariates C_i and analysis is again done conditional on C_i). However, modeling difficulties arise in cases in which C_i contains numerous confounding variables, or continuous covariates, or distinct covariate values for each person in a cluster. The conceptual issues concerning confounding control become somewhat more subtle in the context of interference than in the simple no-interference setting as discussed later in this chapter.

In the econometrics literature the phenomenon of "interference" or "spillover effects" is sometimes referred to as "social interaction." This phenomenon arises in the study of neighborhoods, classrooms, judicial panels, and elsewhere, whenever people interact with one another in such a way that the outcome of one person depends on the exposure of another. The results of this section show that

such "social interactions" may in some instances also be cases of "mechanistic interaction" described in Chapters 9 and 10. The use of the word "interaction" to describe both the joint action of two exposures and also the phenomenon of interference is, however, not simply a linguistic coincidence, but instead, as discussed here, connotes an extensive formal correspondence between various forms of causal interaction on the one hand and specific forms of interference on the other. The theory of causal or mechanistic interactions turns out to provide a theoretical framework and analytical tool by which to reason about different forms of interference. See also Robins et al. (2014) for discussion of relations between interaction, interference and Bell's theorem in quantum mechanics.

15.6. INFERENTIAL CHALLENGES WITH MANY INDIVIDUALS PER CLUSTER

For the most part in this chapter we have considered the setting with just two individuals per household. As noted in previous sections, the methods we have described can be extended to some settings in which there are many individuals per household. The methods in Section 15.2–15.4 could, for example, be extended to settings with multiple individuals per household in which only one person in each household is randomized to exposure but outcome data are available on all individuals in the household or cluster. In this case, the methods described in these sections could be applied by taking an average (or some other summary measure) of the outcomes of all individuals except the first and using this as the outcome of the "second" individual in the methods. Alternatively, if all individuals are randomized to exposure, we could use the methods in Section 15.2-15.4 conditional on the exposure status of all but the first of the individuals. The methods of Section 15.5 concerning testing for specific forms of interference can also be extended to multiple individuals per cluster, and such extensions are described in the Appendix.

However, all of these extensions apply only to fairly special settings. When there are multiple individuals per cluster and all are randomized to exposure and we are interested in the average spillover and individual/direct effects, both the definition of these effects and inference for these effects becomes considerably more complicated. Suppose, for example, we were to compare the effects of a two-stage randomization scheme in which clusters were first randomized to, say, 50% or 30% vaccinated, and then within each cluster each individual was randomized to vaccine with one of these two probabilities. Building on work by Halloran and Struchiner (1995) and Sobel (2006), Hudgens and Halloran (2008) define individual spillover and direct effects by averaging these individual-level effects over all possible randomizations. For example, for the spillover or "indirect effect," we could hold an individual's exposure as fixed and then average over an individual's counterfactual outcomes across all the possible ways of vaccinating 50% of the remaining individuals versus 30% of the remaining individuals. Hudgens and Halloran (2008) refer to this as an "individual average indirect [i.e., spillover] effect." We could average this across all individuals in the cluster to "group average indirect effect," and we could further average this across all clusters to obtain a "population average indirect

effect." The individual/direct effect considered by Hudgens and Halloran (2008) assesses the effect on an individual of changing the exposure while holding the overall proportion of individuals vaccinated fixed in each cluster (e.g., at 30% or 50%; the individual/direct effect could vary according to this proportion). Again this could be averaged over all possible randomization allocations, individuals, and clusters. What Hudgens and Halloran (2008) define as a total effect considers changing the individual's exposure status and also the overall proportion in the cluster vaccinated (e.g., from 30% to 50%), and this too can be averaged over randomization allocations, individuals, and clusters. As in Section 15.2, this "total effect" decomposes into the sum of a spillover/indirect effect and an individual/direct effect. Hudgens and Halloran (2008) also define an "overall effect" that simply compares the overall proportion of those with the outcome with, for example, 50% versus 30% exposed. VanderWeele and Tchetgen Tchetgen (2011b) show that this "overall effect" can be decomposed into the sum of the spillover/indirect effect and a contrast of direct effects where the direct effect contrast comparing two different proportions (e.g., 50% versus 30%), is 0.5 times the direct effect with 50% vaccinated minus 0.3 times the direct effect with 30% vaccinated. See the Appendix for a formal statement of the result.

The formal definitions and notation become more complicated in this setting. Moreover, different randomization schemes give rise to subtle differences in definitions of causal effects: We could, for example, consider randomizing each individual to the exposure with 50% or we could randomly assign vaccine in each cluster so that exactly 50% had the vaccine, and these different randomizations give rise to somewhat different effect definitions (VanderWeele and Tchetgen Tchetgen, 2011b; Tchetgen Tchetgen and VanderWeele, 2012). Statistical inference and the construction of confidence intervals also becomes considerably more difficult when we allow for multiple individuals per cluster and arbitrary patterns of interference. Statistical inference becomes difficult because the averages described above are not taken over individuals who are independent of one another; this gives rise to the forms of dependence that no longer allow for simple appeal to the central limit theorem, which is what usually allows us to construct confidence intervals. Hudgens and Halloran (2008) provide unbiased estimators for population average direct, indirect, overall, and spillover effects, but they show that in this general setting there are no unbiased variance estimator for these effects. Under a further assumption that they call "stratified interference" that an individual's counterfactual outcome depends only on the collection of values of the treatments of other individuals and not who gets what treatment, they are able to give unbiased variance estimators; but because of dependence between individuals, they are not able to use these to derive confidence intervals even under the stratified interference assumption. Tchetgen Tchetgen and VanderWeele (2012) give conservative variance estimators without the stratified interference assumption and, for a binary outcome, derive finite sample confidence intervals for direct, indirect, overall, and total effects without the stratified interference assumption, but these confidence intervals are often overly conservative and very wide. They make no assumptions, however, on the form that the interference takes. Liu and Hudgens (2013) give narrower confidence but

only under much stronger assumptions. See Aronow and Samii (2013) for further discussion of statistical inference under more general interference structures.

Considerable work still needs to be done in this area to make the methods for causal inference under interference with many individuals per cluster easier to employ. Further complications of course arise with observational data when the exposures are not randomized. In general there is a trade-off between tractability of the analysis and the strength of the assumptions made. Most of the work on spillover effects with observational data has made, partially out of necessity, quite strong assumptions, and it is to such observational settings that we now turn.

15.7. SPILLOVER EFFECTS AND OBSERVATIONAL DATA

15.7.1. Addressing Interference Through Functional Form Restrictions

In this section consider a fairly general observational setting such as that considered in Hong and Raudenbush (2006) wherein individuals are clustered in groups such that individuals within groups may influence one another but there is no interference between groups, which was also referred to above as a partial interference assumption (Sobel, 2006; Hudgens and Halloran, 2008). We also make the stratified interference assumption above and further assume, following Hong and Raudenbush, that the potential outcome of each person depends on the exposure received by the other persons in the same cluster only through some known function g (e.g., the mean of the other exposures of the other individuals, or some other summary measure) so that the potential outcome for person j in cluster i could be written as $Y_{ij}\left(a,g\right)$, where a is the individual i's own exposure and g is the summary measure of all the other individuals. If $A_{i(j)}$ denotes the exposures of individuals in cluster i other than individual j, then we are assuming we can summarize the effect on individual j's outcome of other individuals' exposures, $A_{i(j)}$, by some function, $G = g(\mathbf{A}_{i(j)})$. The assumption that the effect of the exposures of all the other individuals can be summarized by a single measure is a fairly strong assumption but, as we will see, it simplifies the analysis considerably.

Suppose that for all i,j, the exposure A_{ij} is determined by unconditional randomization. We will then have that

$$\mathbb{E}[Y(a,g)] = \mathbb{E}[Y(a,g)|A = a, G = g] = \mathbb{E}[Y|A = a, G = g]$$

Hong and Raudenbush (2006) consider a variation on this assumption in the context of observational data. Let C_{ij} denote the covariates for individual j in cluster i. Hong and Raudenbush (2006) assume that

$$\mathbb{E}[Y(a,g)|A = a, G = g, C = c] = \mathbb{E}[Y(a,g)|C = c] \tag{A15.7}$$

and from this it follows that

$$\mathbb{E}[Y(a,g)|C = c] = \mathbb{E}[Y|A = a, G = g, C = c]$$

where the right-hand side can be estimated with observed data. Hong and Raudenbush (2006) also allow C_{ij} to contain cluster level covariates along with cluster aggregates of individual level covariates. Note, however, that (A15.7) requires that $Y_{ij}\left(a,g\right)$ be mean independent of both A_{ij} and $g(A_{i(j)})$ conditional on C_{ij}. If, for each individual, A_{ij} is randomized conditional on C_{ij}, although this will imply that $Y_{ij}\left(a,g\right)$ is mean independent of A_{ij} conditional on C_{ij}, it does not necessarily guarantee that $Y_{ij}\left(a,g\right)$ is mean independent of G conditional on C_{ij}. Let H_{ij} be some known function (possibly a vector) of all covariates C_{ij} for all individuals in cluster i other than individual j. We might, instead of (A15.7), consider

$$\mathbb{E}[Y(a,g)|A=a,G=g,C=c,H=h]=\mathbb{E}[Y(a,g)|C=c,H=h] \qquad \text{(A15.8)}$$

However, once again, with (A15.8), even if, for each individual, A_{ij} were randomized conditional on C,H, this would not guarantee that $Y_{ij}\left(a,g\right)$ is mean independent of G conditional on C,H unless H denoted the entire vector of all covariates for all individuals in cluster i other than individual j. Confounding control becomes more complex in settings with interference since when one individual's outcome is under consideration, control will often need to be made for the covariates of other individuals in the same cluster. See Ogburn and VanderWeele (2014a) for discussion of different causal structures and diagrams for which assumptions (A15.7) or (A15.8) will hold. If assumption (A15.8) did hold, we would have

$$\mathbb{E}[Y(a,g)|C=c,H=h]=\mathbb{E}[Y|A=a,G=g,C=c,H=h]$$

where again the right-hand side can be estimated with observed data. From this one could obtain conditional individual/direct, spillover/indirect, and total effects, namely,

$$\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y_{ij}(a^*,g)|c,h]$$
$$\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y_{ij}(a,g^*)|c,h]$$
$$\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y_{ij}(a^*,g^*)|c,h]$$

Marginal effects, involving counterfactuals of the form $\mathbb{E}[Y(a,g)]$, could be obtained by averaging over the distributions of C and H.

Example. Hong and Raudenbush (2006) considered interference in the context of the effect on reading scores of children of being retained in kindergarten versus being promoted to the first grade. Interference was assumed possible through the dependence of the potential outcomes of reading test scores of one child on whether other children were retained or not. Hong and Raudenbush were principally interested in the effect of a child's being retained and how this varied with being in schools with low retention and versus those with high retention. They used a sample of data from 1080 schools with 471 kindergarten retainees and 10,255 promoted students. In their application, students are clustered in schools. Individual treatment assignment was whether a student is retained (i.e., A_{ij} denotes whether a student is retained). They used a school-level scalar function based on

the proportion of the students that were retained to determine whether a school was a "high-retention" or "low-retention" school; that is, G_i indicates whether a school is "high-retention" $(G = 1)$ or "low-retention" $(G = 0)$. They assumed that interference was possible within schools but not across schools.

Using a propensity-score-based approach, accounting for interference, and assuming that assignment at both the school and the individual level was ignorable given a number of observed individual-level, school-level, and school-aggregated-individual level characteristics, Hong and Raudenbush (2006) obtained estimates of the effect on reading scores of retention. Specifically, in low-retention schools, the effect on reading scores of a student being retained versus being promoted was $\mathbb{E}[Y(1,0) - Y(0,0)] = -8.18$ (95% CI: $-10.02, -6.34$), and in high-retention schools the effect estimate was $\mathbb{E}[Y(1,1) - Y(0,1)] = -8.86$ (95% CI: $-11.56, -6.16$). A standard deviation in reading test scores in this sample is 13.48 points. The direct effects of retention in both high retention and low retention schools appear to be negative.

15.7.2. Sensitivity Analysis for Spillover Effects Under Unmeasured Confounding

We now consider a setting in which causal effects and spillover effects under interference are not identified due to unmeasured confounding. Although we can adjust for measured covariates as discussed above to attempt to control for such confounding, in an observational study, we can never be sure that the control is adequate. Moreover, as discussed above, confounding control becomes even more complex in settings with interference since when one individual's outcome is under consideration, control will often need to be made for the covariates of other individuals in the same cluster (Tchetgen Tchetgen and VanderWeele, 2012; Ogburn and VanderWeele, 2014a). Unmeasured confounding can thus operate either through the unmeasured covariates for the focal individual or for other individuals in the same cluster. Here we will present a simple the sensitivity analysis approach to assess unmeasured confounding in settings with interference and spillover effects.

Suppose now that we have unmeasured confounding by one or more unmeasured confounders U_{ij} and let V_{ij} denote some function (possibly the entire vector) of U_{ij} for all individuals in cluster i other than individual j. Suppose that conditional on observed C, H and unobserved U, V the effects of A and G on the outcome are unconfounded in the sense that

$$\begin{aligned}\mathbb{E}[Y(a,g)|A = a, G = g, C = c, H = h, U = u, V = v] \\ = \mathbb{E}[Y(a,g)|C = c, H = h, U = u, V = v]\end{aligned}$$

so that we could identify causal effect if we had data on U_{ij} for every individual. Without data on U_{ij}, causal effects are not identified. However, we can still employ sensitivity analysis. Let B denote the difference between the causal effect, $\mathbb{E}[Y(a,g)|c,h,u,v]$, and the biased estimand, $\mathbb{E}[Y|a,g,c,h,u,v]$. Very general expressions for the bias factor are given in the Appendix. Here we will consider

an easy-to-use approach under simplifying assumptions. If U_{ij} is a single unmeasured confounder and V is scalar and if the effects of U and V are additive in the sense that $\mathbb{E}(Y|a,g,c,h,u,v) - \mathbb{E}(Y|a,g,c,h,u',v') = \lambda(u-u') + \tau(v-v')$, then (VanderWeele et al., 2014c)

$$B = \lambda\{\mathbb{E}[U|a,g,c,h] - \mathbb{E}[U|a^*,g^*,c,h]\} + \tau\{\mathbb{E}[V|a,g,c,h] - \mathbb{E}[V|a^*,g^*,c,h]\}$$

To use the simplified bias formula, one only needs to specify the effect λ for a one-unit increase in the unmeasured confounder U_{ij}, the effect τ of a one-unit increase in the scalar function of the unmeasured confounders of the other members of the group, V, and how the means of U and V differ when $(A,G) = (a,g)$ versus when $(A,G) = (a^*,g^*)$. Once these sensitivity analysis parameters are specified, the bias factor B can be calculated using the formula above and then B could then be subtracted from the estimate of the causal effect using the observed data $\mathbb{E}[Y|a,g,c,h] - \mathbb{E}[Y|a^*,g^*,c,h]$ to obtain a corrected effect estimate for $\mathbb{E}[Y(a,g)|c,h] - \mathbb{E}[Y(a^*,g^*)|c,h]$. Under this simplified approach because the bias factor involves only the sensitivity analysis parameters and not the observed data, a corrected confidence interval could be obtained by subtracting B from both limits of a confidence interval for $\mathbb{E}[Y|a,g,c,h] - \mathbb{E}[Y|a^*,g^*,c,h]$. See VanderWeele et al. (2014b) for further discussion and for other sensitivity analysis techniques in the context of spillover effects under unmeasured confounding.

Example. Hong and Raudenbush (2006) used a similar expression to that above for the bias factor B to examine the extent to which their estimates, presented in the previous subsection, were robust to potential unmeasured confounding. Their initial estimates of the effect of retention on reading scores were -8.18 (95% CI: $-10.02, -6.34$) in low-retention schools and -8.86 (95% CI: $-11.56, -6.16$) in high-retention schools. They noted that the strongest predictor of current test scores were lagged test scores, but that it was unlikely that there was any unmeasured covariate that would predict their outcomes so strongly. They considered instead whether unmeasured individual and school covariates that had effects on readings scores that were equal to those of the measured covariate with second strongest association with reading scores would suffice to explain away the effect estimates. Using an argument based on a formula similar to that given for the bias factor B above, they report that unmeasured individual and school confounders that had an effect as large as the second most important measured individual and school level covariates would shift the estimate in high-retention schools to $\mathbb{E}[Y(1,1) - Y(0,1)] = -4.25$ (95% CI: $-6.95, -1.54$) and thus not suffice to bring the confidence interval to include 0. However, in low-retention schools, unmeasured individual and school confounders that had an effect as large as the second most important measured covariate would shift the estimate in low-retention schools to $\mathbb{E}[Y(1,0) - Y(0,0)] = -0.60$ (95% CI: $-2.44, 1.24$) and thus would suffice to bring their confidence interval for the effect in low-retention schools below 0. The effects in high-retention schools seem more robust to the possibility of unmeasured confounding.

15.8. SPILLOVER EFFECTS AND SOCIAL NETWORKS

15.8.1. Challenges with Social Network Data

We have assumed throughout that individuals in one cluster do not influence individuals in another. In some settings this may be plausible. In other settings, social relations may be much more complex and there may not be complete independence. Individuals may be related to one another through a series of ties that might be mapped by a social network. Lack of independence creates challenges for statistical inference. The network setting also gives rise to certain conceptual challenges. Within a network setting, interest is often not simply in how the effect of an intervention on one person may affect the outcomes of another, but also how the states or outcomes of an individual propagate through the network, like the type of "contagion effect" discussed in Section 15.4.

Causal inference for social network data is still arguably in its infancy. Within the context of social networks, different forms of causation may be at play. One form is social influence (also called peer effects, relational effects, induction, contagion, or network effects) whereby the behavior, states, and characteristics of one individual in a network may influence behaviors, states, and characteristics of others in the network with whom the first individual shares some form of social tie, directly or indirectly. However, within the context of social networks, another form of causation that may be present is that of network formation (also referred to as friendship formation/selection, homophily, or effects on networks) whereby the behavior, states, and characteristics of various individuals may exert influence on whether and which social ties are present to begin with. The complex interplay between these different forms of causation is in part what makes inferring causation with social network data especially challenging.

More generally if the behavior, states, and characteristics of two individuals with a social tie are found to be correlated, it is possible to envision at least three potential explanations. First, it is possible that the association is due to social influence: One of the persons may have influenced the other, or vice versa, or both. Second, it is possible that the behavior or states are correlated because persons with similar characteristics are more likely to become friends with one another. This phenomenon is sometimes referred to as homophily or selection. Third, it is possible that there is some shared environmental factor that influences the states or behaviors of both individuals so that they are correlated. With cross-sectional data it is essentially impossible to distinguish between these three explanations. Manski (1993) referred to this as the "reflection problem." However, when it is possible to conduct randomized experiments (randomizing either social ties or particular interventions), or when longitudinal data are available, some progress can be made in distinguishing influence, homophily, and environmental confounding. In the remainder of this section, we will give a brief overview of some of the ideas and approaches that have been proposed for causal inference with social network data. A fuller overview can be found in VanderWeele and An (2013). We will begin by discussing different randomized experiments that can be used to draw causal inferences on social networks. We will then devote the following two subsections to an approach to causal

inference with social network data that involves using longitudinal regression analyses; this approach has received considerable attention and has also been the subject of some controversy and so we will consider the various objections and responses that have been put forward in the literature concerning this approach. We will then turn to an additional stochastic actor-oriented approach to observational data and conclude this section with discussion on trying to move toward more formal causal inferences with social network data.

15.8.2. Randomized Experiments Using Social Networks

Some progress can be made in assessing causality in social network settings by using randomized experiments. Experiments might randomize either a specific intervention or social ties. As an example of an experiment in which social contacts were randomly assigned to subjects, Sacerdote (2001) found that when roommates and dormmates were randomly assigned, the academic achievement of a student had significant impact on the academic performance and social activities of other students in the same room. The random assignment of roommates addresses the homophily/selection problem. However, even if we can randomly assign social contacts to subjects, finding a significant association among outcomes does necessarily allow the conclusion of peer influence, because there still is the possibility that the correlation is driven by some contextual factor (environmental confounding) that affects both the subjects and their assigned contacts. For example, even if college roommates are randomly assigned and we find that there is a significant correlation in their academic performance, the correlation may be at least partly generated by their common local circumstances (for example, their shared living conditions—e.g., living in a dorm on a quiet versus a noisy street). One potential solution to this problem is to use lagged outcomes of the randomly assigned contacts to predict the focal subject's outcomes. Indeed any characteristic or covariate that is available prior to randomization could be used for this purpose. If a pre-randomization characteristic is used and associations with peer outcomes are assessed, then homophily/selection is eliminated by randomization, and environmental confounding is effectively eliminated by using characteristics that occurred before either individual was in their shared environment. For example, Kremer and Levy (2008) used student drinking behavior in the year before entering college to predict their roommates' drinking behaviors at a large state university where roommates were randomly assigned and found significant peer effects on drinking.

Other randomized experiments have been employed that randomize not social contact but rather particular interventions as in Sections 15.2–15.6 above, and they make use of the network structure in the design of these interventions. For example, if individuals are partitioned into clusters, and the intervention is to be given to only some portion of individuals in each cluster, the network structure might be used to target the intervention to individuals thought to be influential or central in the network (Valente and Davis, 1999; An, 2011). Alternatively, groups or cliques within a network could be identified using social network data, and the intervention could be targeted to entire cliques rather than having the treated individuals in each

cluster being determined randomly (Wing and Jeffrey, 1999; Valente et al., 2003; An, 2011). See VanderWeele and An (2013) for further discussion. Consider a trial in which (i) some clusters receive no treatment, (ii) some clusters have random treatment assignment to individual, (iii) some clusters have treatment targeted to central individuals, and (iv) some clusters have treatment targeted to particular groups or cliques. If the clusters which are in categories (i)–(iv) above is determined randomly, several different types of causal effects and comparisons can be examined by exploiting such random assignment (An, 2011; VanderWeele and An, 2013). For example, average outcomes in each of the three types of clusters receiving treatment could be compared to those without treatment to assess the overall effects of the different treatment strategies. Further, the average outcomes of untreated subjects in the clusters receiving treatment randomly could be compared to the clusters without treatment to estimate indirect or spillover effects. The average outcomes of treated subjects in the clusters receiving treatment randomly could also be compared to the clusters without treatment to estimate what were called, in Section 15.2, "total effects." The average outcomes in clusters receiving treatment randomly could also be compared to the clusters in which the central individuals or groups of individuals receive treatment to assess whether targeting central individuals or groups of individuals matters. Such comparisons could also potentially be done examining only the treated or the untreated individuals respectively, but this would then not be a causal effect in the traditional sense since, for example, the central individuals may not be comparable to individuals in clusters selected randomly. Such "effects" could be interpreted as the joint effects of the treatment and of selecting particular individuals. Matching methods could potentially be used to attempt to only compare individuals in the various arms who are similar to each other. See An (2011) and VanderWeele and An (2013) for further discussion.

15.8.3. Causal Inference from Observational Social Network Data Using Longitudinal Regression

Although randomized experiments with social network data are becoming increasingly common, much of the available social network data arise from observational contexts. Causal inference from observational social network data in which data are available on only a single social network over time has been more controversial. As noted above, with cross-sectional social network data, when outcomes of peers are associated, it is essentially impossible to distinguish whether this is due to social influence, homophily, or environmental confounding. When longitudinal data are available, some progress can be made but the approaches that have been proposed are subject to numerous limitations.

One prominent approach for observational social network data was used in a series of analyses conducted by Christakis and Fowler and colleagues (e.g., Christakis and Fowler, 2007, 2008; Fowler and Christakis, 2008; Cacioppo et al., 2009; Christakis and Fowler, 2012) claiming that social influence plays an important role in the spread of a variety of health-related attributes, behaviors, and psychological

states including obesity, smoking, happiness, depression, drug use, and even loneliness. The approach essentially consists of regressing one individual's (the ego's) state (e.g., obesity) on another's (the alter's) state, along with the alter's lagged state, the ego's lagged state, and other covariates for the ego. Significant association between the ego's state and the alter's state when also controlling for the ego's and alter's lagged state and other variables is then taken as evidence for a contagion effect.

Suppose individual i names individual k as a friend. Let $Y_i(t)$ and $Y_i(t+1)$ denote the ego's outcome at times t and $t+1$, respectively. Let $Y_k(t)$ and $Y_k(t+1)$ denote the alter's outcome at times t and $t+1$, respectively. Let $X_i(t+1)$ denote the ego's covariates at time t. One might then regress $Y_i(t+1)$ on $Y_i(t), Y_k(t), Y_k(t+1)$ and $X_i(t+1)$ using either repeated measures logistic regression for binary outcomes or repeated measures linear regression for continuous outcomes. The coefficient for $Y_k(t+1)$ in the regression model for $Y_i(t+1)$, which we will call here β, is often taken as a contagion effect (i.e., a measure of social influence), and robust standard errors are typically computed using generalized estimating equations.

Christakis and Fowler (2007) argue that, adjusting for the alter's lagged status (e.g., lagged obesity if obesity is the outcome) helps to control for homophily (Christakis and Fowler, 2007; Carrington et al., 2005). The reasoning is that the latent factor giving rise to homophily would have to explain both the ego's obesity and the alter's obesity via pathways other than through the alter's lagged obesity for such a factor to generate an association in the absence of genuine social influence. Even if we grant adequate control for homophily, interpreting associations, even with longitudinal social network data, as evidence for contagion effects is potentially problematic because of the possibility that a shared environmental factor might in fact affect both the ego's and the alter's state. Christakis and Fowler argue against this as an explanation by noting that the effect estimates for ego-nominated friends are larger than those for alter-nominated friends, which would not occur if the associations were purely due to environmental confounding. Similar analyses for other types of social ties beyond ego-nominated friends, and alter-nominated friends could also be considered such as with mutual friends (person i names person k as a friend and person k names person i as a friend) and with spouses, neighbors, and siblings.

Using this approach, Christakis and Fowler report evidence for social influence for smoking, obesity, alcohol consumption, happiness, loneliness, depression, drug use, and so forth (Christakis and Fowler, 2007, 2008; Fowler and Christakis 2008; Cacioppo et al., 2009; Christakis and Fowler, 2012). Thus, for example, using data from the Framingham Heart Study (Dawber, 1980; Feinlib et al., 1975), Christakis and Fowler (2007) found that an individual's chances of being obese (body mass index > 30) increased by 57% (95% CI: 6–123%) if he or she had a friend who was obese in a given interval. In these analyses they controlled for an ego's age, sex, and education level, the ego's obesity status at the previous time point, and the alter's obesity status at the previous time point. Likewise, using the same data, Christakis and Fowler (2008) report that smoking cessation by a spouse decreased a person's chances of smoking by 67% (95% CI: 59% to 73%).

15.8.4. Objections to Longitudinal Regression for Observational Social Network Data

The analyses of Christakis and Fowler have received substantial attention in the academic literature and in the media, but have come under considerable criticism. The critiques have included incorrect estimation of standard errors (Lyons, 2011), allegedly similar results using the same methodology for factors such as height, acne, and headaches for which social influence seems much less plausible (Cohen-Cole and Fletcher, 2008), inadequate control for homophily (Shalizi and Thomas, 2011), changes in friendship structure giving rise to spurious associations (Noel and Nyhan, 2011), and issues with model inconsistency (Lyons, 2011). Because their longitudinal regression approach has been used often in the literature and is considered quite controversial, we will consider the objections and responses that have been put forward at some length.

More specifically, Shalizi and Thomas (2011) argue that the possibility of latent (unmeasured) homophily threatens the validity of such longitudinal network analyses. As noted above, homophily refers to the tendency of individuals similar to one another to become friends with each other. It may, for example, be the case that two friends simultaneously become obese not because one influences the other but because they both enjoy excessive eating; this shared interest causes them to become friends and also causes them both to become obese over time. When control is not made for variables responsible for homophily in the analysis, it is difficult to attribute the association to social influence rather than homophily (Shalizi and Thomas, 2011). Although control for an alter's lagged obesity, as in Christakis and Fowler (2007), arguably does help somewhat, Shalizi and Thomas (2011) argue that the problem of latent homophily is still present in such analyses. If the latent factor giving rise to the formation of friendship ties affects present obesity even when controlling for past obesity, associations between the ego's and alter's current obesity can arise even when the alter has no social influence on the ego. Shalizi and Thomas (2011) also leverage this point to further critique the argument that Christakis and Fowler use against environmental confounding. Christakis and Fowler argue against environmental confounding as an explanation of their associations by noting that the effect estimates for ego-perceived friends are larger than those for alter-perceived friends, which would not occur if the associations were purely due to environmental confounding. Shalizi and Thomas (2011) show that in the presence of latent homophily, even if there is no unmeasured environmental confounding, the associations comparing ego-perceived friends and alter-perceived friends may differ in magnitude even when there is no social influence. The basic reasoning used by Christakis and Fowler (2007) to argue against environmental confounding in this case breaks down and we are then left with both latent homophily and environmental confounding as possible explanations of associations.

A somewhat related critique was also put forward by Noel and Nyhan (2011) concerning friendship retention. Through simulations, Noel and Nyhan (2011) show that if friends whose characteristics change to become different from one another are also more likely to end the friendship, then this can also lead to bias

and could explain away associations between an ego's and an alter's states. Following the phenomenon on Facebook, they refer to this as the "unfriending" problem. They critique the type of longitudinal social network analysis undertaken by Christakis and Fowler (2007, 2008) on the grounds that such "unfriending" can give rise to spurious associations of the form reported by Christakis and Fowler, even in the absence of social influence.

Yet another important set of critiques was put forward by Lyons (2011). Lyons (2011) argues that when using the repeated measures logistic regression model for binary outcomes, as used by Christakis and Fowler, if, in the network, there is a person i with a tie to person k and that person k has a tie to person $m \neq i$, then, when using contemporaneous ego and alter data, the models themselves imply that the coefficient for social influence equal zero (i.e. $\beta = 0$); similar issues also pertain to linear models. The models themselves effectively contradict the existence of the very effect that Christakis and Fowler want to assess. The issue raised by Lyons is essentially that there are more equations than unknowns. Intuitively, the problem develops because the same variable at the same time period—for example, the ego's state at time $t + 1$—is the dependent variable in one regression and the independent variable in another regression. Lyons argues that the models themselves then effectively contradict the conjecture of social influence that Christakis and Fowler want to assess. If there is in fact social influence, then the models themselves are incorrectly specified.

Lyons (2011) also criticizes the procedures Christakis and Fowler (2007, 2008) use for statistical inference in face of the complex statistical dependence structures that are generated by a social network. Christakis and Fowler (2007, 2008) use a method referred to as generalized estimating equations, clustering on the ego, to take into account the use of multiple time points for the ego. Unfortunately, as Lyons (2011) notes, this is not the only source of dependence in the data. If there is social influence (contagion), then the clusters defined by the ego will not be independent of one another. Moreover, even under the null of no contagion, when contemporaneous ego–alter data are used, the generalized estimating equations standard error is not always valid. In fact, it can be shown that, because Christakis and Fowler (2007, 2008) use contemporaneous data for the ego and the alter, and because one person's state at time $t + 1$ is thus both an outcome in one regression and an independent variable in another, the standard errors for β obtained by Christakis and Fowler (2007, 2008) are too small whenever relationships are reciprocal (e.g., for mutual friends, spouses, siblings, and neighbors; see VanderWeele et al., 2012c).

This array of critiques has shed considerable doubt on the validity of the analyses undertaken by Christakis and Fowler. Although some of these critiques carry substantial weight, some progress has been made in responding to or at least partially circumventing some of these critiques. For now, let us set aside the issues of homophily and environmental confounding, to which we will return later, and suppose that adequate control has been made for these. As noted above, Lyons argued that the models themselves effectively contradict the conjecture of social influence that Christakis and Fowler want to assess. However, an important exception arises when the null hypothesis of no contagion is in fact true. In this case, provided that

homophily and environmental confounding have been properly controlled for, then $\beta = 0$; and, if $\beta = 0$, then the models may be correctly specified, provided, for example, for a binary outcome, that the log odds of the ego's state is indeed linear in the covariates. Under the null hypothesis of no contagion, the problem of model inconsistency effectively vanishes. The estimate and confidence interval for β would not constitute a valid estimate of the contagion effect. However, whether the confidence interval for β contains 0 would constitute a valid test of the null hypothesis of no contagion, again provided the assumptions of no homophily and no environmental confounding conditional on the covariates and that of correct model specification with respect to the covariates held (VanderWeele et al., 2012c). Under these assumptions, we can in theory do testing, but not estimation.

This brings us to another critique of Lyons (2011), that of statistical modeling under the dependence structures that are generated by a social network. Even under the null of no contagion, when contemporaneous ego–alter data are used, the generalized estimating equations standard error is not always valid. Because Christakis and Fowler (2007, 2008) use contemporaneous data for the ego and the alter, the standard errors for β obtained by Christakis and Fowler (2007, 2008) are anti-conservative and the confidence intervals will be too narrow whenever relationships are reciprocal—for example, for mutual friends, spouses, siblings, and neighbors (VanderWeele et al., 2012c). However, for the purposes of testing, both the problem of model inconsistency and the problems of statistical dependence and standard error estimation can be easily addressed if the alter's state is lagged by an additional period in the regressions (VanderWeele et al., 2012c). The argument used by Lyons (2011) to show that the models are inconsistent in the presence of contagion is no longer applicable and, under the null of no contagion/social influence, the clusters defined by the ego are independent of one another leading to valid standard errors when using generalized estimating equations (VanderWeele et al. 2012c).

In fact, Christakis and Fowler (2007, 2008) report, in the online supplement to their papers, that they ran such analyses in which the alter's state was lagged by an additional period and that the results of such analyses were similar to those of their main analyses using contemporaneous data for the ego and alter (i.e., they once again find evidence of significant contagion effects for smoking and obesity).

All of our discussion thus far has assumed that adequate control has been made for homophily and environmental confounding. This assumption is very strong. To partially circumvent this issue, sensitivity analysis, similar to that discussed in Chapter 3, could be used to assess the extent to which an unmeasured factor responsible for homophily or environmental confounding would have to be related to both the ego's and the alter's state in order to substantially alter qualitative and quantitative conclusions (VanderWeele, 2011h). The sensitivity analysis technique is applicable to estimates obtained by lagging the alter's state by an additional period (VanderWeele, 2011h; VanderWeele et al. 2012c). Using the results of Christakis and Fowler (2007, 2008; cf. Fowler and Christakis, 2008; Cacioppo et al., 2009), which are reportedly similar to what is obtained from lagged analyses, VanderWeele (2011h) used such sensitivity analysis techniques to argue that the evidence reported by Christakis and Fowler (2007, 2008) for obesity amongst

mutual friends and for smoking cessation amongst spouses was reasonably robust to potential latent homophily or environmental confounding; associations between other types of relational ties for smoking and obesity and those for happiness and loneliness were considerably less robust. The associations reported by Cohen-Cole and Fletcher (2008) concerning acne and headaches were not at all robust to potential latent homophily and environmental confounding.

The final critique of those discussed above that has not yet been considered is that of the "unfriending" problem (Noel and Nyhan, 2011). Noel and Nyhan argued that unfriending results in changes to social network structures that can lead to spurious associations between egos and alters even in the absence of social influence. However, the simulations of Noel and Nyhan also suggest that the degree of these potential biases depends largely on the extent of "unfriending." In the adolescent AddHealth data mentioned above, friendship retention across waves is only about 50%; and in Noel and Nyhan's simulations, such low retention rates can indeed generate substantial bias. However, in the Framingham Heart Study data used by Christakis and Fowler (2007, 2008), friendship retention is very high and this unfriending problem does not seem, by Noel and Nyhan's own simulations, sufficiently common to result in substantial biases in the analyses of Christakis and Fowler (2007, 2008).

Considerable methodological development still needs to be done concerning such longitudinal social network analyses, perhaps especially in deriving valid estimators of the standard error that are applicable not only under the null hypothesis of no contagion but also in the presence of social influence. However, a number of the existing critiques of previous longitudinal network analysis have at least partially been addressed and the methods that have been used can perhaps at least in some cases give tentative evidence for contagion (for obesity among mutual friends and smoking cessation between spouses) on a social network.

15.8.5. A Stochastic Actor-Oriented Model for Observational Social Network Data

The approach employed by Christakis and Fowler (2007, 2008) makes use of well-established methods for longitudinal data and attempts to control for confounding and homophily by covariate control and the use of lagged states, but it does not explicitly model the mechanism for the selection of social ties itself. An alternative stochastic actor-oriented model has been developed (Snijders, 2001, 2005; Steglich et al., 2010) that models social influence and the selection of social ties jointly.

The model assumes that at each instant an individual may either change the behavior/state under study or change a particular social tie. Such changes occur on the network with specific rates, which may vary across individuals. In the models, such as in Steglich et al. (2010), these events are assumed to follow an exponential distribution. The rates may depend on the states of the individuals in the network, the existing network structure itself, individual actor-level covariates, and dyad-level covariates. Likewise, when an event occurs, the actual changes to either the behavior state or a social tie may depend on the states of the individuals on the network, the existing network structure itself, individual actor-level covariates, and dyad-level

covariates and are also subject to random fluctuation. The magnitude of the effect of each of these components is fit with data.

Such stochastic actor-oriented models are typically too complex to allow for closed-form expressions for the probabilities of particular transitions and also too complex to employ traditional maximum likelihood procedures. The models are instead fit with simulation techniques such as Markov Chain Monte Carlo. Fitting such models can be computationally demanding, which can limit the sample size to which the models can be employed. The fitting procedures of such models can also sometimes fail to converge.

These stochastic actor-oriented models are appealing in that they involve parameters corresponding to both social influence and homophily. They do, however, rely on stronger modeling assumptions. Modeling assumptions need to be made not only regarding the behavior states themselves, as in the Christakis–Fowler approach, but also for processes by which there are changes in social ties. The assumption that there is only one single change allowed at any particular instant might be problematic in cases where subjects make simultaneous changes in both behavior and social ties. For example, in the case of smoking, it is likely that some subjects sever their ties to smokers while stopping smoking at the same time. It is unclear how these models perform in the context of simultaneous changes. Another strong assumption made by the stochastic actor-oriented model is that actors have full information of their local networks. Yet another strong assumptions made by the stochastic actor-oriented model is that actors' decisions to make changes in either behavior or social ties are not reactive (i.e., not taking into account other actors' potential reactions) and also that the co-evolution of network and behavior follows a Markov process in which only the immediate past states matter. This assumption greatly simplifies the mathematics, but would be in tension with social relationships and social behaviors that may have longer-lasting impact on future network and behavior change. Finally, the stochastic actor-oriented models are also subject to the same limitations concerning environmental confounding as the longitudinal models of Christakis and Fowler. This issue of environmental confounding needs to be critically investigated and evaluated when using stochastic actor-oriented models.

15.8.6. Prospects and Challenges

Considerable work remains to be done in providing a more rigorous foundation for causal inference from observational social network data. As was noted above, in the longitudinal social network analyses currently being employed, the estimation procedures used for statistical inference are often valid only for testing, not estimation. Further work is required in developing variance estimators that are applicable in the presence of social influence and when statistical dependence may be present between the states of all individuals within a social network. As noted above, even in a simpler setting with well-defined clusters, formal statistical inference in the presence of social influence can be challenging when there are more

than two individuals per cluster, and these issues are further complicated within the context of a social network.

At a more conceptual level, further theoretical development remains to be done in attempting to formulate longitudinal social network analyses within a counterfactual framework. Within the counterfactual or potential outcomes framework, causation is generally conceived of in terms of counterfactual contrasts, and the counterfactuals are themselves generally tied to hypothetical interventions. Within the context of studying the possibility of social influence in a social network of a state, such as obesity, say, it is not entirely clear how to appropriately tie discussion of causation to such hypothetical interventions or what such hypothetical interventions might be. Moreover, different possible interventions (e.g., exercise, diet, or liposuction) may have different effects in terms of influencing other individuals within the network (cf. Hernán and VanderWeele, 2011). One of the key advantages of experimental studies of interventions on social networks is that it is clear what it is that the causal effects estimated actually correspond to. With observational social network data concerning a particular trait, state, or characteristic, the "exposure" of interest is often not well-defined and does not necessarily clearly correspond to a particular intervention. Such issues pertain to both the longitudinal models of Christakis and Fowler (2007, 2008) and the stochastic actor-oriented models of Snijders (2001, 2005; Steglich et al., 2010). Work could also be done formalizing the confounding/selection assumptions required to give the parameters of these models a causal interpretation.

Work also remains to be done in further explicating the relationship between the phenomenon referred to here as "interference," on the one hand, and that which is often called "contagion," on the other. The former term is generally used for settings in which the exposure of one individual may influence the outcomes of another, whereas "contagion" is typically used for the phenomenon whereby the outcome or state of one individual influences the same outcome or state of another individual. Contagion may be one mechanism by which interference occurs as discussed in Section 15.4 in the infectious disease context. The issue applies more generally, however. Consider a study of obesity in which a particular weight loss intervention is assigned to certain persons within a network. The intervention may affect other persons to whom the intervention was not explicitly assigned in at least two distinct ways. First, information from the weight loss intervention may be passed from one individual to another, leading to weight loss even among those to whom the intervention was not explicitly assigned. Second, the intervention may lead to weight loss for those to whom it was explicitly assigned, and such weight loss may influence the norms of other persons to whom the intervention was not assigned, motivating them also to lose weight. The second mechanism might be conceived of as one of "contagion" whereas the first might be conceived of as one of "direct interference" (i.e., "direct" with respect to, not through, the obesity state of the person assigned the intervention). Further work remains to be done in better explicating and formalizing the relationship between these concepts when data are available on an entire social network. The literature on social networks has increased dramatically over the past several years, but considerable work remains to be done in formalizing causal inference on such social networks.

15.9. DISCUSSION

In this chapter we have discussed the final topic of this book: social interaction and spillover effects. We have seen that although causal inference is more challenging when the exposure of one person can affect the outcome of others, progress at least in certain settings is still possible and certain analyses concerning spillover effects can be reasonably straightforward to conduct. We have seen that a number of the methods from previous chapters including methods for assessing mechanistic interaction, methods for direct and indirect effects, and methods for principal stratification have essentially direct application to the spillover effect setting. Nonetheless, in more general cases, the analysis of spillover effects and statistical inference for spillover effects can be quite challenging and considerable methodological development is still needed in this area, especially with regard to methodological approaches that can be easily applied. We saw that assumptions about the form that interference takes can simplify the analysis quite considerably. In general, progress with observational data will require such simplifying assumptions. The assumptions that we employed with observational data were that the spillover effect of the exposure of all other individuals in a cluster on a focal individual depended only on some simple summary measure of the other individuals' exposures. Yet further complications arise when interest lies in causal inference from observational social network data. Although a few approaches have been employed, these make very strong assumptions and have been highly controversial. The study of social interactions will likely be of interest in many settings in which outcomes arise through human interactions. Some progress can be made, but analysis is challenging, and further methodological work remains to be done.

16

Mediation and Interaction: Future and Context

The concluding chapter of this book is divided into two sections. In the first section, we will consider the current state of methodology for questions of mediation and interaction, what types of inferences and analyses are and are not possible, areas of inquiry in which special care must be taken, and how future methodological development might address certain deficiencies in the methods currently available. In the second section of this chapter we will once again take a broader view, as was done in Chapter 1, of issues of causation and explanation and will situate some of the discussion in this book within the philosophical discourse on these issues.

16.1. THE PRESENT STATE OF METHODS AND FUTURE METHODOLOGICAL DEVELOPMENT

This book has covered a broad range of concepts and methods for assessing mediation and interaction. Methodology for assessing these phenomena has expanded dramatically over the past 10 years. New developments are likely to continue to rapidly emerge for some time to come. In the concluding sections of the various chapters of this book we discussed some of the open questions and needs in methodological development. In this section we will review and highlight some of these open areas of inquiry and will attempt to synthesize what we can and cannot learn from the current state of methodology available and where further development is most critical.

Chapter 2 provided a summary of concepts for mediation analysis and regression-based methods to estimate direct and indirect effects. One of the topics discussed in the chapter was how study design was related to questions of mediation and mechanisms and whether control for past values of exposure, mediator,

and outcome was important. Further theoretical, simulation-based, and empirical inquiry into these questions would be valuable. Too often mediation analyses have been undertaken with cross-sectional data. Longitudinal data helps, to a certain extent, to circumvent some of the issues of feedback and reverse causation that may arise in cross-sectional studies. However, how best to design and analyze studies in which mediation and mechanisms are of interest is still largely an open question, one that we touched upon again in Chapters 3 and 6, but one that is largely still unaddressed.

Chapter 3 discussed sensitivity analysis. As was discussed in both Chapter 2 and Chapter 3, analyses assessing mediation make strong assumptions, assumptions that are often considerably stronger than those used in assessing overall causation. Many of these assumptions have been ignored in practice in empirical research within the biomedical and social sciences. As a result, numerous of the empirical studies that have undertaken to assess mediation are likely wrong, some possibly quite severely. Even understanding the assumptions about confounding and knowing when and how they might be violated can help ensure more accurate inferences. However, as discussed in Chapter 3, sensitivity analysis can also be very useful in assessing how strongly the assumptions would have to be violated in order to undermine the qualitative conclusions being drawn. A number of methods for assessing the sensitivity of results both to unmeasured confounding and to measurement error were provided. Considerable methodological work remains to be done still, however, especially in developing sensitivity analysis methods for unmeasured confounding for natural direct and indirect effects which have easy-to-interpret parameters in complex settings allowing for interaction. Perhaps the greatest danger currently faced when mediation analysis is being employed empirically in actual practice is the present discrepancy between what software packages will currently allow in terms of estimation on the one hand versus sensitivity analysis on the other. The SAS, SPSS, and Stata macros described in Chapter 2 (Valeri and VanderWeele, 2013) currently allow for estimation in a number of settings, but have not yet automated any sensitivity analysis procedures for unmeasured confounding. The simulation-based R and Stata commands described in Chapter 2 (Imai et al., 2010a, b) have some sensitivity analysis features for unmeasured confounding built in but allow for estimation of direct and indirect effects in far more settings than in which they allow for sensitivity analysis. None of these software packages yet provides any sort of sensitivity analysis for measurement error. In some settings it is possible to carry out simple sensitivity analysis for unmeasured confounding and measurement error by hand, but in other settings this is considerably more difficult, and, moreover, in a number of settings the techniques have not yet even been developed. Even when techniques become available, until they are built into easy-to-use software packages, macros, and commands, their use is likely to remain very limited. From the perspective of ensuring correct qualitative inferences about mediation and indirect and direct effects, the development and software implementation of easy-to-use sensitivity analysis techniques is perhaps of the highest priority for methodological development concerning mediation.

Chapters 4 and 5 discussed mediation analysis with time-to-event outcomes and with multiple mediators, respectively. In just the last few years, considerable

progress has been made in both areas and a number of methods are now available. In certain settings, the mathematics rendered analytic formulae for direct and indirect effects difficult or impossible to obtain, however. In some cases, it was possible to address this limitation by using various weighting approaches to confounding control, rather than using regression adjustment. However, as was discussed in these chapters and also further in Chapter 7, these weighting approaches would often also result in loss of efficiency of the estimators. Further work needs to be done in understanding when this loss of efficiency occurs, what if anything can be done about this, when the weighting estimators are to be avoided, and whether other more efficient (possibly multiply robust) estimators can be employed and easily implemented. Methodological research also needs to be done for settings in which the mediator itself is a time-to-event variable.

In Chapter 6 we discussed mediation analysis with time-varying exposures and mediators. Controlled direct effects were relatively straightforward to estimate in this setting. However, natural direct and indirect effects—useful in assessing the extent of mediation—were considerably more challenging. A fairly general nonparametric result was given, but a great deal of work still needs to be done on developing methods for this setting that can be implemented in a straightforward way with data. As we discussed in that chapter, some of these developments may come from taking methods for mediation that have been proposed in the social science literature and reformulating them within the counterfactual framework, extending their applicability and clarifying the underlying assumptions. As noted above, longitudinal designs will often be desirable and even necessary for assessing mediation, and developing adequate methods to assess mediation with longitudinal data with time-varying exposures and mediators will be a critical task in the years ahead.

Chapter 7 covered a range of selected topics in mediation analysis. Of those discussed, the most pressing need is arguably the development of power and sample-size calculations for direct and indirect effects. Most of the literature on this topic is restricted to simulations and often very simple cases. Development of analytic power and sample-size formulae for standard settings with binary, continuous, count, and time-to-event outcomes and mediators will be of considerable importance and should be a priority in the years ahead. Understanding the role of exposure–mediator interaction in determining the power to detect direct and indirect effects may also be important. Power and sample-size calculations are essential in planning studies and in ensuring that resources are adequate to test relevant hypotheses of interest. The lack of such tools in the current literature is a major limitation.

As noted in Chapter 8, each of the topics in that chapter could quite possibly have given rise to a book of its own, and so we cannot discuss all the various developments that might be possible or desirable in each of these areas. From the perspective of mediation, what is perhaps most crucial, and what was emphasized in this book, is clarifying how each of these topics—principal stratification, surrogates, instrumental variables, and Mendelian randomization—differs from mediation analysis. Each of these topics is important. Each attempts to answer a

different set of questions. Each uses different analytic tools. Understanding these differences is essential. There are of course some similarities in the concepts and approaches and analytic tools as well. This has, at times, led to conflation of the concepts. It will be important in the years ahead to make clear within the various empirical research communities when the different concepts and analytic tools are to be employed, what different questions they answer, and how the concepts are related and differ.

Chapter 9 gave a broad overview of concepts of interaction and of interaction analysis. Much of this material is well established. One area that was touched upon in that chapter where there have been a number of recent important developments (but for which considerably more work remains to be done) is methods to identify subgroups to target treatments using multiple covariates. Detecting treatment effect heterogeneity has been of interest to researchers in the biomedical and social sciences for decades and methods to detect such heterogeneity across levels of a single covariate are well established and relatively straightforward. With multiple covariates, strata often become too small to have adequate power to detect such heterogeneity. However, more recently, methods propose combining covariates into a single regression-based predicted-treatment-effect score that can be used as a single covariate for identifying subgroups. While the approaches that have been proposed are very promising, statistical inference, avoiding overfitting, and ensuring adequate outof-sample performance and model robustness are challenging problems. Some progress has been made at this in the past few years, but considerable work remains to be done. The potential payoff of such methodological development, however, is vast, as questions about such effect heterogeneity are pervasive across disciplines. With increasing policy emphasis on the individual tailoring of treatment, from personalized medicine to personalized education, further methodological contributions in this area could prove very important indeed.

Chapter 10 discussed what sorts of inferences were possible concerning different mechanistic forms of interaction. In the past decade it has been shown that inferences about these mechanistic forms of interaction were in fact possible to draw empirically, from data, in a wide range of contexts that had hitherto been thought impossible. Theoretical results, empirical tests, and analytical tools have been developed for a number of different settings. While some methodological development remains to be done in this area, perhaps a bigger question for the future of methods on this topic is whether, and when, we actually learn something that is of use from a policy or scientific perspective when we detect such mechanistic interaction. There are now a number of studies that have identified the presence of mechanistic interaction between different exposures for various outcomes. But it is not yet clear what use has, or even might be, made of conclusions of this form. Clarifying when these conclusions of mechanistic interaction are of interest, and why, may be an important task for methodological inquiry in this area.

Chapter 11 discussed bias analysis for interaction and provided tools to assess how unmeasured confounding and measurement error might bias interaction analyses and when such analyses were robust. While the techniques for unmeasured

confounding in interaction analyses extended to fairly general settings, the current results and tools for measurement error were much more limited. A couple of robustness results, for interaction estimates and for tests, to measurement error were given, but only very limited techniques for sensitivity analysis were described. Much work thus remains to be done in this area. Moreover, the results for measurement error also assumed that the distributions of the exposures were independent. As was described in that chapter, interaction analyses are robust to certain forms of measurement error, but understanding exactly when and how to correct such analyses when they are not robust will be important topics for methodological development in the years ahead.

Chapter 12 concerned issues of independence of exposures, boosting power, and multiple testing in interaction analyses, issues that arise frequently in the genetics literature. Much of this material is now well established and used routinely in genetics research. An interesting question concerning this methodology is the extent to which it can and should be used outside of genetics. For example, while the case-only estimator for interaction is now used fairly frequently in the genetics literature, it could potentially be employed in numerous other fields as well (whenever the distributions of the two exposures are independent). The estimator is powerful and easy to implement. It would be interesting to see whether other fields might benefit from its use, or from the use of joint tests, or whether other disciplines should make more use of correction for multiple testing when examining interactions. There may be lessons for other fields to learn from the genetics literature in these areas.

Chapter 13 described power and sample-size calculations for interaction analyses. Tools and formulae for number of different settings were given. While this material is relatively straightforward and the potential for use is quite general, an interesting methodological question concerns whether, when, and in what applied contexts we really do have adequate power to hope to detect interaction. Such concerns may be especially important in the social sciences. When units are schools or neighborhoods, often the sample sizes available in any given study are quite small. It was clear from the discussion in Chapter 13, however, that to have good power to detect interaction, sample sizes need to be quite large. This discrepancy raises the question, in settings with small sample sizes, as to when tests for effect heterogeneity are found to be statistically significant, whether this is often likely to be a false positive. Further reflection ought to be given to whether it is really possible to achieve sample sizes adequate to attain reasonable power to detect effect heterogeneity. It may be possible that in most settings this is feasible at, say, the child level but not at the school level. Further thought concerning these issues may be important in study design and resource allocation.

Chapter 14 presented a framework and a decomposition to unite many of the concepts of mediation and interaction discussed in this book. The fourfold decomposition given in that chapter allows an investigator to assess how much of an effect is due to just mediation, to just interaction, to both mediation and interaction, and to neither mediation nor interaction. Further reflection could be given to in what settings this fourfold decomposition contributes over and above the methods that just assess mediation or just assess interaction. Further methodological work could

be devoted to determining which, and in what settings, each of the four components of the decomposition are robust to unmeasured confounding and measurement error; further work could also be devoted to sensitivity analysis techniques for each of the four components when biases are present.

Chapter 15 presented a number of developments on spillover effects, interference, social interaction, and social networks in which one person's exposure or outcome could affect those of others. Inference in these settings is challenging, and many of the methodological developments in this area have come only very recently. Accordingly, much work remains to be done in this area. Further methodological research is needed in settings with multiple individuals per cluster, in settings with observational data, in valid statistical inference for spillover effects, and in extending various results and approaches to social networks. While notable progress has been made in this area over the last 10 years, much more still remains.

The previous chapters of this books have described a broad range of the methods that are currently available for assessing mediation and interaction. As can be seen from the present discussion, considerable methodological development remains to be done and the array of methods may look quite different 10 years from now than it does today. There are many open questions and areas of inquiry that require more exploration and development. As this occurs, applied empirical researchers will hopefully be better equipped with tools to accurately assess the phenomena of mediation and interaction.

16.2. PHILOSOPHICAL QUESTIONS

In Chapter 1 we discussed the relationship between causation and explanation with a focus on what could be learned empirically from data—that is, the logic behind how we go about making causal inferences from observational data. To conclude this book, we will here, in this section, return to some of these same questions concerning our understanding of causation and explanation but with a focus that is more conceptual and philosophical.

16.2.1. Causation, Explanation, Natural Laws, and Hume

This book has been concerned with causation and, more specifically, with issues of explanation and mechanism in the context of causal phenomena. As discussed in Chapter 1, although it is difficult to give a complete characterization of what is meant by causation or of the conditions under which the proposition "X caused Y (in this particular instance)" is considered to be true, it is easier to give sufficient conditions for this, and also to even empirically reason about propositions of the form "X causes Y (in general, i.e. in at least some instances)." We have, in this book, taken the perspective that if an outcome were to occur if the exposure were set to be present, but would not occur if the exposure were set to be absent, then, in this case, we would say that the exposure caused the outcome. This condition may not be necessary for causation, but, as discussed in Chapter 1, it is generally considered to be sufficient for attributing causation.

But what is causation itself? This was a question the eighteenth-century philosopher David Hume considered. Hume sought to evaluate our intuitive notions of causation and whether these notions had any empirical justification. Hume claimed that causation is simply a relation between experiences. He argued that it is not empirically verifiable that the cause produces an effect, but only that the experienced event called the cause is invariably followed by the experienced event called the effect. There is nothing guaranteeing that the next time the cause is observed, the effect will be also. However, from our experience, we find ourselves expecting it to be. We feel that there is some sort of necessary connection between the cause and the effect, and we project our feelings of necessity onto the objects themselves.

Hume (1739) suggested that we usually ascribe a causal relationship when three conditions are met: (1) spatial and temporal contiguity—the cause and the effect are present at the same point in time and space; (2) temporal succession—the presence of the cause precedes the effect; and (3) constant conjunction—whenever we observe the cause, we also observe the effect. However, Hume argued that the conjunction observed between cause and effect is not logically necessary. It is a contingent fact that a particular effect follows its cause. There is no logical contradiction in observing the cause without the effect. However, it is because we in fact do always observe the cause followed by the effect that we ascribe a causal relation between the two events.

Hume's account is sometimes criticized in that it does not distinguish between accidental generalization and genuine causation (cf. Psillos, 2002). We may have a series of objects or events that satisfy Hume's three criteria simply by accidental coincidence. For example, perhaps every time a person walks into the city hall, the city hall bell rings; but it may be the case that the person enters every day precisely at noon, which is also precisely when the bells begin to ring. The person's walking into the city hall and the bell's ringing may satisfy Hume's criteria of spatial and temporal contiguity, temporal succession, and constant conjunction, but we would not say that the person's walking into the city hall is the cause of the bell's ringing. Or consider another similar such series: Perhaps every time a specific person parks below a certain acorn tree, an acorn falls on the car immediately prior to the car's moving upon its departure. There may only ever be three such instances in which the car is ever parked below the tree; but if in all three instances an acorn does, by chance, fall upon the car, then this would seemingly satisfy Hume's three criteria, under which, Hume suggests, we would attribute causation of the car's moving to the acorn. Examples such as this, involving accidental generalizations, seem to demonstrate that Hume's criteria of spatial and temporal contiguity, temporal succession, and constant conjunction are not sufficient to ascribe a causal relation.

Hume's position has been revised and restated in attempt to distinguish between causation and accidental generalization. Entailment by the so-called "laws of nature" is sometimes used as the criterion distinguishing between causation and accidental generalizations (Mill, 1911; Ramsey, 1928; Lewis, 1973; Psillos, 2002). The laws of nature might be understood as those statements that describe the regularities that we observe amongst physical phenomena and that we would continue to observe even in contrary-to-fact scenarios. Science catalogues these laws

of nature. It does so by hypothesis and experimentation; it seeks a minimal set of regularities with maximum explanatory power.[1]

Causation might then be seen as a relation between events in accordance with the laws of nature. We might say "X had an effect on Y" (which we again take as a sufficient condition for causation) when (i) the event X occurred, (ii) the event X, the state of the universe, and the laws of nature jointly entail the event Y, and (iii) the absence of event X, the state of the universe, and laws of nature jointly would entail the absence of event Y. Under this conceptualization, causation is a derivative concept to the laws of nature. Causation is defined in reference to these laws. Indeed this derivative nature of our causal concepts is reflected, to some extent, in the concepts employed by physics. Physics makes reference to various laws of nature. Physics does not typically make reference to notions of causation. If we conceive of causation as entailment by the state of the universe (in the presence versus absence of some event) and the laws of nature, then physics, in describing the laws of nature and their implications, does not need to rely on causal concepts. Causal concepts are derivative; they are not as fundamental.

Hume's position of there being no actual "necessary connection" when we attribute a causal relationship might then be restated as the position that there is nothing guaranteeing that these so-called laws of nature will continue to hold (cf. Psillos, 2009). There is no reason why the laws of physics must hold. From all of our past observation they do seem to hold; but that they do at each moment is a contingent fact.

Suppose we were to grant this adaptation of Hume's position; does this then threaten our counterfactual approach? The counterfactual framework is essentially agnostic to Hume's characterization of causation and to the status of the laws of nature. The counterfactual framework does not offer a characterization of causation that circumvents Hume's position, nor is it threatened by Hume's account or its more modern adaptations. The counterfactual framework does not require a "necessary connection" to be present between cause and effect; it does not posit that the laws of nature must continue to always hold. If effect estimates are extrapolated to other or future contexts it would be presupposing these laws of nature do not shift but the framework itself is not definitively tied to a particular notion of causation. It is simply a framework for thinking about the outcomes that might have occurred under two or more different actions. The counterfactual framework does, however, require some account of counterfactual statements themselves, a point to which we will return later.

16.2.2. Human Action and Noncausal Forms of Explanation

Our focus in this book has been on *causal* explanation. However, there are of course other forms of explanation that do not make reference to causes. In Chapter 1, for example, we noted that logical or mathematical explanations were noncausal forms

1. Note that a distinction might be drawn between the actual laws and our knowledge of them. Regularities that are proposed, tested extensively, and not falsified are provisionally accepted within the scienfitic community as true. If further evidence later emerges contradicting such laws, they are then rejected or modified so as to cohere with the existing evidence.

of explanation; and there are other forms of noncausal explanation as well. In what is now considered a fairly classic classification of "causes," Aristotle, in his *Physics* and again in his *Metaphysics*, distinguished between what he viewed as four different types of causes: material causes, formal causes, efficient causes, and final causes. Aristotle described the material cause as that out of which the object is made; the formal cause as that into which the object is made; the efficient cause as that which makes the object; and the final cause that for which the object is made. Aristotle gives as an example a bronze statue. The material cause (that out of which the object is made) is bronze; the formal cause (that into which the object is made) might be a solider—the statue may take the form of a soldier; the efficient cause (that which makes the object) is the sculptor; and the final cause (that for which the object is made) might be to serve as a memorial for those who died at war. Each of Aristotle's "causes" offers some form of explanation or answers a specific question: Out of what?... Into what...? By whom or what...? For what purpose...?

This book and the causal inference literature in statistics, epidemiology, and the social sciences focus on what Aristotle called "efficient causes." Science in general focuses on efficient causes and perhaps, to a certain extent, material and formal causes. However, we only really use "cause" today to refer to efficient causes and perhaps sometimes final causes. We generally wouldn't today say that the "cause of the statue is bronze" or the "cause of the statue is a solider." We do, however, still sometimes use "cause" in reference to final causes. We might say "he died for a good cause" or "I am doing this for the cause of the uninsured ... or for the cause of science."

Our use of "cause" in these cases has led some philosophers to question whether final causes might be analyzed in terms of efficient causes. Davidson (1963, 1980), for example, has suggested that a desire and a belief taken together cause human action. For example, if it is the case that "I want the light on" and it is also the case that "I believe that flipping the switch will turn on the light," then I might well proceed by flipping the switch, upon which the light goes on. Perhaps then this belief and desire taken together cause me to flip the switch. It is indeed the case that both my belief and my desire precede my action.

However, a closer look at our language suggests that the analysis of human action in terms of efficient causes may be deficient (cf. Hacker, 1996; Kenny, 1976; Wittgenstein, 1953). Consider the two statements: "He raised his arm" and "His arm went up." This is the same event, but the two statements consist of different descriptions of this one event. If we ask "Why did his arm go up?," we might answer and explain this by various physiological processes—that is, his arm went up because certain neurons in the brain were firing or certain muscles were contracting. If we ask "Why did he raise his arm?", we would respond by explaining by his intent; he raised his arm in order to turn on the light or in order to answer a question. If we responded to the question "Why did he raise his arm?" with "He raised his arm because certain muscles were contracting," we might receive the reply "No ... no ... that is not what I meant." If we said "Certain muscle contractions caused him to raise his arm," we might think this was happening involuntarily; the response might be, "You mean he didn't intentionally raise his arm?"

In general, when we are explaining human action, such action is (at least in conversation) typically explained by reasons and intent and not by causes (in the efficient sense).

More generally, how an event is described determines, in part, how it can be explained. Consider two statements: "She is playing the piano" and "She is practicing for a concert." These are two descriptions of the same event. We can explain the statement "She is playing the piano" by offering the further statement that "She is practicing for a concert." However, we cannot explain the statement "She is practicing for a concert" by saying "She is playing the piano." Even though these are the same events, they are events under different descriptions, and only one explains the other. What counts as an explanation for an event depends on how the event is described.

Similarly, whether an event has an efficient cause depends, in part, on how it is described. The statement "Her fingers hit the piano keys" can be explained physiologically; it can be explained in terms of efficient causes. However, what is described in "She is practicing for a concert" is human action and arguably the event described as such has no efficient cause, though it might have a final cause. For example, we might be able to explain her practicing for a concert by "She wants to win the competition." In general, when an event is described as human action, everyday language does not typically allow for explanations in terms of efficient causes, but instead only in terms of final causes. The language we use seems to suggest that human actions, described as such, have no efficient cause. Human actions can be explained, but, in our ordinary language, such human actions are not "caused." We might ask whether this is an idiosyncrasy of the language we use or whether it points to the nature of human action itself .... Can human action always be redescribed so as to have an efficient cause? Is there something beyond mere physiology giving rise to our actions? Do humans have, in some sense, freewill in such a way that the state of the universe in conjunction with the laws of nature do not wholly determine human actions?

16.2.3. Counterfactuals, Interventions, and Human Agency

Our discussion of final causes and human action in relation to efficient causes also brings us to an interesting question concerning the relationships between human agency, causation, and counterfactuals. In this book, as well as within the causal inference framework that has come to dominate in statistics, epidemiology, and the social sciences, causation is typically conceived of in terms of contrasts of counterfactual outcomes. These counterfactual outcomes are themselves typically conceived of as the outcomes under hypothetical interventions; and the hypothetical interventions that give rise to counterfactuals usually consist of some human action; for example, a person takes drug A versus drug B.

We might however, then, further ask if, by relating the meaning of a counterfactual to human action, we have assumed that it is possible for humans to act in ways other than they did. If humans are not free (for example, if determinism is true, so that a complete description of the world entailed all that followed), then humans

could not act other than they did unless something else were other than it was. If this is so, then it is not entirely clear that we can reasonably relate the meaning of counterfactuals to the idea of human agency. If there is no human freedom in a sense that contradicts determinism, we arguably ought to abandon our distinction between counterfactuals involving human intervention and other counterfactuals. There would seem to be no basis for the distinction. That we generally feel more comfortable with counterfactuals if we can propose a hypothetical intervention may again suggest that there is perhaps something special about human action. It is difficult for us to reason in ways that do not take the idea of human freedom seriously. Relating the meaning of counterfactuals to human agency only makes sense if human action is in some way qualitatively different from other events.

If we were to abandon the position of trying to relate the meaning of counterfactuals to human freedom, we would be left with the question as to whether counterfactual statements have any meaning at all. David Lewis, who is often credited with having re-introduced the contemporary philosophical discourse on counterfactual conceptualizations of causation, provided an analysis of counterfactuals in terms of possible worlds (Lewis, 1973). He argued that the statement "If X had not happened, then Y would not have happened" is to be considered true if, in each possible world W1 in which X did not happen and Y did happen, there is another possible world W2 in which X did not happen and Y did not happen that is closer to the actual world than W1. The theory requires the construct of alternative "possible worlds" and some metric to assess closeness. Sometimes closeness is perhaps obvious; the possible world in which one person's middle name is spelled "Jon" rather than "John" is arguably closer to the actual world than in the possible world in which, is the year 1800, a meteorite hit the earth, extinguishing all human life. But other cases are not so clear. Is the possible world in which a meteorite hit the earth, extinguishing all human life in 1800, closer to the actual world than the same event occurring in 1801?

Lewis' account requires the existence of possible worlds and some metric by which to judge distance from the actual world. The question of whether the conception of possible worlds is reasonable is controversial. What exactly is meant by a possible world, and what does or does not qualify? Likewise, formally defining a metric to compare a possible world, whatever is meant by them is problematic and difficult. A fair bit of the philosophical debate on Lewis' description of counterfactual statements has centered around these issues of the meaning and status of possible world and metrics by which to compare them.

Alternatively, returning to the account of causation given in terms of the laws of nature as described in Section 16.2.1, perhaps the ambiguity of counterfactual statements arises from failure to specify precisely what statement it is we are in fact making. In Section 16.2.1 we noted that one possible account of the statement "X had an effect on Y" was that the statement is true if (i) the event X occurred, (ii) the event X, the state of the universe, and the laws of nature jointly entail the event Y, and (iii) the absence of event X, the state of the universe, and laws of nature jointly would entail the absence of event Y. Along the same lines, we might similarly suggest that the statement "If X had not happened, then Y would not have happened" is to be considered true if, in the absence of event X, the state of the

universe and laws of nature jointly would entail the absence of event Y. In practice, when we make a statement such as "If X had not happened, then Y would not have happened" we cannot, and do not, fully specify the state of the universe beyond X. When we make such a statement, much of what we take as "the state of the universe other than the event X" is presumed. We assume that, for the most part, the state of the universe (other than the event X) is similar to what it at present is. However, what precisely also would have to be different if the event X were to not have occurred will, at least sometimes, be unclear. There may be several different sets of circumstances—different states of the universe—in which the event X would be absent, and these different states may have different implications for whether the event Y would occur.

The counterfactual statement "If X had not happened, then Y would not have happened" is thus ambiguous. It is ambiguous to the extent that (i) there are multiple states of the universe consistent with the absence of event X, (ii) it is neither the case that the laws of nature entail the presence of Y in all of these states, nor the case that the laws of nature entail the absence of Y in all of these states,[2] and (iii) the speaker has not clarified which state of the universe is in view at least to the extent that would suffice to fix entailment of Y, or its absence, by the laws of nature. In some cases, although there may be multiple states of the universe consistent with the absence of event X, it is possible that all of these would entail either the presence or the absence of the event Y, and then the statement is no longer ambiguous. In other cases, there may be multiple states of the universe consistent with the absence of event X, and it may be the case that the laws of nature entail neither the presence of Y nor the absence of Y in all of these states; but if the speaker further specifies the states of the universe that are in view sufficiently so that the laws of nature and the state of the universe thus specified do entail either the presence or the absence of Y, then, once again, the counterfactual statement "If X had not happened, then Y would not have happened" is no longer ambiguous. We can thus, to a certain extent, clarify the ambiguity of this counterfactual statement by further specifying the state or states of the universe in view, at least to the extent necessary for an analysis of a particular event. Said another way, the consequence of the event X not occurring may depend on other aspects of the state of the universe; there may be heterogeneity in the event Y across these different states—the states in some sense modify implications of the absence of event X. For the counterfactual statement to be unambiguous, the speaker must sufficiently specify the state of the universe in view so that that heterogeneity is eliminated.

In summary, we might thus consider conceptualizing counterfactual statements in terms of entailment, of an event or its absence, by the state of the universe and the laws of nature, much as we did with our sufficient conditions for causation (cf. VanderWeele and Hernán, 2012). The ambiguity of counterfactual statements again then potentially arises from failure to sufficiently specify the state of the universe

2. If the laws of nature entailed the presence of Y in all of these states, then the counterfactual "If X had not happened, then Y would not have happened" would be false. If the laws of nature entailed the absence of Y in all of these state, then the counterfactual "If X had not happened, then Y would not have happened" would be true.

that is in view. If we return to the question as to why we are sometimes more comfortable with counterfactual statements grounded in human action being otherwise than it was, it may be the case that human actions, at least on the surface of things, seem sufficiently free that we have an easier time imagining only one specific action being different, and nothing else. With regard to what else would have had to be different if a human action had been different, there seems to be less ambiguity than there is if some other aspect of the state of the universe had been different. It is easier to imagine the rest of the universe being just as it is if a patient took pill A rather than pill B than it is trying to imagine what else in the universe would have had to be different if the temperature yesterday had been 30 degrees rather than 40.

16.2.4. Chains of Causation and Arguments for a First Cause

In Part I (on mediation) of this book we were essentially interested in chains of causation. We were interested in settings in which an exposure affected a mediator, which in turn affected an outcome, and the extent to which the effect of an exposure on an outcome was due to the exposure's effect on the mediator. Such chains of causation are typically used in our reasoning to understand mechanisms, informally in everyday life and more quantitatively as in this book. Such chains of causation will, in most cases, be relative to the scope of explanation in which we are interested. We might seek to understand a mechanism between a particular exposure and outcome; that is, we might seek to the importance of a mediator in this relationship. However, once we have established that a specific mediator is important, we might sometimes further question what the mechanisms are for the relationship between the exposure and the mediator we have identified. The analytic complexities of such settings were discussed to some extent in Chapter 5. We might seek more and more detailed levels of explanation.

Chains of causation and explanation can of course run in the other as well. We may have established that an exposure has an effect on an outcome, and we might then want to seek what the cause of the exposure is. Why did the exposure itself come about? We might then further ask, What was the cause of the cause? And such questioning can potentially continue regressively on and on. Tracing such chains of causation backwards brings us to another set of philosophical debates that have been important and influential historically, debates over what is sometimes called an argument for a "first cause." The form of argument is also sometimes called the "cosmological argument for the existence of God." The argument concerns the implications of causation itself. The argument appeared in ancient Greek thought in Plato and Aristotle, among others; the argument was stated in its most popular form by the philosopher and theologian Thomas Aquinas; and in the last couple of decades there has been renewed interested, in the philosophical literature, in this so-called cosmological argument and its modern variants. We will begin with Aquinas' form of the argument, present some of the debates concerning the argument during the enlightenment in the correspondence between Clarke and Leibniz, and conclude with some of the contemporary debate on the status of the argument. See also the Rowe (1997) and Reichenbach (2013) for further related discussion.

In his work *Summa Theologica*, Aquinas offers five arguments for the existence of God. The first three of these arguments are all closely related forms of the argument for a first cause. All three concern the possibility of an infinite regress. The first concerns changes, the second concerns causes, and the third concerns entities. Thus, for example, his first argument concerning changes runs roughly as follows: (1) Certain things change; (2) every change is brought about by a change in something else; (3) there cannot be an infinite regress of changes. Therefore, there must have been something that brought about the first change. That which brought about the first change is what Aquinas conceived of as God. The argument essentially denies the possibility of an infinite regress of changes, and this is where it is potentially open to objection. It is not clear why we cannot have an infinite regress of changes. Not only might we be able to conceive of an infinite regress of changes, but we can potentially even conceive of an infinite regress of changes in a finite time, with each previous change occurring temporally half way between the subsequent one and the time origin.

Aquinas' second and third arguments follow a similar form. His second argument concerns causes rather than changes: (1) Every event must have an (efficient) cause; (2) every cause must thus have been brought about by a previous cause; (3) there cannot be an infinite regress of causes. Therefore, there must have been a first cause; and we might identify this first cause as God. As before, objection might be made to the impossibility of an infinite regress of causes. Additionally, with causation, the position that every event has a cause is sometimes viewed as controversial. We considered the debate over whether human action has an efficient cause; likewise within modern physics it is not clear if certain quantum events have causes.

Aquinas' third argument involves existence and can be formulated as follows: (1) Every being that exists or ever did exist is either a dependent being or a self-existent being; (2) not every being can be a dependent being whose existence was caused by another being because we cannot have an infinite regress of being. Therefore there exists a self-existent being, which again we might identify as God. As before, with the third argument also, we might question whether an infinite regress of being is in fact impossible. It, however, be more difficult to conceive of an infinite regress of being—of matter—than of changes.

These questions and this argument was picked up again in a correspondence Leibniz in Germany and Clarke in England (Clarke and Leibniz, 1717). The correspondence between Leibniz and Clarke takes the argument a step further, however. In it, the question is posed that, even if we grant the possibility of an infinite regress of being, we might still ask, What explains the fact that there are dependent beings at all? Debate over this question is related to what is sometimes called the "Principle of Sufficient Reason." This Principle of Sufficient Reason can be stated as the claim that: "No contingent fact can be true unless there be a sufficient reason why it is so" (cf. Leibniz, 1714; Clarke and Leibniz, 1717). A contingent fact is one that may have been either true or false. If this Principle of Sufficient Reason is true, then, if every being were dependent, there would be a fact about the existence of things without a sufficient reason, namely that there are dependent

beings. From this it could be argued that there once again must be a self-existent being.

If the Principle of Sufficient Reason, stated as such, is true, then the argument for a self-existent being perhaps does appear to follow. However, whether the Principle of Sufficient Reason is true is itself an interesting and difficult question, and some of the contemporary philosophical literature has taken up this question. Some of the literature objects to this Principle of Sufficient Reason on the grounds that it is not clear what evidence there is in support of it or why one should consent to it. Other philosophical literature has attempted to come up with counterexamples to this Principle of Sufficient Reason. There have also been attempts to weaken the Principle of Sufficient Reason so that the conclusion of the cosmological argument still holds, but so that, with the weaker version, it is more difficult to disagree with the principle itself and also more difficult to come up with counterexamples against the principle.[3] There continues to be philosophical debate both over whether these arguments are successful and whether the premises are reasonable (Oppy, 2000; Davey and Clifton, 2001; Gale and Pruss, 2002; Reichenbach, 2013).

The arguments concerning a first cause or a self-existent being do, to a certain extent, have analogues in physics as well. The argument is sometimes made that if everything that begins to exist has a cause of its existence, and if the universe began to exist (as the Big Bang Theory would suggest), then the universe itself must have a cause of its existence. The cause of the universe's existence is sometimes again identified as God. This specific form of argument concerning the beginning of the universe is sometimes referred to as the "kalam cosmological argument" (Craig and Smith, 1993). Both premises of the argument have been contested. Phenomena in quantum physics are sometimes thought to constitute an exception to the premise that everything that begins to exist has a cause of its existence. And although modern physics suggests an origin of the universe with the "Big Bang" such that the universe began to exist about 13–14 billion years ago, some have posited the possibility of an oscillating universe (Musser, 2004). There is contemporary debate in the physics community whether our current understanding of physics is compatible with an infinite series of expanding and contracting universes, with at least some evidence, due to constraints on entropy, seeming to suggest not (Baum and Frampton, 2007; cf. Craig and Sinclair, 2009) and that the universe did have a beginning. Further discussion of these debates is given in Craig and Smith (1993), Craig and Sinclair (2009), and Reichenbach (2013).

On a more intuitive level, a lot of the debate on these points seems to come down to the question, Why is there something rather than nothing? Why does anything exist at all? To some, it may seem more reasonable to assume that the existence of the universe has some cause or explanation, which we might call "God," than that the universe is without cause or explanation. The position that God is self-explanatory

3. One such weakening of the Principle of Sufficient Reason, for example, can be stated as follows: No contingent fact *concerning the existence of things* can be true unless there be a sufficient reason why it is so. Another, weaker statement of the Principle of Sufficient Reason can be stated as: For every true contingent proposition, it is possible that there is an explanation for that proposition (Gale and Pruss, 1999).

may seem more reasonable than the position that the universe is self-explanatory. These are questions that have been debated for hundreds, even thousands, of years. They are questions that challenge the limits of human inquiry and reason. They are questions of explanation that are among the most difficult. And for me, they are also questions of explanation that are among the most important and that point to what is of most importance.

APPENDIX

Technical Details and Proofs

A.1. EXPLANATION AND MECHANISM

In this section we will provide a very brief introduction to the "potential outcomes" or "counterfactual" framework. Consider the setting in which we are interested in assessing the effect of some exposure A on an outcome Y. Within the "counterfactual" or "potential outcomes" framework (Neyman, 1923; Rubin, 1974; Robins, 1986; Pearl, 1995) we let Y_a denote the outcome that would have occurred had A been set to a. If the exposure is binary, then there are two such potential outcomes, Y_0 and Y_1. For each individual, we only get to observe the outcome corresponding to the exposure level that actually occurred for that individual. If the individual actually had exposure level $A = 0$, then we observe Y_0 and we will not in general know the value of Y_1 for that individual. If the individual actually had exposure level $A = 1$, then we observe Y_1 and we will not in general know the value of Y_0 for that individual. That we do in fact observe one of potential outcomes for each individual is sometimes referred to as the "consistency assumption"; stated formally, it is that for an individual who has actual exposure $A = a$ the actual outcome Y is equal to Y_a. Although we only get to observe one of the potential outcomes for each individual, at least hypothetically we could conceive of both potential outcomes. We could then define the causal effect for an individual as the difference between the two potential outcomes: $Y_1 - Y_0$. We would say that the exposure had an effect on the outcome for that individual if the difference, $Y_1 - Y_0$, were nonzero.

Because we generally know only one of the two potential outcomes for each individual, we cannot in general calculate the individual level causal effect. However, we might hope to be able to estimate that effect on average for a population, $\mathbb{E}[Y_1 - Y_0]$. Even to be able to do this, we need to assume that the potential outcomes are comparable across the exposure groups. Often this will be an implausible assumption and so instead we assume that at least within strata of some set of measured covariates C, the different exposure groups are comparable in their potential outcomes. This assumption is sometimes referred to as an "exchangeability" assumption, an "ignorable treatment assignment" assumption, a "no-unmeasured-confounding" assumption, or an "exogeneity" assumption.

More formally, we will use the notation $A \perp\!\!\!\perp B|C$ to denote that A is independent of B conditional on C. The no-unmeasured-confounding assumption

(again also called "ignorability," "exchangeability," or "exogeneity") can be stated as $Y_a \perp\!\!\!\perp A|C$—that is, that within strata of the covariates C the potential outcomes Y_a are independent of the exposure level actually received. If this were the case, the groups that actually had exposure level $A = 1$ and $A = 0$ would be comparable in their potential outcomes Y_0—that is, in terms of what would have occurred had exposure been set to level 0. Similarly, we would have that the groups that actually had exposure level $A = 1$ and $A = 0$ would be comparable in their potential outcomes Y_1—that is, in terms of what would have occurred had exposure been level 1. If this no-unmeasured-confounding assumption holds, then we can obtain average causal effects from the observed data within strata of covariates since

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|c] &= \mathbb{E}[Y_1|c] - \mathbb{E}[Y_0|c] \\ &= \mathbb{E}[Y_1|A = 1, c] - \mathbb{E}[Y_0|A = 0, c] \\ &= \mathbb{E}[Y|A = 1, c] - \mathbb{E}[Y|A = 0, c]\end{aligned}$$

where the second equality follows by the no-unmeasured-confounding assumption that $Y_a \perp\!\!\!\perp A|C$ and the third equality follow from the consistency assumption. Whereas the quantity $\mathbb{E}[Y_1 - Y_0|c]$ is a contrast of potential outcomes, the final expression in the display equation above, $\mathbb{E}[Y|A = 1, c] - \mathbb{E}[Y|A = 0, c]$, is a quantity we can estimate from the observed data. Under the no-unmeasured-confounding assumption, these two quantities are equal; we can estimate the average causal effect within strata of covariates using just the observed data: $\mathbb{E}[Y_1 - Y_0|c] = \mathbb{E}[Y|A = 1, c] - \mathbb{E}[Y|A = 0, c]$. This is the causal effect within stratum of covariates $C = c$. If we wanted to estimate the average causal effect for the population, $\mathbb{E}[Y_1 - Y_0]$, we could again do this under the no-unmeasured-confounding assumption by $\mathbb{E}[Y_1 - Y_0] = \sum_c \mathbb{E}[Y_1 - Y_0|c]P(c) = \sum_c \{\mathbb{E}[Y|A = 1, c] - \mathbb{E}[Y|A = 0, c]\}P(c)$, where once again this final expression is one that we can empirically estimate from the observed data. If the exposure is randomized, then we have that $Y_a \perp\!\!\!\perp A$ so that the no unmeasured confounding assumption holds without conditioning on any covariates at all. In this case we can estimate the average causal effect just by comparing observed outcomes across the exposure groups: $\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$.

With observational data for which the exposure is not randomized, we typically try to collect data on and stratify by, or otherwise control for, a sufficiently rich set of covariates C so that the no-unmeasured-confounding assumption is plausible. However, with observational data we never know whether the assumption is in fact satisfied or how severely it might be violated. Sensitivity analysis techniques, the topic of Chapter 3, can be useful in assessing how strong an unmeasured confounder would have to affect both the exposure and the outcome in order to alter conclusions. If we control for the measured covariates c under a regression model such as $\mathbb{E}[Y|a, c] = \alpha_0 + \alpha_1 a + \alpha_2' c$, then under the assumption of no-unmeasured-confounding, conditional on measured covariates c, the average causal effect of the exposure A on outcome Y is given by α_1 since $\mathbb{E}[Y_1 - Y_0|c] = \mathbb{E}[Y|A = 1, c] - \mathbb{E}[Y|A = 0, c] = \{\alpha_0 + \alpha_1 1 + \alpha_2' c\} - \{\alpha_0 + \alpha_1 0 + \alpha_2' c\} = \alpha_1$. Thus if the regression model is correctly specified, the regression coefficient for the exposure can be interpreted as the average causal effect if there is no unmeasured confounding conditional on covariates C.

As noted in the text, here we will use the terms "potential outcome" and "counterfactual outcome" interchangeably. Some authors (e.g., Rubin, 2005) contrast the use of the term "potential outcome" and "counterfactual outcome" and prefer to use the term "potential outcome" within the context of the outcomes under each of two potential exposures, states, or interventions. If there are two potential exposure states, 0 and 1, then Y_0 would be used to denote the outcome under exposure 0, and Y_1 would be used to denote the outcome under exposure 1. If in fact exposure 0 takes place, then the outcome Y_0 occurs; it is what actually occurred; it is not counterfactual or contrary to fact. If the actual exposure was 0, then the outcome that would have taken place under exposure 1, Y_1, is counterfactual; that is, it is the outcome that would have occurred if, contrary to fact, something had taken place, namely the exposure being 1, other than what actually did. If exposure 1 takes place, then the outcome Y_1 occurs; and the outcome that would have taken place under exposure 0, Y_0, is counterfactual. Thus, under Rubin's terminology, only one of the two outcomes is "counterfactual" or contrary to fact. Rubin also prefers using the terminology "potential outcomes" because, prior to any exposure taking place, neither of the outcomes are contrary to fact. Other authors use "counterfactual outcomes" for both Y_0 and Y_1. In this use of terminology, the view is that some exposure, either 0 or 1, will naturally occur and this will lead to the actual outcome Y. The variables Y_0 and Y_1 are viewed as those that would have taken place had there been an *intervention* to set the exposure to 0 or to 1, respectively. Thus, with respect to the outcome that would have naturally occurred, Y, both Y_0 and Y_1, outcomes that would have occurred under some intervention, are viewed as "counterfactual." This is because Y_0 and Y_1 are the outcomes that would have taken place if, contrary to fact, there had been an intervention to set the exposure to 0 or 1, respectively. Under this perspective, it is also assumed that if the exposure that naturally occurs is 0, then $Y = Y_0$; and if the exposure that naturally occurs is 1, then $Y = Y_1$—that is, that if the exposure *naturally takes* a particular value, then the actual outcome is equal to the outcome that would have occurred *under an intervention* to set the exposure for that individual to that value. As noted above, this is sometimes referred to as the "consistency" assumption and is mentioned above and discussed further in Section A.7.1 and in the text in Section 7.2. The assumption is also effectively embedded in Rubin's formulation of the potential outcomes framework and is a component of what he refers to as the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980, 1986). We will discuss this assumption in detail in Appendix A7.1. See Chapters 7 and 15 or VanderWeele and Hernán (2013) for further discussion.

A.2. MEDIATION: INTRODUCTION AND REGRESSION-BASED APPROACHES

A.2.1. Definitions and Identification

Let Y_a denote a subject's outcome if exposure A were set, possibly contrary to fact, to a. Let M_a denote a subject's counterfactual value of the intermediate M if exposure A were set to the value a. Finally, let Y_{am} denote a subject's counterfactual value for Y if A were set to a and M were set to m. We make an assumption,

sometimes referred to as composition, that $Y_a = Y_{aM_a}$—that is, the value of Y that would occur if A were set to a is equal to the value of Y that would occur if A were set to a and M were set to what it would have been if A were set to a. We also make an assumption, sometimes referred to as the consistency assumption, that when $A = a$, the counterfactual outcomes Y_a and M_a are, respectively, equal to the observed outcomes Y and M. We likewise assume that when $A = a$ and $M = m$, the counterfactual outcome Y_{am} is equal to the observed Y. Further discussion of the interpretation of these consistency assumptions in the context of mediation is given elsewhere (VanderWeele, 2009b; VanderWeele and Vansteelandt, 2009).

Robins and Greenland (1992) and Pearl (2001) gave the following definitions for controlled direct effects and natural direct and indirect effects based on interventions on the mediator M. The controlled direct effect of exposure A on outcome Y comparing $A = a$ with $A = a^*$ and setting M to m is defined by $Y_{am} - Y_{a^*m}$ and measures the effect of A on Y not mediated through M—that is, the effect of A on Y after intervening to fix the mediator to some value m. In contrast to controlled direct effects, natural direct effects fix the intermediate variable for each individual to the level it naturally would have been under—for example, the absence of exposure. The natural direct effect of exposure A on outcome Y comparing $A = a$ with $A = a^*$ intervening to set M to what it would have been if exposure had been $A = a^*$ is formally defined by $Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$. Essentially, the natural direct effect assumes that the intermediate M is set to M_{a^*}, the level it would have been for each individual had exposure been a^*, and then compares the outcomes with exposure set to a versus a^* (with the intermediate set to this level M_{a^*}). Corresponding to a natural direct effect is a natural indirect effect. The natural indirect effect comparing $A = a$ with $A = a^*$ and intervening to set exposure A to a is formally defined by $Y_{aM_a} - Y_{aM_{a^*}}$. The natural indirect effect assumes that exposure is set to some level $A = a$ and then compares what would have happened if the mediator were set to what it would have been if exposure had been a versus what would have happened if the mediator were set to what it would have been if exposure had been a^*. We can also consider the average values of the controlled direct effect and natural direct and indirect effects, either for a population, $\mathbb{E}[Y_{am} - Y_{a^*m}]$, $\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}]$, and $\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}]$, or conditional on covariates $C = c$, $\mathbb{E}[Y_{am} - Y_{a^*m}|c]$, $\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]$, and $\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c]$.

A total effect can be decomposed into a natural direct and indirect. To see this, note that the total effect $Y_a - Y_{a^*}$ can be written as $Y_a - Y_{a^*} = Y_{aM_a} - Y_{a^*M_{a^*}} = (Y_{aM_a} - Y_{aM_{a^*}}) + (Y_{aM_{a^*}} - Y_{a^*M_{a^*}})$, where the first expression in the sum is the natural indirect or mediated effect and the second expression is the natural direct effect. The decomposition is obtained by adding and subtracting $Y_{aM_{a^*}}$. An important difference between controlled and natural direct effects is that the effect decomposition above works for natural direct and indirect effects but not for controlled direct effects. If one subtracts a controlled direct effect from a total effect, the resulting quantity cannot in general be interpreted as an indirect effect unless there is no interaction at the individual level between the effects of the exposure and the mediator on the outcome (Robins, 2003; Kaufman et al., 2004) in which case controlled direct effects and natural direct effects are equivalent since $Y_{am} - Y_{a^*m}$ will be constant for all values of m and thus $Y_{am} - Y_{a^*m} = Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$.

The effects above are sometimes referred to as "pure" (natural) direct effects and "total" natural indirect effects (Robins and Greenland, 1992). An alternative decomposition can be obtained by adding and subtracting $Y_{a^*M_a}$ instead of $Y_{aM_{a^*}}$. We then have $Y_a - Y_{a^*} = Y_{aM_a} - Y_{a^*M_{a^*}} = (Y_{aM_a} - Y_{a^*M_a}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}})$ where the first expression in the sum is sometimes referred to as the "total" (natural) direct effect and the second expression is the "pure" (natural) indirect effect. These two different decompositions essentially arise from different ways of accounting for interaction as discussed in Chapters 7 and 14.

With a binary outcome, we would likewise define direct and indirect effects on a risk ratio or odds ratio scale (VanderWeele and Vansteelandt, 2010). On the odds ratio scale, the total effect conditional on $C = c$ is given by $OR^{TE}_{a,a^*|c} = \frac{P(Y_a=1|c)/\{1-P(Y_a=1|c)\}}{P(Y_{a^*}=1|c)/\{1-P(Y_{a^*}=1|c)\}}$. The controlled direct effect on the odds ratio scale conditional on $C = c$ is given by $OR^{CDE}_{a,a^*|c}(m) = \frac{P(Y_{am}=1|c)/\{1-P(Y_{am}=1|c)\}}{P(Y_{a^*m}=1|c)/\{1-P(Y_{a^*m}=1|c)\}}$. The natural direct effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NDE}_{a,a^*|c}(a^*) = \frac{P(Y_{aM_{a^*}}=1|c)/\{1-P(Y_{aM_{a^*}}=1|c)\}}{P(Y_{a^*M_{a^*}}=1|c)/\{1-P(Y_{a^*M_{a^*}}=1|c)\}}$. The natural indirect effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NIE}_{a,a^*|c}(a) = \frac{P(Y_{aM_a}=1|c)/\{1-P(Y_{aM_a}=1|c)\}}{P(Y_{aM_{a^*}}=1|c)/\{1-P(Y_{aM_{a^*}}=1|c)\}}$. On a risk ratio scale conditional on $C = c$, the total effect is given by $RR^{TE}_{a,a^*|c} = \frac{P(Y_a=1|c)}{P(Y_{a^*}=1|c)}$, the controlled direct effect is given by $RR^{CDE}_{a,a^*|c}(m) = \frac{P(Y_{am}=1|c)}{P(Y_{a^*m}=1|c)}$, and the natural direct effect is given by $RR^{NDE}_{a,a^*|c}(a^*) = \frac{P(Y_{aM_{a^*}}=1|c)}{P(Y_{a^*M_{a^*}}=1|c)}$. The natural indirect effect on the risk ratio scale conditional on $C = c$ is given by $RR^{NIE}_{a,a^*|c}(a) = \frac{P(Y_{aM_a}=1|c)}{P(Y_{aM_{a^*}}=1|c)}$. The total effect then decomposes into the product of the natural direct and indirect effects on the odds ratio or risk ratio scale: $OR^{TE}_{a,a^*|c} = OR^{NIE}_{a,a^*|c}(a) \times OR^{NDE}_{a,a^*|c}(a^*)$ and $RR^{TE}_{a,a^*|c} = RR^{NIE}_{a,a^*|c}(a) \times RR^{NDE}_{a,a^*|c}(a^*)$.

We will use the notation $A \perp\!\!\!\perp B|C$ to denote that A is independent of B conditional on C. Total effects are identified if, conditional on some set of measured covariates C, the effect of exposure A on outcome Y is unconfounded given C; in counterfactual notation, this is $Y_a \perp\!\!\!\perp A|C$. Controlled direct effects are identified if control is made for a set of covariates C that includes all confounders of not only the exposure–outcome relationship but also the mediator–outcome relationship. In counterfactual notation, we require that for all a and m,

$$Y_{am} \perp\!\!\!\perp A|C \qquad \textbf{(A.2.1)}$$

$$Y_{am} \perp\!\!\!\perp M|\{A, C\} \qquad \textbf{(A.2.2)}$$

Proposition 2.1 (Robins, 1986; cf. Pearl, 2001):
If assumptions (A2.1) and (A2.2) hold, then average controlled direct effects conditional on C are identified and given by

$$\mathbb{E}[Y_{am} - Y_{a^*m}|c] = \mathbb{E}[Y|a, m, c] - \mathbb{E}[Y|a^*, m, c]$$

Proof:
We have

$$\begin{aligned}\mathbb{E}[Y_{am} - Y_{a^*m}|c] &= \mathbb{E}[Y_{am}|a,c] - \mathbb{E}[Y_{a^*m}|a^*,c] \text{ by (A2.1)} \\ &= \mathbb{E}[Y_{am}|a,m,c] - \mathbb{E}[Y_{a^*m}|a^*,m,c] \text{ by (A2.2)} \\ &= \mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] \text{ by consistency.} \quad \blacksquare\end{aligned}$$

From Proposition 2.1 we have that the average controlled direct effect for a population is given by $\mathbb{E}[Y_{am} - Y_{a^*m}] = \sum_c \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(c)$.

Natural direct and indirect effects will be identified if, in addition to assumptions (A2.1) and (A2.2), the following two assumptions hold; that is, for all a, a^*, and m, we have

$$M_a \perp\!\!\!\perp A|C \qquad \textbf{(A.2.3)}$$

$$Y_{am} \perp\!\!\!\perp M_{a^*}|C \qquad \textbf{(A.2.4)}$$

Assumption (A2.3) can be interpreted as: conditional on C, there is no unmeasured confounding of the exposure-mediator relationship. On a causal diagram interpreted as a set of nonparametric structural equations (Pearl, 2009), if assumption (A2.2) holds, then assumption (A2.4) will hold if there is no variable L that is affected by the exposure A and that itself affects both M and Y—that is, no effect of exposure A that confounds the mediator–outcome relationship. If, however, there is an effect of the exposure that confounds the mediator–outcome relationship as in Figure 2.2, in the text then natural direct and indirect effects will not in general be identified irrespective of whether data is available on L or not (Avin et al., 2005), except under strong assumptions about no-interaction at the individual level (Robins, 2003). To see the relationship between assumption (A2.4) stated as the counterfactual independence $Y_{am} \perp\!\!\!\perp M_{a^*}|C$ and this assumption restated as the absence of mediator–outcome confounders affected by the exposure, consider first Figure 2.1. The nonparametric structural equations (Pearl, 2009) for Y and for M would be given by $Y = f_Y(C,A,M,\varepsilon_Y)$ and $M = f_M(C,A,\varepsilon_M)$, where ε_Y and ε_M are independent. Conditional on $C = c$, the counterfactual Y_{am} is given by $Y_{am} = f_Y(c,a,m,\varepsilon_Y)$ and the counterfactual M_{a^*} is given by $M_{a^*} = f_M(c,a^*,\varepsilon_M)$. Conditional on $C = c$, the variables Y_{am} and M_{a^*} are independent since Y_{am} is simply a function of ε_Y, and M_{a^*} is simply a function of ε_M, and thus the independence of Y_{am} and M_{a^*} follows from the independence of ε_Y and ε_M. Now consider Figure 2.2. The nonparametric structural equations for Y and for M and L would be given by $Y = f_Y(C,A,L,M,\varepsilon_Y)$, $M = f_M(C,A,L,\varepsilon_M)$, and $L = f_L(C,A,\varepsilon_L)$, where ε_Y, ε_M, ε_L are mutually independent. Conditional on $C = c$, the counterfactual Y_{am} is given by $Y_{am} = f_Y(c,a,m,L,\varepsilon_Y) = f_Y(c,a,m,f_L(c,a,\varepsilon_L),\varepsilon_Y)$ and the counterfactual M_{a^*} is given by $M_{a^*} = f_M(c,a^*,L,\varepsilon_M) = f_M(c,a^*,f_L(c,a^*,\varepsilon_L),\varepsilon_M)$. In general, Y_{am} and M_{a^*} will not be independent conditional on $C = c$ since both are functions of ε_L. Thus assumption (A2.4) would hold on nonparametric structural equations in Figure 2.1 but not in Figure 2.2. Unless otherwise specified, we will interpret all causal diagrams as non-parametric structural equation models as in Pearl (2009).

Note that if exposure A is randomized, then assumptions (A2.1) and (A2.3) will hold automatically, but assumptions (A2.2) and (A2.4) may not.

Proposition 2.2 (Pearl, 2001):
If assumptions (A2.1)–(A2.4) hold, then the average natural direct effect conditional on C is identified and is given by

$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \sum_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c)$$

and the average natural indirect effect conditional on C is identified and is given by

$$\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = \sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}$$

Proof:
We have

$$\begin{aligned}
\mathbb{E}[Y_{aM_{a^*}}|c] &= \sum_m \mathbb{E}[Y_{am}|c, M_{a^*} = m]P(M_{a^*} = m|c) \quad \text{by iterated expectations} \\
&= \sum_m \mathbb{E}[Y_{am}|c]P(M_{a^*} = m|a^*,c) \quad \text{by (A2.4) and (A2.3)} \\
&= \sum_m \mathbb{E}[Y_{am}|a,c]P(M = m|a^*,c) \quad \text{by (A2.1) and consistency} \\
&= \sum_m \mathbb{E}[Y_{am}|a,m,c]P(m|a^*,c) \quad \text{by (A2.2)} \\
&= \sum_m \mathbb{E}[Y|a,m,c]P(m|a^*,c) \quad \text{by consistency}
\end{aligned}$$

If we apply this result and replace a with a^*, we get $\mathbb{E}[Y_{a^*M_{a^*}}|c] = \sum_m \mathbb{E}[Y|a^*,m,c]$ $P(m|a^*,c)$, AND from this it follows that the average natural direct effect is given by

$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \sum_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c)$$

If we apply this result and replace a^* with a, we get $\mathbb{E}[Y_{aM_a}|c] = \sum_m \mathbb{E}[Y|a,m,c]$ $P(m|a,c)$, and from this it follows that the average natural indirect effect is given by

$$\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = \sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}. \quad \blacksquare$$

The expression in Proposition 2.2 for the natural indirect effect, $\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = \sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}$, is sometimes referred to as the "mediation formula" (Pearl, 2012). From Proposition 2.2, the natural direct and indirect effects for a population are given by $\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}] = \sum_{m,c}\{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c)P(c)$. and $\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = \sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}$, respectively.

A.2.2. Regression Methods for Direct and Indirect Effects

Using the empirical formulae in Propositions 2.1 and 2.2, we can derive controlled direct effects and natural direct and indirect effects for any statistical models specified for $\mathbb{E}[Y|a,m,c]$ and $P(m|a,c)$. In this section, we will derive such controlled direct effects and natural direct and indirect effects for some standard linear and logistic regression models for the outcome and for the mediator, wherein the outcome models allow for potential exposure–mediator interaction. Although we will

consider specific regression models in this section, the counterfactual approach to mediation is very flexible and similar effects could be derived for any other models for $\mathbb{E}[Y|a,m,c]$ and $P(m|a,c)$.

Proposition 2.3 (VanderWeele and Vansteelandt, 2009):
If assumptions (A2.1)–(A2.4) hold and if Y and M are continuous and the following regression models for Y and M are correctly specified:

$$\mathbb{E}[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

then the average controlled direct effect and the average natural direct and indirect effects, conditional on $C = c$, are given by

$$\mathbb{E}[Y_{am} - Y_{a^*m}|c] = (\theta_1 + \theta_3 m)(a - a^*)$$
$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$$
$$\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

with standard errors for these effects given by

$$\sqrt{\Gamma \Sigma \Gamma'}|a - a^*|$$

where

$$\Sigma \equiv \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{pmatrix}$$

with Σ_β and Σ_θ the covariance matrices for the estimators $\widehat{\beta}$ of $\beta \equiv (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}$ of $\theta \equiv (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4')'$ and $\Gamma \equiv (0, 0, 0', 0, 1, 0, m, 0')$ for the controlled direct effect, $\Gamma \equiv (\theta_3, \theta_3 a^*, \theta_3 C', 0, 1, 0, \beta_0 + \beta_1 a^* + \beta_2' C, 0')$ for the natural direct effect and $\Gamma \equiv (0, \theta_2 + \theta_3 a, 0', 0, 0, \beta_1, \beta_1 a, 0')$ for the natural indirect effect, where $0'$ denotes a row vector of the dimension of C, containing only zeroes.

Proof:
If the regression models are correctly specified and assumptions (A2.1) and (A2.2) hold then we could compute the controlled direct effect as follows:

$$\begin{aligned}\mathbb{E}[Y_{am} - Y_{a^*m}|c] &= \mathbb{E}[Y|c,a,m] - \mathbb{E}[Y|c,a^*,m] \\ &= (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) \\ &\quad -(\theta_0 + \theta_1 a^* + \theta_2 m + \theta_3 a^* m + \theta_4' c) \\ &= (\theta_1 a + \theta_3 am - \theta_1 a^* - \theta_3 a^* m) \\ &= \theta_1(a - a^*) + \theta_3 m(a - a^*)\end{aligned}$$

Under assumptions (A2.1)–(A2.4) we could compute natural direct effects by

$$\begin{aligned}&\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] \\ &= \sum\nolimits_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c)\end{aligned}$$

$$
\begin{aligned}
&= \sum\nolimits_m \left\{ (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) \right. \\
&\quad \left. -(\theta_0 + \theta_1 a^* + \theta_2 m + \theta_3 a^* m + \theta_4' c) \right\} P(m|c, a^*) \\
&= \sum\nolimits_m \{\theta_1 a + \theta_2 m + \theta_3 am) - (\theta_1 a^* + \theta_2 m + \theta_3 a^* m)\} P(m|c, a^*) \\
&= \{\theta_1 a + \theta_2 \mathbb{E}[M|a^*, c] + \theta_3 a \mathbb{E}[M|a^*, c]) \\
&\quad -(\theta_1 a^* + \theta_2 \mathbb{E}[M|a^*, c] + \theta_3 a^* \mathbb{E}[M|a^*, c])\} \\
&= \{\theta_1 a + \theta_2 (\beta_0 + \beta_1 a^* + \beta_2' c) + \theta_3 a (\beta_0 + \beta_1 a^* + \beta_2' c)) \\
&\quad -(\theta_1 a^* + \theta_2 (\beta_0 + \beta_1 a^* + \beta_2' c) + \theta_3 a^* (\beta_0 + \beta_1 a^* + \beta_2' c))\} \\
&= \{\theta_1 a + \theta_3 a (\beta_0 + \beta_1 a^* + \beta_2' c) - (\theta_1 a^* + \theta_3 a^* (\beta_0 + \beta_1 a^* + \beta_2' c))\} \\
&= (\theta_1 + \theta_3 \beta_0 + \theta_3 \beta_1 a^* + \theta_3 \beta_2' c)(a - a^*)
\end{aligned}
$$

and we could compute natural indirect effects by

$$
\begin{aligned}
&\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}} | c] \\
&\quad = \sum\nolimits_m \mathbb{E}[Y|a, m, c] \{P(m|a, c) - P(m|a^*, c)\}. \\
&\quad = \sum\nolimits_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) P(m|c, a) \\
&\qquad - \sum\nolimits_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) P(m|c, a^*) \\
&\quad = (\theta_0 + \theta_1 a + \theta_2 \mathbb{E}[M|a, c] + \theta_3 a \mathbb{E}[M|a, c] + \theta_4' c) \\
&\qquad - \sum\nolimits_c (\theta_0 + \theta_1 a + \theta_2 \mathbb{E}[M|a^*, c] + \theta_3 a \mathbb{E}[M|a^*, c] + \theta_4' c) \\
&\quad = (\theta_0 + \theta_1 a + \theta_2 (\beta_0 + \beta_1 a + \beta_2' c) + \theta_3 a (\beta_0 + \beta_1 a + \beta_2' c) + \theta_4' c) \\
&\qquad - (\theta_0 + \theta_1 a + \theta_2 (\beta_0 + \beta_1 a^* + \beta_2' c) \\
&\qquad + \theta_3 a (\beta_0 + \beta_1 a^* + \beta_2' c) + \theta_4' c) \\
&\quad = \theta_2 \beta_1 (a - a^*) + \theta_3 \beta_1 a (a - a^*)
\end{aligned}
$$

Let Σ_β and Σ_θ be the covariance matrices for the estimators $\widehat{\beta}$ of $\beta \equiv (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}$ of $\theta \equiv (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4')'$, then the covariance matrix of $(\beta', \hat{\theta}')'$ is

$$
\Sigma \equiv \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{pmatrix}
$$

which can be seen upon noting that

$$
\begin{aligned}
\mathrm{Cov}(\beta, \hat{\theta}) &= E\left\{ \mathrm{Cov}(\beta, \hat{\theta} | M, A, C) \right\} + \mathrm{Cov}\left\{ \mathbb{E}(\beta | M, A, C), \mathbb{E}(\hat{\theta} | M, A, C) \right\} \\
&= 0 + \mathrm{Cov}\,(\beta, \theta) = 0
\end{aligned}
$$

where we use the fact that β is a function of M, A, and C only. Standard errors of the controlled and natural direct and indirect effects in can then be obtained using the Delta method as

$$
\sqrt{\Gamma \Sigma \Gamma'} |a - a^*|
$$

with $\Gamma \equiv (0,0,0',0,1,0,m,0')$ for the controlled direct effect in (9), $\Gamma \equiv (\theta_3,\theta_3 a^*,\theta_3 C',0,1,0,\beta_0+\beta_1 a^*+\beta_2' C,0')$ for the natural direct effect and $\Gamma \equiv (0,\theta_2+\theta_3 a,0',0,0,\beta_1,\beta_1 a,0')$ for the total natural indirect effect. ■

Proposition 2.4 (VanderWeele and Vansteelandt, 2010):
If assumptions (A2.1)–(A2.4) hold and if Y is dichotomous and rare and M continuous and the following regression models for Y and M are correctly specified:

$$\text{logit}\{P(Y=1|a,m,c])\} = \theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c$$
$$\mathbb{E}[M|a,c] = \beta_0+\beta_1 a+\beta_2' c$$

with M conditionally normally distributed given A, C with conditional variance σ^2, then the average controlled direct effect and the average natural direct and indirect effects are given by

$$\begin{aligned} OR^{CDE}(m) &= \exp\{(\theta_1+\theta_3 m)(a-a^*)\} \\ OR^{NDE} &\approx \exp\{(\theta_1+\theta_3\beta_0+\theta_3\beta_1 a^*+\theta_3\beta_2' C+\theta_3\theta_2\sigma^2)(a-a^*) \\ &\quad +0.5\theta_3^2\sigma^2(a^2-a^{*2})\} \\ OR^{NIE} &\approx \exp\{(\theta_2\beta_1+\theta_3\beta_1 a)(a-a^*)\} \end{aligned}$$

with the approximation holding to the extent that the outcome Y is rare and with standard errors for the log of these effects given by

$$\sqrt{\Gamma\Sigma\Gamma'}|a-a^*|$$

where

$$\Sigma \equiv \begin{pmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_\theta & 0 \\ 0 & 0 & \Sigma_{\sigma^2} \end{pmatrix}$$

with Σ_β, Σ_θ, and Σ_{σ^2} the covariance matrices for the estimators $\widehat{\beta}$ of $\beta \equiv (\beta_0,\beta_1,\beta_2')'$ and $\hat{\theta}$ of $\theta \equiv (\theta_0,\theta_1,\theta_2,\theta_3,\theta_4')'$ and $\hat{\sigma}^2$ of σ^2 and with $\Gamma \equiv (0,0,0',0,1,0,m,0',0)$ for the log of the controlled direct effect odds ratio, $\Gamma \equiv (0,\theta_2+\theta_3 a,0',0,0,\beta_1,\beta_1 a,0',0)$ for the log of the natural indirect effect odds ratio, and $\Gamma \equiv (\theta_3,\theta_3 a^*,\theta_3 c,0,1,\theta_3\sigma^2,\beta_0+\beta_1 a^*+\beta_2' c+\theta_2\sigma^2+\theta_3\sigma^2(a+a^*),0',\theta_3\theta_2+0.5\theta_3^2(a+a^*))$ for the log of the natural direct effect odds ratio, where $0'$ denotes a row vector of the dimension of c, containing zeros only.

Proof:
We have that

$$\begin{aligned} OR^{CDE}_{a,a^*|c}(m) &= \frac{P(Y=1|a,m,c)/\{1-P(Y=1|a,m,c)\}}{P(Y=1|a^*,m,c)/\{1-P(Y=1|a^*,m,c)\}} \\ &= \frac{\exp\{\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c\}}{\exp\{\theta_0+\theta_1 a^*+\theta_2 m+\theta_3 a^* m+\theta_4' c\}} \\ &= \exp\{(\theta_1+\theta_3 m)(a-a^*)\}. \end{aligned}$$

For the natural direct and indirect effect odds ratios, we have that logit $\{P(Y_{aM_{a^*}} = 1|c)\}$

$$
\begin{aligned}
&\approx \log\{P(Y_{aM_{a^*}} = 1|c)\} \\
&= \log\left\{\int P(Y_{am} = 1|c, M_{a^*} = m)P(M_{a^*} = m|c)dm\right\} \\
&= \log\left\{\int P(Y_{am} = 1|c)P(M_{a^*} = m|c)dm\right\} \quad \text{by (A2.4)} \\
&= \log\left\{\int P(Y = 1|a, m, c)P(M = m|a^*, c)dm\right\} \quad \text{by (A2.1)–(A2.3)} \\
&\approx \log\left\{\int \exp(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c)P(M = m|a^*, c)dm\right\} \\
&= \log\left\{\exp(\theta_0 + \theta_1 a + \theta_4' c)\int \exp\{(\theta_2 + \theta_3 a)m\}P(M = m|a^*, c)dm\right\} \\
&= \log\left\{\exp(\theta_0 + \theta_1 a + \theta_4' c)\mathbb{E}[e^{(\theta_2+\theta_3 a)M}|a^*, c]\right\} \\
&= \log\left\{\exp(\theta_0 + \theta_1 a + \theta_4' c)\exp((\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c)\right. \\
&\quad \left. + \frac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2)\right\} \\
&= \theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) \\
&\quad + \frac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2
\end{aligned}
$$

Similarly, we have that

$$
\begin{aligned}
&\text{logit}\{P(Y_{aM_a} = 1|c)\} \\
&\quad = \theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2
\end{aligned}
$$

Thus for the natural indirect effect odds ratio we have

$$
\begin{aligned}
&\log\{OR^{NIE}_{a,a^*|c}(a)\} \\
&\quad = \log\left[\frac{P(Y_{aM_a} = 1|c)/\{1 - P(Y_{aM_a} = 1|c)\}}{P(Y_{aM_{a^*}} = 1|c)/\{1 - P(Y_{aM_{a^*}} = 1|c)\}}\right] \\
&\quad = \text{logit}\{P(Y_{aM_a} = 1|c)\} - \text{logit}\{P((Y_{aM_{a^*}} = 1|c)\} \\
&\quad \approx \theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2 \\
&\qquad - \{\theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2\} \\
&\quad = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)
\end{aligned}
$$

Exponentiating gives $OR^{NIE}_{a,a^*|c}(a) \approx \exp\{(\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)\}$, and this completes the proof for the expression for the natural indirect effect odds ratio.

For the natural direct effect odds ratio, we have that

$$\text{logit}\{P(Y_{a^*M_{a^*}} = 1|c)\} = \theta_0 + \theta_1 a^* + \theta_4' c + (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^*)^2\sigma^2$$

and thus

$$\begin{aligned}
&\log\{OR^{NDE}_{a,a^*|c}(a^*)\} \\
&= \log\left[\frac{P(Y_{aM_{a^*}} = 1|c)/\{1 - P(Y_{aM_{a^*}} = 1|c)\}}{P(Y_{a^*M_{a^*}} = 1|c)/\{1 - P(Y_{a^*M_{a^*}} = 1|c)\}}\right] \\
&= \text{logit}\{P(Y_{aM_{a^*}} = 1|c)\} - \text{logit}\{P((Y_{a^*M_{a^*}} = 1|c)\} \\
&\approx \theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2 \\
&\quad - \{\theta_0 + \theta_1 a^* + \theta_4' c + (\theta_2 + \theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c) + \tfrac{1}{2}(\theta_2 + \theta_3 a^*)^2\sigma^2\} \\
&= \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})
\end{aligned}$$

Exponentiating gives

$$OR^{NDE}_{a,a^*|c}(a^*) \approx \exp\left[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right]$$

Suppose that the resulting estimates $\hat{\beta}$ of $\beta \equiv (\beta_0, \beta_1, \beta_2')'$, $\hat{\theta}$ of $\theta \equiv (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4')'$, and $\hat{\sigma}^2$ of σ^2 have covariance matrices Σ_β, Σ_θ, and Σ_{σ^2}. Note that under a linear regression for the mediator, if we let *RSS* denote the residual sum of square, an unbiased estimate of $\hat{\sigma}^2$ is given by $RSS/(n-p)$ where n is the sample size and p is the number of parameters in the regression model; the variance of $\hat{\sigma}^2$ can be estimated by $\frac{2\hat{\sigma}^4}{n-p}$. Then standard errors of the log of the controlled direct effect odds ratios and natural direct and indirect effect odds ratios can be obtained using the Delta method as

$$\sqrt{\Gamma \sum \Gamma'}|a - a^*|$$

with

$$\sum \equiv \begin{pmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_\theta & 0 \\ 0 & 0 & \Sigma_{\sigma^2} \end{pmatrix}$$

and with $\Gamma \equiv (0, 0, 0', 0, 1, 0, m, 0', 0)$ for the log of the controlled direct effect odds ratio, $\Gamma \equiv (0, \theta_2 + \theta_3 a, 0', 0, 0, \beta_1, \beta_1 a, 0', 0)$ for the log of the natural indirect effect

odds ratio, and $\Gamma \equiv (\theta_3, \theta_3 a^*, \theta_3 c, 0, 1, \theta_3 \sigma^2, \beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 \sigma^2 + \theta_3 \sigma^2 (a + a^*), 0', \theta_3 \theta_2 + 0.5\theta_3^2 (a + a^*))$ for the log of the natural direct effect odds ratio. ■

Proposition 2.5 (Valeri and VanderWeele, 2013):
If assumptions (A2.1)–(A2.4) hold and if Y is continuous and M binary with the following regression models for Y and M correctly specified:

$$\mathbb{E}[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M=1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

then the average controlled direct effect and the average natural direct and indirect effects are given by

$$\mathbb{E}[Y_{am} - Y_{a^*m}|c] = (\theta_1 + \theta_3 m)(a - a^*)$$
$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \{\theta_1 (a - a^*)\} + \{\theta_3 (a - a^*)\} \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]}$$
$$\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = (\theta_2 + \theta_3 a) \left\{ \frac{\exp[\beta_0 + \beta_1 a + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]} \right\}$$

with standard errors for the controlled direct effect and natural direct effect given by

$$\sqrt{\Gamma \Sigma \Gamma'} |a - a^*|$$

where

$$\Sigma \equiv \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{pmatrix}$$

with Σ_β and Σ_θ the covariance matrices for the estimators $\widehat{\beta}$ of $\beta \equiv (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}$ of $\theta \equiv (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4')'$ and with $\Gamma = (0, 0, 0', 0, 1, 0, m, 0')$ for the controlled direct effect and $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the natural direct effect, where

$$d_1 = \frac{\theta_3 \exp[[\beta_0 + \beta_1 a^* + \beta_2' c](1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]) - \theta_3 \{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]\}^2}{(1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c])^2}$$
$$d_2 = \frac{\theta_3 a^* \exp[\beta_0 + \beta_1 a^* + \beta_2' c](1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]) - \{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]\}^2}{(1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c])^2}$$
$$d_3 = \frac{\theta_3 c' \exp[\beta_0 + \beta_1 a^* + \beta_2' c](1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]) - \{\exp[\beta_0 + \beta_1 a^* \beta_2' c]\}^2}{(1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c])^2}$$
$$d_4 = 0$$

$$d_5 = 1$$
$$d_6 = 0$$
$$d_7 = \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}$$
$$d_8 = 0^{'}$$

Standard errors of the natural indirect are given by

$$\sqrt{\Gamma \Sigma \Gamma^{'}}$$

where if we let

$$Q = \frac{\exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\{1 + \exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\} - \{\exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}^2}{\{1 + \exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}^2}$$
$$B = \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\} - \{\exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}^2}{\{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}^2}$$
$$K = \frac{\exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + \exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$
$$D = \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

then $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$, where

$$d_1 = \{\theta_2 + \theta_3 a\}[Q - B]$$
$$d_2 = \{\theta_2 + \theta_3 a\}[aQ - a^* B]$$
$$d_3 = \{\theta_2 + \theta_3 a\} c^{'} [Q - B]$$
$$d_4 = 0$$
$$d_5 = 0$$
$$d_6 = K - D$$
$$d_7 = a[K - D]$$
$$d_8 = 0^{'}$$

Proof:
For the natural direct effect we have that

$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^* M_{a^*}} | C = c]$$
$$= \sum\nolimits_m \{\mathbb{E}[Y|c, a, m] - \mathbb{E}[Y|c, a^*, m]\} P(m|c, a^*)$$

$$= \sum_m \{(\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) - (\theta_0 + \theta_1 a^* + \theta_2 m + \theta_3 a^* m + \theta_4' c)\} P(m|c, a^*)$$
$$= \sum_m \{(\theta_1 a + \theta_2 m + \theta_3 am) - (\theta_1 a^* + \theta_2 m + \theta_3 a^* m)\} P(m|c, a^*)$$
$$= \{\theta_1 (a - a^*)\} + \{\theta_3 (a - a^*)\} \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]}$$

For the natural indirect effect we have

$$\begin{aligned}
&\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}} | C = c] \\
&= \sum_m \mathbb{E}[Y|c, a, m] \{P(m|c, a) - P(m|c, a^*)\} \\
&= \sum_m (\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c) \{P(m|c, a) - P(m|c, a^*)\} \\
&= (\theta_2 + \theta_3 a) \{\mathbb{E}[M|a, c] - \mathbb{E}[M|a^*, c]\} \\
&= (\theta_2 + \theta_3 a) \left\{ \frac{\exp[\beta_0 + \beta_1 a + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]} \right\}
\end{aligned}$$

The standard errors follow by the delta method as in Propositions 2.3 and 2.4. ■

Proposition 2.6. (Valeri and VanderWeele, 2013):
If assumptions (A2.1)–(A2.4) hold and if Y and M are binary, and if the outcome Y is rare with the following regression models for Y and M are correctly specified:

$$\text{logit}\{P(Y = 1|a, m, c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M = 1|a, c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

then the average controlled direct effect and the average natural direct and indirect effects are given by

$$\begin{aligned}
OR^{CDE}(m) &= (\theta_1 + \theta_3 m)(a - a^*) \\
OR^{NDE} &= \frac{\exp(\theta_1 a)\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}{\exp(\theta_1 a^*)\{1 + \exp(\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c)\}} \\
OR^{NIE} &= \frac{\{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2' c)\}}{\{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}
\end{aligned}$$

Standard errors of the controlled and natural direct and indirect effects are given by

$$\sqrt{\Gamma \Sigma \Gamma'}$$

where

$$\Sigma \equiv \begin{pmatrix} \Sigma_\beta & 0 \\ 0 & \Sigma_\theta \end{pmatrix}$$

with Σ_β and Σ_θ the covariance matrices for the estimators $\widehat{\beta}$ of $\beta \equiv (\beta_0, \beta_1, \beta_2')'$ and $\hat{\theta}$ of $\theta \equiv (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4')'$ and with $\Gamma = (0, 0, 0^{'}, 0, (a - a^*), 0, m(a - a^*), 0^{'})$ for the controlled direct effect, $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the logarithm of the natural direct effect where letting

$$Q = \frac{\exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + \exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

$$B = \frac{\exp[\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + \exp[\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

then

$$\begin{aligned} d_1 &= Q - B \\ d_2 &= a^*(Q - B) \\ d_3 &= c^{'}(Q - B) \\ d_4 &= 0 \\ d_5 &= (a - a^*) \\ d_6 &= Q - B \\ d_7 &= aQ - a^*B \\ d_8 &= 0^{'} \end{aligned}$$

and $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$ for the logarithm of the natural indirect effect where letting

$$Q = \frac{\exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + \exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$

$$B = \frac{\exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + \exp[\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

$$K = \frac{\exp[\beta_0 + \beta_1 a + \beta_2^{'} c]}{\{1 + \exp[\beta_0 + \beta_1 a + \beta_2^{'} c]\}}$$

$$D = \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]}{\{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2^{'} c]\}}$$

then

$$\begin{aligned} d_1 &= (D + Q) - (K + B) \\ d_2 &= a^*[D - B] + a[Q - K] \end{aligned}$$

$$
\begin{aligned}
d_3 &= c^{'}[(D+Q)-(K+B)] \\
d_4 &= 0 \\
d_5 &= 0 \\
d_6 &= Q-B \\
d_7 &= a[Q-B] \\
d_8 &= 0^{'}
\end{aligned}
$$

Proof:
For the natural direct effect we have

$$
\begin{aligned}
&\exp\left[\log\left\{\frac{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}{P(Y_{a^*M_{a^*}}=1|c)/(1-P(Y_{a^*M_{a^*}}=1|c))}\right\}\right] \\
&= \exp\left[\text{logit}\{P(Y_{aM_{a^*}}=1|c)\}-\text{logit}\{P(Y_{a^*M_{a^*}}=1|c)\}\right] \\
&\approx \exp\left[\log\{P(Y_{aM_{a^*}}=1|c)\}-\log\{P(Y_{a^*M_{a^*}}=1|c)\}\right] \\
&= \exp\left[\log\{\textstyle\sum_m\{\mathbb{E}[Y|c,a,m]\}P(m|c,a^*)\}-\log\{\sum_m\{\mathbb{E}[Y|c,a^*,m]\}P(m|c,a^*)\}\right] \\
&\approx \exp\left[\log\left\{\frac{\exp(\theta_0+\theta_1 a+\theta_4^{'}c)+\exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{1+\exp(\beta_0+\beta_1 a^*+\beta_2^{'}C)}\right\}\right. \\
&\quad \left.-\log\left\{\frac{\exp(\theta_0+\theta_1 a^*+\theta_4^{'}c)+\exp(\theta_0+\theta_1 a^*+\theta_4^{'}c+\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{1+\exp(\beta_0+\beta_1 a^*+\beta_2^{'}c)}\right\}\right] \\
&= \left\{\frac{\exp(\theta_0+\theta_1 a+\theta_4^{'}c)+\exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{\exp(\theta_0+\theta_1 a^*+\theta_4^{'}c)+\exp(\theta_0+\theta_1 a^*+\theta_4^{'}c+\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c)}\right\} \\
&= \left\{\frac{\exp(\theta_1 a)(1+\exp[\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c])}{\exp(\theta_1 a^*)(1+\exp[\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2^{'}c])}\right\}
\end{aligned}
$$

where the fourth equality holds because, since the outcome is rare, we have $\mathbb{E}[Y|c,a,m]=\frac{\exp(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4^{'}c)}{1+\exp(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4^{'}c)}\approx\exp(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4^{'}c)$.
For the natural indirect effect we have

$$
\begin{aligned}
&\exp\left[\log\left\{\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\right\}\right] \\
&= \exp\left[\text{logit}\{P(Y_{aM_a}=1|c)\}-\text{logit}\{P(Y_{aM_{a^*}}=1|c)\}\right] \\
&\approx \exp\left[\log\{P(Y_{aM_a}=1|c)\}-log\{P(Y_{aM_{a^*}}=1|c)\}\right] \\
&= \exp\left[\log\left\{\textstyle\sum_m\{\mathbb{E}[Y|c,a,m]\}P(m|c,a)\right\}-\log\left\{\sum_m\{\mathbb{E}[Y|c,a,m]\}P(m|c,a^*)\right\}\right] \\
&\approx \exp\left[\log\left\{\frac{\exp(\theta_0+\theta_1 a+\theta_4^{'}c)+\exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2^{'}c)}{1+\exp[\beta_0+\beta_1 a+\beta_2^{'}c]}\right\}\right. \\
&\quad \left.-\log\left\{\frac{\exp(\theta_0+\theta_1 a+\theta_4^{'}c)+\exp(\theta_0+\theta_1 a+\theta_4^{'}c+\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2^{'}c)}{1+\exp[\beta_0+\beta_1 a^*+\beta_2^{'}c]}\right\}\right]
\end{aligned}
$$

$$= \frac{[1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)][1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2' c)]}{[1+\exp(\beta_0+\beta_1 a+\beta_2' c)][1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)]}$$

The standard errors follow by the delta method as in Propositions 2.3 and 2.4. ■

A.2.3. Equivalence of the Product and Difference Methods for a Continuous Outcome and for a Rare Binary Outcome

Proposition 2.7 (MacKinnon et al., 1995):
Suppose that the model for the mediator is

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

and the outcome follows the linear regression model:

$$\mathbb{E}[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_4' c$$

Suppose also a model is fit for the outcome with just the exposure, not the mediator:

$$\mathbb{E}[Y|a,m,c] = \phi_0 + \phi_1 a + \phi_4' c$$

The difference method uses $\phi_1 - \theta_1$ as a measure of the indirect effect; the product method uses $\beta_1\theta_2$. If all of the models are correctly specified, then product and difference methods coincide, that is, $\phi_1 - \theta_1 = \beta_1\theta_2$.

Proof:
By the model for the outcome without the mediator we have

$$\mathbb{E}[Y|a,c] = \phi_0 + \phi_1 a + \phi_4' c$$

and we also have

$$\begin{aligned}\mathbb{E}[Y|a,c] &= \mathbb{E}[\mathbb{E}[Y|a,M,c]] \\ &= \theta_0 + \theta_1 a + \theta_2 \mathbb{E}[M|a,c] + \theta_4' c \\ &= \theta_0 + \theta_1 a + \theta_2\{\beta_0 + \beta_1 a + \beta_2' c\} + \theta_4' c \\ &= \{\theta_0 + \theta_2\beta_0\} + \{\theta_1 + \theta_2\beta_1\}a + \{\theta_4' + \theta\beta_2'\}c\end{aligned}$$

Because this holds for all a, we must have $\phi_1 = \{\theta_1 + \theta_2\beta_1\}$ and thus $\phi_1 - \theta_1 = \theta_2\beta_1$. ■

Proposition 2.8 (VanderWeele and Vansteelandt, 2010):
Suppose that the model for the mediator is:

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

with M normally distributed conditional on A and C with conditional variance equal to some constant σ^2 and suppose the outcome follows an logistic regression model:

$$\text{logit}\{P(Y=1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$

Suppose also a model is fit for the outcome with just the exposure, not the mediator:

$$\text{logit}\{P(Y=1|a,c)\} = \phi_0 + \phi_1 a + \phi_4' c$$

The difference method uses $\phi_1 - \theta_1$ as a measure of the indirect effect; the product method uses $\beta_1\theta_2$. If the outcome is rare and all of the models are correctly specified, then product and difference methods approximately coincide, that is, $\phi_1 - \theta_1 \approx \beta_1\theta_2$.

Proof:
Under the rare outcome assumption, we must have $\phi_0 + \phi_1 a + \phi_2' c = \text{logit}(P(Y = 1|a,c)) \approx \log\{P(Y=1|a,c)\}$ and thus we have that

$$\begin{aligned}
&\exp\{\phi_0 + \phi_1 a + \phi_2' c\} \\
&\quad \approx P(Y=1|a,c) \\
&\quad = \mathbb{E}[P(Y=1|a,c,M)|a,c] \\
&\quad \approx \mathbb{E}[\exp\{\theta_0 + \theta_1 a + \theta_2 M + \theta_4' c\}|a,c] \\
&\quad = \exp(\theta_0 + \theta_1 a + \theta_4' c)\mathbb{E}[\exp(\theta_2 M)|a,c] \\
&\quad = \exp(\theta_0 + \theta_1 a + \theta_4' c)\exp\{\theta_2(\beta_0 + \beta_1 a + \beta_2' c) + \tfrac{1}{2}{\theta_2}^2\sigma^2\} \\
&\quad = \exp\{(\theta_0 + \tfrac{1}{2}{\theta_2}^2\sigma^2 + \beta_0\theta_2) + (\theta_1 + \theta_2\beta_1)a + (\theta_4 + \theta_2\beta_2)'c\}
\end{aligned}$$

Since this holds for all a, we must have that $\phi_1 \approx (\theta_1 + \theta_2\beta_1)$ and thus $\phi_1 - \theta_1 \approx \theta_2\beta_1$. ■

A.2.4. The Product Method as a Valid Test of the Presence of Any Mediated Effect

Proposition 2.9 (VanderWeele, 2011b):
Assume some model (M1) for the mediator conditional on the exposure A and the covariates C with a single parameter β_1 for the effect of the exposure on the mediator. Assume another model (M2) for the outcome conditional on the exposure A, the mediator M, and the covariates C with a single parameter θ_2 for the effect of the mediator on the outcome. Suppose, that on some causal diagram interpreted as non-parametric structural equation (cf. Pearl, 2009), assumptions (A2.1)–(A2.4) hold, then if $\beta_1\theta_2 \neq 0$, there must be individuals for whom there is a natural indirect effect.

Proof:
Suppose assumptions (A2.1)–(A2.4) hold; that is, for all a, a^*, m, (A2.1) $Y_{am} \perp\!\!\!\perp A|C$, (A2.2) $Y_{am} \perp\!\!\!\perp M|\{A,C\}$, (A2.3) $M_a \perp\!\!\!\perp A|C$, and (A2.4) $Y_{am} \perp\!\!\!\perp M_{a^*}|C$. On any causal diagram interpreted as nonparametric structural equation models for which (A2.4) holds, it also follows that $(Y_{am}, Y_{am^*}) \perp\!\!\!\perp (M_a, M_{a^*})|C$ (cf. Pearl, 2009). If in models (M1) and (M2) we have that $\theta_2\beta_1 \neq 0$, then from this it follows that $\theta_2 \neq 0$ and $\beta_1 \neq 0$. If $\beta_1 \neq 0$, then by assumption (A2.3) it follows that A has

an effect on M in the sense that for some a and a^* there are individuals $\omega \in \Theta_1$ such that $M_a(\omega) - M_{a^*}(\omega) \neq 0$. Let $m = M_a(\omega)$ and $m^* = M_{a^*}(\omega)$. If $\theta_2 \neq 0$, then by assumptions (A2.1) and (A2.2) it follows that M has an effect on Y with A fixed at a in the sense that there are individuals $\omega \in \Theta_2$ such that $Y_{am}(\omega) - Y_{am^*}(\omega) \neq 0$. Since $(Y_{am}, Y_{am^*}) \perp\!\!\!\perp (M_a, M_{a^*})|C$, it follows that there are individuals $\omega \in \Theta_1 \cap \Theta_2$ and thus for $\omega \in \Theta_1 \cap \Theta_2$, $0 \neq Y_{am}(\omega) - Y_{am^*}(\omega) = Y_{aM_a}(\omega) - Y_{aM_{a^*}}(\omega)$, that is, $Y_{aM_a}(\omega) \neq Y_{aM_{a^*}}(\omega)$, so there are some individuals for whom the natural indirect effect is nonzero. ■

A.3. SENSITIVITY ANALYSIS FOR MEDIATION

A.3.1. Sensitivity Analysis for Unmeasured Confounding for Total Effects on the Difference Scale

Suppose that Y is dichotomous, ordinal, or continuous and that A is categorical, ordinal, or continuous. For causal contrasts, we compare expected potential outcomes (i.e., counterfactual outcomes) for any two treatment levels, a and a^*, of A where a^* is taken as the reference. The average causal effect in the total population and among those receiving treatment $A = a$ or $A = a^*$ are given respectively by $\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})$, $\mathbb{E}(Y_a|a) - \mathbb{E}(Y_{a^*}|a)$ and $\mathbb{E}(Y_a|a^*) - \mathbb{E}(Y_{a^*}|a^*)$. The causal effects conditional on $C = c$ are given by $\mathbb{E}(Y_a|c) - \mathbb{E}(Y_{a^*}|c)$, $\mathbb{E}(Y_a|a,c) - \mathbb{E}(Y_{a^*}|a,c)$, and $\mathbb{E}(Y_a|a^*,c) - \mathbb{E}(Y_{a^*}|a^*,c)$. Suppose that the effect of A on Y is unconfounded given (U, C), where again U is unmeasured; that is, in counterfactual notation we assume that $Y_a \perp\!\!\!\perp A|C,U$. We then have that the true causal effects conditional on $C = c$ are given by

$$\mathbb{E}(Y_a|c) - \mathbb{E}(Y_{a^*}|c) = \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a^*,c,u)\}P(u|c)$$
$$\mathbb{E}(Y_a|a,c) - \mathbb{E}(Y_{a^*}|a,c) = \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a^*,c,u)\}P(u|c,a)$$
$$\mathbb{E}(Y_a|a^*,c) - \mathbb{E}(Y_{a^*}|a^*,c) = \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a^*,c,u)\}P(u|c,a^*)$$

The bias due to not controlling for the unmeasured confounder U is thus given by the difference between the observed average outcome differences, adjusted for C, and the true causal effect. Let $B_{add}(c)$, $B_a(c)$, and $B_{a^*}(c)$ denote the relevant bias when the target population is the total group, or those exposed to a or a^* respectively:

$$B_{add}(c) = \mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c) - \{\mathbb{E}(Y_a|c) - \mathbb{E}(Y_{a^*},c)\}$$
$$B_a(c) = \mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c) - \{\mathbb{E}(Y_a|a,c) - \mathbb{E}(Y_{a^*}|a,c)\}$$
$$B_{a^*}(c) = \mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c) - \{\mathbb{E}(Y_a|a^*,c) - \mathbb{E}(Y_{a^*}|a^*,c)\}$$

The marginal causal effects are given by adjusting for both C and U:

$$\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*}) = \sum_c \sum_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a^*,c,u)\}P(u|c)P(c)$$
$$\mathbb{E}(Y_a|a) - \mathbb{E}(Y_{a^*}|a) = \sum_c \sum_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a^*,c,u)\}P(u|c,a)P(c|a)$$
$$\mathbb{E}(Y_a|a^*) - \mathbb{E}(Y_{a^*}|a^*) = \sum_c \sum_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a^*,c,u)\}P(u|c,a^*)P(c|a^*)$$

If adjustment is made for C but not U, we would obtain the following expressions for the average outcome differences adjusted for C when the target population is the total group, or those exposed to a or a^*respectively:

$$\sum_c \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\}P(c)$$
$$\sum_c \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\}P(c|a)$$
$$\sum_c \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\}P(c|a^*)$$

The bias due to not controlling for the unmeasured confounder U is thus given by the difference between the observed average outcome differences, adjusted for C, and the true causal effect. Let B_{add}, B_a, and B_{a^*} denote the relevant bias when the target population is the total group, or those exposed to a or a^* respectively:

$$B_{add} = \sum_c \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\}P(c) - \{\mathbb{E}(Y_a) - \mathbb{E}(Y_{a^*})\}$$
$$B_a = \sum_c \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\}P(c|a) - \{\mathbb{E}(Y_a|a) - \mathbb{E}(Y_{a^*}|a)\}$$
$$B_{a^*} = \sum_c \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\}P(c|a^*) - \{\mathbb{E}(Y_a|a^*) - \mathbb{E}(Y_{a^*}|a^*)\}$$

Note that $B_{add} = \sum_c B_{add}(c)P(c)$, $B_a = \sum_c B_a(c)P(c|a)$, and $B_{a^*} = \sum_c B_{a^*}(c)$ $P(c|a^*)$.

Proposition 3.1 (VanderWeele and Arah, 2011):
If $Y_a \perp\!\!\!\perp A|C,U$, and if u' is any chosen reference value for the unmeasured confounder U, then the conditional bias formulae are given by

$$B_{add}(c) = \sum_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|c)\}$$
$$- \sum_u \{\mathbb{E}(Y|a^*,c,u) - \mathbb{E}(Y|a^*,c,u')\}\{P(u|a^*,c) - P(u|c)\}$$
$$B_a(c) = \sum_u \{\mathbb{E}(Y|a^*,c,u) - \mathbb{E}(Y|a^*,c,u')\}\{P(u|a,c) - P(u|a^*,c)\}$$
$$B_{a^*}(c) = \sum_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|a^*,c)\}$$

and the marginal bias formulae are given by $B_{add} = \sum_c B_{add}(c)P(c)$, $B_a = \sum_c B_a(c)P(c|a)$, and $B_{a^*} = \sum_c B_{a^*}(c)P(c|a^*)$.

Proof of Proposition 3.1
We have that

$$\begin{aligned}
B_a(c) &= \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\} - \{\mathbb{E}(Y_a|a,c) - \mathbb{E}(Y_{a^*}|a,c)\} \\
&= \sum\nolimits_u \mathbb{E}(Y|a,c,u)P(u|a,c) - \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c) \\
&\quad - \sum\nolimits_u \mathbb{E}(Y_a|a,c,u)P(u|a,c) + \sum\nolimits_u \mathbb{E}(Y_{a^*}|a,c,u)P(u|a,c)P \\
&= \sum\nolimits_u \mathbb{E}(Y_{a^*}|a,c,u)P(u|a,c) - \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c) \\
&= \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)P(u|a,c) - \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c) \\
&= \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)\{P(u|a,c) - P(u|a^*,c)\} \\
&= \sum\nolimits_u \{\mathbb{E}(Y|a^*,c,u) - \mathbb{E}(Y|a^*,c,u')\}\{P(u|a,c) - P(u|a^*,c)\}
\end{aligned}$$

The proof for $B_{a^*}(c)$ is similar. For $B_{add}(c)$ we have that

$$\begin{aligned}
B_{add}(c) &= \{\mathbb{E}(Y|a,c) - \mathbb{E}(Y|a^*,c)\} - \{\mathbb{E}(Y_a|c) - \mathbb{E}(Y_{a^*}|c)\} \\
&= \sum\nolimits_u \mathbb{E}(Y|a,c,u)P(u|a,c) - \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c) \\
&\quad - \sum\nolimits_u \mathbb{E}(Y_a|c,u)P(u|c) + \sum\nolimits_u \mathbb{E}(Y_{a^*}|c,u)P(u|c) \\
&= \sum\nolimits_u \mathbb{E}(Y|a,c,u)P(u|a,c) - \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c) \\
&\quad - \sum\nolimits_u \mathbb{E}(Y_a|a,c,u)P(u|c) + \sum\nolimits_u \mathbb{E}(Y_{a^*}|a^*,c,u)P(u|c) \\
&= \sum\nolimits_u \mathbb{E}(Y|a,c,u)\{P(u|a,c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \mathbb{E}(Y|a^*,c,u)\{P(u|a^*,c) - P(u|c)\} \\
&= \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \{\mathbb{E}(Y|a^*,c,u) - \mathbb{E}(Y|a^*,c,u')\}\{P(u|a^*,c) - P(u|c)\}
\end{aligned}$$

Since $B_{add} = \sum_c B_{add}(c)P(c)$, $B_a = \sum_c B_a(c)P(c|a)$, and $B_{a^*} = \sum_c B_{a^*}(c)P(c|a^*)$, this completes the proof. ■

In order to use these bias formulaes, one would need to specify (i) the relation between U and Y, among those with treatment level $A = a$ and $A = a^*$, conditional on C, that is, $\{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}$ and $\{\mathbb{E}(Y|a^*,c,u) - \mathbb{E}(Y|a^*,c,u')\}$, and (ii) how the distribution of the unmeasured confounder U among those with treatment level $A = a$ and $A = a^*$ compares with the overall distribution of U, conditional on C, that is, $\{P(u|a,c) - P(u|c)\}$ and $\{P(u|a^*,c) - P(u|c)\}$. Once the bias factor is calculated it can be subtracted from the estimate controlling only for measured covariates C to obtain a corrected estimator. Proposition 3.1 is quite general and encompasses a number of more specific sensitivity-analysis techniques in the

literature (cf. VanderWeele and Arah, 2011). It does not assume any particular method, model, or functional form assumptions.

Corollary (VanderWeele and Arah, 2011). If $Y_a \perp\!\!\!\perp A|C,U$ and if u' is any chosen reference value for the unmeasured confounder U and the relationship between U and Y—that is, $\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')$—does not vary across strata of A, then

$$B_{add}(c) = \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|a^*,c)\}$$
$$B_{add} = \sum\nolimits_c \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|a^*,c)\}P(c)$$

If in addition U is binary, then

$$B_{add}(c) = \delta(c)\gamma\,(c)$$
$$B_{add} = \sum\nolimits_c \delta(c)\gamma\,(c)P(c)$$

where $\delta(c) = \mathbb{E}(Y|a,c,U=1) - \mathbb{E}(Y|a,c,U=0)$ and $P(U=1|a,c) - P(U=0|a^*,c)$.

Proof: If $\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')$ does not vary across strata of A, then

$$\begin{aligned} B_{add}(c) &= \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|c)\} \\ &\quad - \sum\nolimits_u \{\mathbb{E}(Y|a^*,c,u) - \mathbb{E}(Y|a^*,c,u')\}\{P(u|a^*,c) - P(u|c)\} \\ &= \sum\nolimits_u \{\mathbb{E}(Y|a,c,u) - \mathbb{E}(Y|a,c,u')\}\{P(u|a,c) - P(u|a^*,c)\} \end{aligned}$$

If U is binary and we take $u' = 0$ as the reference value, this becomes $\{\mathbb{E}(Y|a,c,U=1) - \mathbb{E}(Y|a,c,U=0)\}\{P(U=1|a,c) - P(U=1|a^*,c)\}$. ■

This simple formula $B_{add}(c) = \delta(c)\gamma\,(c)$ has been obtained previously (Cochran, 1938; Draper and Smith, 1981; Lin et al., 1998) but under much stronger assumptions. Note that the corollary above does not assume any particular method, model, or functional form assumptions. Note that, for the causal effect conditional on $C = c$, under the simplifying assumptions of the corollary, once the sensitivity parameters are specified, the standard error of the bias-corrected estimator is the same as that of the original estimator since the sensitivity parameters are fixed. Because the standard errors of the original and bias-adjusted estimates are the same, the bias factor can be subtracted not only from the estimate itself but also from both limits of a confidence interval to obtain a corrected confidence interval.

A.3.2. Sensitivity Analysis for Unmeasured Confounding for a Total Effect on a Ratio Scale

For binary Y, other measures of effect such as the risk ratio or odds ratio may be of interest. The conditional causal risk ratio in the total population or among

those receiving treatment a or a^* are defined respectively by $\mathbb{E}(Y_a|c)/\mathbb{E}(Y_{a^*}|c)$, $\mathbb{E}(Y_a|a,c)/\mathbb{E}(Y_{a^*}|a,c)$, and $\mathbb{E}(Y_a|a^*,c)/\mathbb{E}(Y_{a^*}|a^*,c)$. The conditional causal odds ratios in the total population or among those receiving treatment a or a^* can be defined similarly; for example, the conditional causal odds ratio in the total population is defined by

$$\frac{\mathbb{E}(Y_a|c)/\{1-\mathbb{E}(Y_a|c)\}}{\mathbb{E}(Y_{a^*}|c)/\{1-\mathbb{E}(Y_{a^*}|c)\}}$$

We can define bias factors $B_{mult}(c)$, $B_a^{RR}(c)$ and $B_{a^*}^{RR}(c)$ corresponding to the ratios between the risk ratios conditional on C and the true conditional causal risk ratios respectively in the total population or among those receiving treatment a or a^*:

$$B_{mult}(c) = \frac{\mathbb{E}(Y|a,c)/\mathbb{E}(Y|a^*,c)}{\mathbb{E}(Y_a|c)/\mathbb{E}(Y_{a^*}|c)}$$

$$B_a^{RR}(c) = \frac{\mathbb{E}(Y|a,c)/\mathbb{E}(Y|a^*,c)}{\mathbb{E}(Y_a|a,c)/\mathbb{E}(Y_{a^*}|a,c)}$$

$$B_{a^*}^{RR}(c) = \frac{\mathbb{E}(Y|a,c)/\mathbb{E}(Y|a^*,c)}{\mathbb{E}(Y_a|a^*,c)/\mathbb{E}(Y_{a^*}|a^*,c)}$$

We can define the conditional odds-ratio bias factors analogously. For the bias factors for the conditional causal risk ratio, we then have the following result, expressing the biases in terms of the relationship between the unmeasured confounder(s) U and the outcome Y and the relationship between U and treatment A.

Proposition 3.2 (VanderWeele and Arah, 2011):
If $Y_a \perp\!\!\!\perp A|C,U$ and if u' is any chosen reference value for the unmeasured confounder U, then

$$B_{mult}(c) = \frac{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|a,c)}{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|c)} \Big/ \frac{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a^*,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|c)}$$

$$B_a^{RR}(c) = \frac{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a^*,c)}$$

$$B_{a^*}^{RR}(c) = \frac{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|a,c)}{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|a^*,c)}$$

Proof of Proposition 3.2
We have that

$$B_a^{RR}(c) = \frac{\mathbb{E}(Y|a,c)/\mathbb{E}(Y|a^*,c)}{\mathbb{E}(Y_a|a,c)/\mathbb{E}(Y_{a^*}|a,c)}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}(Y_{a^*}|a,c)}{\mathbb{E}(Y|a^*,c)} \\
&= \frac{\sum_u \mathbb{E}(Y_{a^*}|a,c,u)P(u|a,c)}{\sum_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c)} \\
&= \frac{\sum_u \mathbb{E}(Y|a^*,c,u)P(u|a,c)}{\sum_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c)} \\
&= \frac{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a^*,c)}
\end{aligned}
$$

The proof for $B^{RR}_{a^*}(c)$ is similar. For $B_{mult}(c)$ we have that

$$
\begin{aligned}
B_{mult}(c) &= \frac{\mathbb{E}(Y|a,c)/\mathbb{E}(Y|a^*,c)}{\mathbb{E}(Y_a|c)/\mathbb{E}(Y_{a^*}|c)} \\
&= \frac{\sum_u \mathbb{E}(Y|a,c,u)P(u|a,c)}{\sum_u \mathbb{E}(Y_a|c,u)P(u|c)} \Big/ \frac{\sum_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c)}{\sum_u \mathbb{E}(Y_{a^*}|c,u)P(u|c)} \\
&= \frac{\sum_u \mathbb{E}(Y|a,c,u)P(u|a,c)}{\sum_u \mathbb{E}(Y|a,c,u)P(u|c)} \Big/ \frac{\sum_u \mathbb{E}(Y|a^*,c,u)P(u|a^*,c)}{\sum_u \mathbb{E}(Y|a^*,c,u)P(u|c)} \\
&= \frac{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|a,c)}{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|c)} \Big/ \frac{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a^*,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|c)}.
\end{aligned}
$$

■

Corollary (Schlesselman, 1978). If $Y_a \perp\!\!\!\perp A|C,U$ and if U is binary and $\gamma = \mathbb{E}(Y|a,c,U=1)/\mathbb{E}(Y|a,c,U=0)$ does not vary across strata of a, then

$$
B_{mult}(c) = \frac{1+(\gamma-1)P(U=1|a,c)}{1+(\gamma-1)P(U=1|a^*,c)}
$$

Proof:
We have

$$
\begin{aligned}
&\frac{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|a,c)}{\sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|c)} \Big/ \frac{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a^*,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|c)} \\
&= \sum_u \frac{\mathbb{E}(Y|a,c,u)}{\mathbb{E}(Y|a,c,u')}P(u|a,c) \Big/ \sum_u \frac{\mathbb{E}(Y|a^*,c,u)}{\mathbb{E}(Y|a^*,c,u')}P(u|a^*,c) \\
&= \frac{\gamma\, P(U=1|a,c)+P(U=0|a,c)}{\gamma\, P(U=1|a^*,c)+P(U=0|a^*,c)} \\
&= \frac{1+(\gamma-1)P(U=1|a,c)}{1+(\gamma-1)P(U=1|a^*,c)}.
\end{aligned}
$$

■

Once we have calculated the bias factor $B_{mult}(c)$, we can simply estimate our risk ratio controlling only for C and we divide our estimate by $B_{mult}(c)$ to get the corrected estimate for risk ratio—that is, what we would have obtained if we had

adjusted for U a well. Under the simplifying assumptions of the corollary, we can also obtain corrected confidence intervals by dividing both limits of the confidence interval by $B_{mult}(c)$ since the bias factor depends only on the sensitivity analysis parameters and not the data. The results also hold approximately on the odds ratio scale if the outcome is rare. For further results on marginal risk ratios and odds ratio with a common outcome, see VanderWeele and Arah (2011).

A.3.3. Sensitivity Analysis for Unmeasured Confounding for Controlled Direct Effects

We consider conditional controlled direct effects, $\mathbb{E}[Y_{am} - Y_{a^*m}|c]$, and let $B^{CDE}_{add}(m|c)$ denote the difference between (i) the estimate of the controlled direct effect conditional on C, that is, $\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]$, and (ii) the true controlled direct effect:

$$B^{CDE}_{add}(m|c) = \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\} - \mathbb{E}[Y_{am} - Y_{a^*m}|c]$$

Proposition 3.3 (VanderWeele, 2010a):
Suppose that for all a and m, $Y_{am} \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$ then for any reference level u' of U we have $B^{CDE}_{add}(m|c) =$

$$\sum_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|a,c)\}$$
$$-\sum_u \{\mathbb{E}[Y|a^*,m,c,u] - \mathbb{E}[Y|a^*,m,c,u']\}\{P(u|a^*,m,c) - P(u|a^*,c)\}$$

Moreover, if U is binary, with $U \perp\!\!\!\perp A|C$ and for a particular value m, $\gamma_m = \mathbb{E}(Y|a,c,m,U=1) - \mathbb{E}(Y|a,c,m,U=0)$ is constant across strata of a, then

$$B^{CDE}_{add}(m|c) = \delta_m \gamma_m$$

where $\delta_m = P(U=1|a,m,c) - P(U=1|a^*,m,c)$.

Proof:
We have that,

$$\begin{aligned}
B^{CDE}_{add}(m|c) &= \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\} - \mathbb{E}[Y_{am} - Y_{a^*m}|c] \\
&= \sum_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c) - \sum_u \mathbb{E}[Y|a^*,m,c,u]P(u|a^*,m,c) \\
&\quad - \sum_u \mathbb{E}[Y|a,m,c,u]P(u|a,c) + \sum_u \mathbb{E}[Y|a^*,m,c,u]P(u|a^*,c) \\
&= \sum_u \mathbb{E}[Y|a,m,c,u]\{P(u|a,m,c) - P(u|a,c)\} \\
&\quad - \sum_u \mathbb{E}[Y|a^*,m,c,u]\{P(u|a^*,m,c) - P(u|a^*,c)\} \\
&= \sum_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|a,c)\} \\
&\quad - \sum_u \{\mathbb{E}[Y|a^*,m,c,u] - \mathbb{E}[Y|a^*,m,c,u']\}\{P(u|a^*,m,c) - P(u|a^*,c)\}
\end{aligned}$$

where the second equality follows because $Y_{am} \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$ and the final equality follows because for a fixed reference value u' of U, $\mathbb{E}[Y|a,m,c,u']$ and $\mathbb{E}[Y|a^*,m,c,u']$ are constants and thus $\sum_u \mathbb{E}[Y|a,m,c,u']P(u|a,m,c) = \mathbb{E}[Y|a,m,c,u'] = \sum_u \mathbb{E}[Y|a,m,c,u']P(u|a,c)$ and similarly $\sum_u \mathbb{E}[Y|a^*,m,c,u']P(u|a^*,m,c) = \mathbb{E}[Y|a^*,m,c,u'] = \sum_u \mathbb{E}[Y|a^*,m,c,u']$ $P(u|a^*,c)$.

If U is binary with $U \perp\!\!\!\perp A|C$ and if $\gamma_m = \mathbb{E}(Y|a,c,m,U=1) - \mathbb{E}(Y|a,c,m,U=0)$ is constant across strata of a, let $u' = 0$. We have

$$\begin{aligned}
B^{CDE}_{add}(m|c) &= \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|a,c)\} \\
&\quad - \sum\nolimits_u \{\mathbb{E}[Y|a^*,m,c,u] - \mathbb{E}[Y|a^*,m,c,u']\}\{P(u|a^*,m,c) \\
&\quad - P(u|a^*,c)\} \\
&= \{\mathbb{E}[Y|a,m,c,U=1] - \mathbb{E}[Y|a,m,c,U=0]\}\{P(U=1|a,m,c) \\
&\quad - P(U=1|a,c)\} - \{\mathbb{E}[Y|a^*,m,c,U=1] \\
&\quad - \mathbb{E}[Y|a^*,m,c,U=0]\}\{P(U=1|a^*,m,c) - P(U=1|a^*,c)\} \\
&= \gamma_m\{P(U=1|a,m,c) - P(U=1|c)\} - \gamma_m\{P(U=1|a^*,m,c) \\
&\quad - P(U=1|c)\} \\
&= \gamma_m\delta_m. \quad \blacksquare
\end{aligned}$$

On the risk ratio scale (which will also approximate odds ratios if the outcome is rare) the bias factor compares the true controlled direct effect risk ratio, $P(Y_{am}|c)/P(Y_{a^*m}|c)$, with its estimate, $P(Y|a^*,m,c)/P(Y|a,m,c)$, that is,

$$B^{CDE}_{mult}(m|c) = \frac{P(Y|a,m,c)}{P(Y|a^*,m,c)} / RR^{CDE}_{a,a^*|c}(m)$$

Proposition 3.4 (VanderWeele, 2010a): Suppose that for all a and m, $Y_{am} \perp\!\!\!\perp A|C$ and $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$ then for any reference level u' of U we have that

$$B^{CDE}_{mult}(m|c) = \frac{\sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|a,m,c)}{\sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|a,c)} \Big/ \frac{\sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|a^*,m,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|a^*,c)}$$

If U is binary with $U \perp\!\!\!\perp A|C$ and if $\gamma_m = \frac{P(Y|a,m,c,U=1)}{P(Y|a,m,c,U=0)} =$ is constant across strata of a, then

$$B^{CDE}_{mult}(m|c) = \frac{1 + (\gamma_m - 1)P(U=1|a,m,c)}{1 + (\gamma_m - 1)P(U=1|a^*,m,c)}$$

Proof: We have for any reference level u' that

$$\begin{aligned}
B^{CDE}_{mult}(m|c) &= \frac{\mathbb{E}(Y|a,m,c)/\mathbb{E}(Y|a^*,m,c)}{\mathbb{E}(Y_{am}|c)/\mathbb{E}(Y_{a^*m}|c)} \\
&= \frac{\sum_u \mathbb{E}(Y|a,m,c,u)P(u|a,m,c)}{\sum_u \mathbb{E}(Y_{am}|c,u)P(u|a,c)} \Big/ \frac{\sum_u \mathbb{E}(Y|a^*,m,c,u)P(u|a^*,m,c)}{\sum_u \mathbb{E}(Y_{a^*m}|c,u)P(u|a^*,c)}
\end{aligned}$$

$$= \frac{\sum_u \mathbb{E}(Y|a,m,c,u)P(u|a,m,c)}{\sum_u \mathbb{E}(Y|a,m,c,u)P(u|a,c)} \Big/ \frac{\sum_u \mathbb{E}(Y|a^*,m,c,u)P(u|a^*,m,c)}{\sum_u \mathbb{E}(Y|a^*,m,c,u)P(u|a^*,c)}$$
$$= \frac{\sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|a,m,c)}{\sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|a,c)} \Big/ \frac{\sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|a^*,m,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|a^*,c)}$$

If U is binary with $U \perp\!\!\!\perp A|C$ and if $\gamma_m = \frac{P(Y|a,m,c,U=1)}{P(Y|a,m,c,U=0)} =$ is constant across strata of a, we let $u' = 0$ and we have

$$B^{CDE}_{mult}(m|c) = \frac{\sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|a,m,c)}{\sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|c)} \Big/ \frac{\sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|a^*,m,c)}{\sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|c)}$$
$$= \sum_u \frac{\mathbb{E}(Y|a,m,c,u)}{\mathbb{E}(Y|a,m,c,u')}P(u|a,m,c) \Big/ \sum_u \frac{\mathbb{E}(Y|a^*,m,c,u)}{\mathbb{E}(Y|a^*,m,c,u')}P(u|a^*,m,c)$$
$$= \frac{\gamma\, P(U=1|a,m,c) + P(U=0|a,m,c)}{\gamma\, P(U=1|a^*,m,c) + P(U=0|a^*,m,c)}$$
$$= \frac{1+(\gamma-1)P(U=1|a,m,c)}{1+(\gamma-1)P(U=1|a^*,m,c)}. \quad \blacksquare$$

A.3.4. Sensitivity Analysis for Unmeasured Confounding for Natural Direct and Indirect Effects

For natural direct and indirect effects we define the bias factor on the difference scale, conditional on covariates $C = c$ by

$$B^{NDE}_{add}(c) = \sum_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c) - \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]$$
$$B^{NIE}_{add}(c) = \sum_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\} - \mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c]$$

Proposition 3.5 (VanderWeele, 2010a):
Suppose $Y_{am} \perp\!\!\!\perp A|C$, $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$, $M_a \perp\!\!\!\perp A|C$, $Y_{am} \perp\!\!\!\perp M_{a^*}|\{C,U\}$, and $U \perp\!\!\!\perp A|C$, then for any reference level u' of U the bias formula for the natural direct effect is given by

$$B^{NDE}_{add}(c) = \sum_m \sum_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|a^*,m,c)\}P(m|a^*,c)$$

and the bias formula for the natural indirect effect is given by

$$B^{NDE}_{add}(c) = -\sum_m \sum_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|a^*,m,c)\}P(m|a^*,c)$$

Proof:
For the natural direct effect we have that

$$
\begin{aligned}
B_{add}^{NDE}(c) &= \sum\nolimits_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}P(m|a^*,c) - \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]P(u|a^*,m,c)P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(m|a^*,c,u)P(u|c) \\
&\quad + \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c) \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]\frac{P(m|a^*,c,u)P(u|a^*,c)}{P(m|a^*,c)}P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\frac{P(u|a^*,m,c)P(m|a^*,c)}{P(u|a^*,c)}P(u|c) \\
&\quad + \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c) \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\frac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}P(m|a^*,c) \\
&\quad + \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c) \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\{P(u|a,m,c) - \frac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a^*,m,c,u]\{P(u|a^*,c) - P(u|c)\}P(m|a^*,c,u) \\
&= \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) \\
&\quad - \frac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a^*,c) \\
&\quad - P(u|c)\}P(m|a^*,c,u)
\end{aligned}
$$

where the second equality follows because $Y_{am} \perp\!\!\!\perp A|\{C,U\}$, $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$, $M_a \perp\!\!\!\perp A|\{C,U\}$ and $Y_{am} \perp\!\!\!\perp M_{a^*}|\{C,U\}$. Since, $U \perp\!\!\!\perp A|C$ we have

$$
B_{add}^{NDE}(c) = \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c)
$$

$$
\begin{aligned}
&- \frac{P(u|a^*,m,c)P(u|c)}{P(u|c)}\}P(m|a^*,c) \\
&- \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|c) \\
&- P(u|c)\}P(m|a^*,c,u) \\
&= \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) \\
&- P(u|a^*,m,c)\}P(m|a^*,c).
\end{aligned}
$$

For the natural indirect effects we have that

$$
\begin{aligned}
B_{add}^{NIE}(c) &= \sum\nolimits_m \mathbb{E}[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\} - \mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c)\{P(m|a,c) - P(m|a^*,c)\} \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(m|a,c,u)P(u|c) \\
&\quad + \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(m|a^*,c,u)P(u|c) \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c)P(m|a,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)\} \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\frac{P(u|a,m,c)}{P(u|a,c)}P(m|a,c)P(u|c) \\
&\quad + \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\frac{P(u|a^*,m,c)}{P(u|a^*,c)}P(m|a^*,c)P(u|c) \\
&= \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\{P(u|a,m,c) - \frac{P(u|c)}{P(u|a,c)}P(u|a,m,c)\}P(m|a,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \mathbb{E}[Y|a,m,c,u]\{P(u|a,m,c) \\
&\quad - \frac{P(u|c)}{P(u|a^*,c)}P(u|a^*,m,c)\}P(m|a^*,c) \\
&= \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) \\
&\quad - \frac{P(u|c)}{P(u|a,c)}P(u|a,m,c)\}P(m|a,c) \\
&\quad - \sum\nolimits_m \sum\nolimits_u \{\mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u']\}\{P(u|a,m,c) \\
&\quad - \frac{P(u|c)}{P(u|a^*,c)}P(u|a^*,m,c)\}P(m|a^*,c)
\end{aligned}
$$

where the second equality follows because $Y_{am} \perp\!\!\!\perp A|\{C,U\}$, $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$, $M_a \perp\!\!\!\perp A|\{C,U\}$ and $Y_{am} \perp\!\!\!\perp M_{a^*}|\{C,U\}$. Since $U \perp\!\!\!\perp A|C$ we have

$$\begin{aligned}
B_{add}^{NIE}(c) &= \sum_m \sum_u \mathbb{E}[Y|a,m,c,u] \left\{ P(u|a,m,c) - \frac{P(u|c)}{P(u|c)} P(u|a,m,c) \right\} P(m|a,c) \\
&\quad - \sum_m \sum_u \mathbb{E}[Y|a,m,c,u] \\
&\quad \times \left\{ P(u|a,m,c) - \frac{P(u|c)}{P(u|c)} P(u|a^*,m,c) \right\} P(m|a^*,c) \\
&= - \sum_m \sum_u \mathbb{E}[Y|a,m,c,u] \{ P(u|a,m,c) - P(u|a^*,m,c) \} P(m|a^*,c) \\
&= - \sum_m \sum_u \{ \mathbb{E}[Y|a,m,c,u] - \mathbb{E}[Y|a,m,c,u'] \} \{ P(u|a,m,c) \\
&\quad - P(u|a^*,m,c) \} P(m|a^*,c).
\end{aligned}$$

This completes the proof. ■

Proposition 3.6:
Consider a binary outcome and binary mediator and with regression models

$$\begin{aligned}
\text{logit}\{P(Y=1|a,m,c)\} &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\
\text{logit}\{P(M=1|a,c)\} &= \beta_0 + \beta_1 a + \beta_2' c
\end{aligned}$$

Suppose now that there were an unmeasured binary U that confounded only the mediator–outcome relationship so that $Y_{am} \perp\!\!\!\perp A|C$, $Y_{am} \perp\!\!\!\perp M|\{A,C,U\}$, $M_a \perp\!\!\!\perp A|C$, $Y_{am} \perp\!\!\!\perp M_{a^*}|\{C,U\}$, and $U \perp\!\!\!\perp A|C$ and such that U had constant effect over a,m,c and let

$$\begin{aligned}
\gamma &= \frac{P(Y=1|a,m,c,U=1)}{P(Y=1|a,m,c,U=0)} \\
B_0 &= \frac{1+(\gamma-1)P(U=1|a,M=0,c)}{1+(\gamma-1)P(U=1|a^*,M=0,c)} \\
B_1 &= \frac{1+(\gamma-1)P(U=1|a,M=1,c)}{1+(\gamma-1)P(U=1|a^*,M=1,c)} \\
B_2 &= \frac{1+(\gamma-1)P(U=1|a^*,M=1,c)}{1+(\gamma-1)P(U=1|a^*,M=0,c)}
\end{aligned}$$

Then let

$$\begin{aligned}
\theta_1^\dagger &= \theta_1 - \log(B_0) \\
\theta_2^\dagger &= \theta_2 - \log(B_2) \\
\theta_3^\dagger &= \theta_3 - \log(B_1) + \log(B_0)
\end{aligned}$$

Then corrected natural direct and indirect effects are given by

$$\log\{OR^{NDE}\} = \frac{\exp(\theta_1^\dagger a)\{1+\exp(\theta_2^\dagger+\theta_3^\dagger a+\beta_0+\beta_1 a^*+\beta_2' c)\}}{\exp(\theta_1^\dagger a^*)\{1+\exp(\theta_2^\dagger+\theta_3^\dagger a^*+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

$$\log\{OR^{NIE}\} = \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\theta_2^\dagger+\theta_3^\dagger a+\beta_0+\beta_1 a+\beta_2' c)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\theta_2^\dagger+\theta_3^\dagger a+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

Proof:
For a binary outcome and a binary mediator, we have by Proposition 3.4, for the true controlled direct effect with $m=0$, namely $\exp(\theta_1^\dagger)$, that $\exp(\theta_1)/\exp(\theta_1^\dagger) = B_0$. From this it follows that $\theta_1^\dagger = \theta_1 - \log(B_0)$. By Proposition 3.4, for the true controlled direct effect with $m=1$, $\exp(\theta_1^\dagger+\theta_3^\dagger)$, we have that $\exp(\theta_1+\theta_3)/\exp(\theta_1^\dagger+\theta_3^\dagger) = B_1$ and thus $\theta_3^\dagger = \theta_1+\theta_3-\theta_1^\dagger-\log(B_1) = \theta_1+\theta_3-\{\theta_1-\log(B_0)\}-\log(B_1) = \theta_3-\log(B_1)+\log(B_0)$. Finally, by the corollary of Proposition 3.2, we have that $\exp(\theta_2)/\exp(\theta_2^\dagger) = B_2$ from which it follows $\theta_2^\dagger = \theta_2 - \log(B_2)$. This, in conjunction with the formulaes for binary mediator and outcomes in Proposition 2.6, then gives the result. ■

Proposition 3.7 (Imai et al., 2010a):
Suppose the exposure A is binary and the following models are fit to the data:

$$\begin{aligned} Y &= \phi_0 + \phi_1 A + \phi_4' C + \epsilon_1 \\ M &= \beta_0 + \beta_1 A + \beta_2' C + \epsilon_2 \\ Y &= \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4' C + \epsilon_3 \end{aligned}$$

and that the error terms ϵ_2 and ϵ_3 are correlated with correlation ρ then if assumptions (A2.1) and (A2.3) hold the natural indirect effect is given by

$$\mathbb{E}[Y_{1M_1} - Y_{1M_0}] = \frac{\beta_1\sigma_1(1)}{\sigma_2(1)}(\widetilde{\rho}_a - \rho\sqrt{\{1-\widetilde{\rho}_a^2\}/(1-\rho^2)})$$

where $\sigma_i(a) = \mathrm{Var}(\epsilon_i|A=a)$ and $\widetilde{\rho}_a = \mathrm{Cov}(\epsilon_1,\epsilon_2|A=a)$.

Proof:
Under assumptions (A2.1) and (A2.3), we can identify ϕ_0, ϕ_1, ϕ_4' and $\beta_0, \beta_1 A, \beta_2' C$. From $M = \beta_0 + \beta_1 A + \beta_2' C + \epsilon_2$ and $Y = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4' C + \epsilon_3$ we have that

$$\begin{aligned} Y &= \theta_0 + \theta_1 A + \theta_2(\beta_0 + \beta_1 A + \beta_2' C + \epsilon_2) \\ &\quad + \theta_3 A(\beta_0 + \beta_1 A + \beta_2' C + \epsilon_2) + \theta_4' C + \epsilon_3 \\ &= (\theta_0 + \theta_2\beta_0) + (\theta_1 + \theta_2\beta_1 + \theta_3\beta_0 + \theta_3\beta_1)A \\ &\quad + (\theta_2\beta_2' + \theta_4')C + \theta_3\beta_2' AC + (\theta_2 + \theta_3 A)\epsilon_2 + \epsilon_3 \end{aligned}$$

Under assumptions (A2.1) and (A2.3) we have $0 = \mathbb{E}[\epsilon_1|A=1] = (\theta_2+\theta_3)\mathbb{E}[\epsilon_2|A=1] + \mathbb{E}[\epsilon_3|A=1]$. We thus have $\phi_0 = (\theta_0+\theta_2\beta_0)$, $\phi_1 = (\theta_1+$

$\theta_2\beta_1 + \theta_3\beta_0 + \theta_3\beta_1)$, and $\phi_4' = (\theta_2\beta_2' + \theta_4')$. If we can identify θ_2 and θ_3, then we would have $\theta_0 = \phi_0 - \theta_2\beta_0, \theta_1 = \phi_1 - (\theta_2\beta_1 + \theta_3\beta_0 + \theta_3\beta_1), \theta_4' = \phi_4' - \theta_2\beta_2'$.

We moreover have that $\sigma_1(a) = \text{Var}(\epsilon_1|A = a) = \text{Var}\{(\theta_2 + \theta_3 a)\epsilon_2 + \epsilon_3|A = a\} = (\theta_2 + \theta_3 a)^2\sigma_2(a)^2 + (\theta_2 + \theta_3 a)\sigma_2(a)\sigma_3(a)\widetilde{\rho}_a + \sigma_3(a)^2$, which gives us two equations, one for $a = 1$ and one for $a = 0$. We likewise have that $\widetilde{\rho}_a\sigma_1(a)\sigma_2(a) = \text{Cov}(\epsilon_1, \epsilon_2|A = a) = \text{Cov}\{(\theta_2 + \theta_3 a)\epsilon_2 + \epsilon_3, \epsilon_2|A = a\} = (\theta_2 + \theta_3 a)\sigma_2(a)^2 + \rho\sigma_2(a)\sigma_3(a)$, which again gives us two equations, one for $a = 1$ and one for $a = 0$. Since $\widetilde{\rho}_a, \sigma_1(a), \sigma_2(a)$ can be estimated from the data, once ρ is fixed we have four equations with four unknowns: $\theta_2, \theta_3, \sigma_3(0), \sigma_3(1)$. Solving these equations and then using $\mathbb{E}[Y_{1M_1} - Y_{1M_0}] = (\theta_2\beta_1 + \theta_3\beta_1)$ gives the result. ■

Proposition 3.8 (Emsley and VanderWeele, 2014): Suppose A and M are binary and that A is randomized and that conditional on some set of possibly unmeasured baseline variables W (A2.1)–(A2.4) hold. Suppose further that the mediator M does not interact with the exposure A or the confounders W in the sense that the following model is correctly specified: $\mathbb{E}[Y|a, m, w] = f(a, w) + g(m)$. Let $\gamma = \mathbb{E}[Y_{m=1} - Y_{m=0}]$ denote the effect of M on Y obtained from a different secondary study in which M was randomized. If $\mathbb{E}[Y_{m=1} - Y_{m=0}]$ is the same in the two studies, then in the primary study we have that the natural indirect and direct effects are given by

$$\begin{aligned}\mathbb{E}[Y_{1M_1} - Y_{1M_0}] &= \gamma\{\mathbb{E}(M|A = 1) - \mathbb{E}(M|A = 0)\}\\ \mathbb{E}[Y_{1M_0} - Y_{0M_0}] &= \{\mathbb{E}(Y|A = 1) - \mathbb{E}(Y|A = 0)\} - \gamma\{\mathbb{E}(M|A = 1)\\ &\quad - \mathbb{E}(M|A = 0)\}\end{aligned}$$

where the expectations are taken in the primary study.

Proof:
Since M is binary, without loss of generality we may assume $\mathbb{E}[Y|a, m, w] = f(a, w) + \theta m$ for some θ. Moreover, we have in the primary study

$$\begin{aligned}\mathbb{E}[Y_{m=1} - Y_{m=0}] &= \mathbb{E}[Y_{m=1}|a] - \mathbb{E}[Y_{m=0}|a]\\ &= \mathbb{E}[Y_{am=1}|a] - \mathbb{E}[Y_{am=0}|a]\\ &= \sum\nolimits_w \{\mathbb{E}[Y_{am=1}|a, w] - \mathbb{E}[Y_{am=0}|a, w]\}P(w|a)\\ &= \sum\nolimits_w \{\mathbb{E}[Y_{am=1}|a, M = 1, w] - \mathbb{E}[Y_{am=0}|a, M = 0, w]\}P(w|a)\\ &= \sum\nolimits_w \{\mathbb{E}[Y|a, M = 1, w] - \mathbb{E}[Y|a, M = 0, w]\}P(w|a)\\ &= \sum\nolimits_w \{f(a, w) + \theta - f(a, w)\}P(w|a)\\ &= \theta\end{aligned}$$

where the first equality holds by randomization of A, the second and the fifth by consistency, and the fourth by assumption (A2.2). From this it follows that $\theta = \gamma$.

Under assumptions (A2.1)–(A2.4), we have in the primary study that

$$\begin{aligned}
\mathbb{E}[Y_{1M_1} - Y_{1M_0}] &= \sum\nolimits_{m,w} \mathbb{E}[Y|A=1,m,w]\{P(m|A=1,w) \\
&\quad - P(m|A=0,w)\}P(w) \\
&= \sum\nolimits_{m,w} (f(a,w) + \gamma\, m)\{P(m|A=1,w) - P(m|A=0,w)\}P(w) \\
&= \sum\nolimits_{w} \gamma\, \{\mathbb{E}(M|A=1,w) - \mathbb{E}(M|A=0,w)\}P(w) \\
&= \sum\nolimits_{w} \gamma\, \{\mathbb{E}(M|A=1,w)P(w|A=1) \\
&\quad - \mathbb{E}(M|A=0,w)P(w|A=0)\} \\
&= \gamma\, \{\mathbb{E}(M|A=1) - \mathbb{E}(M|A=0)\}
\end{aligned}$$

where the first equality follows by the mediation formula and the fourth because A is randomized. We thus also have

$$\begin{aligned}
\mathbb{E}[Y_{1M_0} - Y_{0M_0}] &= \mathbb{E}[Y_1 - Y_0] - \mathbb{E}[Y_{1M_1} - Y_{1M_0}] \\
&= \{\mathbb{E}(Y|A=1) - \mathbb{E}(Y|A=0)\} \\
&\quad - \gamma\, \{\mathbb{E}(M|A=1) - \mathbb{E}(M|A=0)\}. \quad \blacksquare
\end{aligned}$$

A.4. MEDIATION ANALYSIS WITH SURVIVAL DATA

A.4.1. Definitions for Mediation in a Survival Context

We will let A denote an exposure of interest, T a time-to-event outcome, M a mediator, and C a set of covariates. We will let T_a denote the counterfactual event time if A had been set to a; likewise we let T_{am} denote the counterfactual event time if A had been set to a and M had been set to m. We let M_a be the counterfactual value of the mediator if A had been set to a. With these definitions we can also consider nested counterfactual event times. For example, $T_{aM_{a^*}}$ is an individual's event time if the exposure had been set to a and the mediator had been set to the level it would have been had exposure been a^*. We assume composition, that is, $T_a = T_{aM_a}$. For an arbitrary time-to-event variable V we will let $S_V(t)$ denote the survival function at time t, that is, $S_V(t) = P(V > t)$; the survival function conditional on covariates $C = c$ can likewise be defined as $S_V(t|c) = P(V > t|c)$. We will use $\lambda_V(t)$ and $\lambda_V(t|c)$ for the hazard or conditional hazard at time t, that is, the instantaneous rate of the event conditional on $V \geq t$.

For survival data within the context of mediation analysis there are multiple ways or scales by which we might decompose a total effect comparing exposure levels a and a^* into direct and indirect effects. For example, if we were to consider the survival functions, we could decompose a comparison of the survival functions $S_{T_a}(t)$

and $S_{T_{a^*}}(t)$ as follows:

$$S_{T_a}(t) - S_{T_{a^*}}(t) = \left[S_{T_{aM_a}}(t) - S_{T_{aM_{a^*}}}(t)\right] + \left[S_{T_{aM_{a^*}}}(t) - S_{T_{a^*M_{a^*}}}(t)\right]$$

where the first expression in brackets is the natural indirect effect on the survival function scale and the second is the natural direct effect on the survival function scale. We could alternatively but similarly decompose the overall difference in hazards as the sum of natural indirect and direct effects on the hazard scale:

$$\lambda_{T_a}(t) - \lambda_{T_{a^*}}(t) = \left[\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)\right] + \left[\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)\right]$$

We could, however, also consider other effect decompositions. We could, for example, consider a decomposition in terms of mean survival times:

$$\mathbb{E}(T_a) - \mathbb{E}(T_{a^*}) = \left[\mathbb{E}(T_{aM_a}) - \mathbb{E}(T_{aM_{a^*}})\right] + \left[\mathbb{E}(T_{aM_{a^*}}) - \mathbb{E}(T_{a^*M_{a^*}})\right]$$

Or if we let let Q_a and Q_{am} denote the median counterfactual survival time if A had been set to a or if A had been set to a and M had been set to m, respectively, then we have the decomposition

$$Q_a - Q_{a^*} = [Q_{aM_a} - Q_{aM_{a^*}}] + [Q_{aM_{a^*}} - Q_{a^*M_{a^*}}]$$

One could also consider using the difference in log-survival function, or log-hazards, or log-expected survival times, and so on. For example, with log-hazard one has the decomposition

$$\begin{aligned}\log\{\lambda_{T_a}(t)\} - \log\{\lambda_{T_{a^*}}(t)\} = &\left[\log\{\lambda_{T_{aM_a}}(t)\} - \log\{\lambda_{T_{aM_{a^*}}}(t)\}\right] \\ &+ \left[\log\{\lambda_{T_{aM_{a^*}}}(t)\} - \log\{\lambda_{T_{a^*M_{a^*}}}(t)\}\right]\end{aligned}$$

which, by exponentiating, can also be written

$$\lambda_{T_a}(t)/\lambda_{T_{a^*}}(t) = \left[\lambda_{T_{aM_a}}(t)/\lambda_{T_{aM_{a^*}}}(t)\right] \times \left[\lambda_{T_{aM_{a^*}}}(t)/\lambda_{T_{a^*M_{a^*}}}(t)\right]$$

so that the hazard ratio is the product of the natural indirect and direct effect hazard ratios. All of the above measures could also be considered conditional on strata of covariates $C = c$. With each of these potential decompositions on the difference scale, one could calculate a "proportion mediated" by taking a ratio of the natural indirect effect to the sum of the natural direct and indirect effects (i.e., the total effect). These measures of the proportion mediated may vary across scales. Also, depending on the specific survival model, the natural direct and indirect effects may be analytically tractable on certain scales but not on others.

Assumptions for the identification of the above effects are analogous to (A2.1)–(A2.4) above with Y_a and Y_{am} replaced by T_a and T_{am}: (A4.1) $T_{am} \perp\!\!\!\perp A|C$, (A4.2) $T_{am} \perp\!\!\!\perp M|\{A, C\}$, (A4.3) $M_a \perp\!\!\!\perp A|C$, (A4.4) $T_{am} \perp\!\!\!\perp M_{a^*}|C$; that is, conditional on C, no unmeasured exposure–outcome, mediator–outcome, and exposure–mediator confounding and no effect of the exposure that itself confounds the mediator–outcome relationship.

A.4.2. Mediation with Accelerated Failure Time Models

Proposition 4.1 (VanderWeele, 2011b):
Suppose assumptions (A4.1)–(A4.4) hold and that the mediator M is continuous and follows a linear regression model

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

with M conditionally normally distributed given A, C with conditional variance σ^2. Suppose that T follows an accelerated failure time model:

$$\log(T) = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4' C + \nu\varepsilon$$

then natural direct and indirect effects on the log expected survival time ratio scale are given by

$$\log\{\mathbb{E}(T_{aM_a}|c)\} - \log\{\mathbb{E}(T_{aM_{a^*}}|c)\} = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$
$$\log\{\mathbb{E}(T_{aM_{a^*}}|c)\} - \log\{\mathbb{E}(T_{a^*M_{a^*}}|c)\} = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})$$

with standard errors as given in Proposition 2.4.

Proof:
We have that

$$\begin{aligned}
\mathbb{E}(T_{aM_{a^*}}|c) &= \int \mathbb{E}[T_{am}|c, M_{a^*} = m]\, dP_{M_{a^*}}(m|c) \\
&= \int \mathbb{E}[T_{am}|c] dP_{M_{a^*}}(m|c) \\
&= \int \mathbb{E}[T|a,m,c] dP_M(m|a^*,c) \\
&= \int \mathbb{E}[e^{\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c+\nu\varepsilon}]\, dP_M(m|a^*,c) \\
&= e^{\theta_0+\theta_1 a+\theta_4' c}\mathbb{E}[e^{\nu\varepsilon}]\mathbb{E}[e^{\theta_2 M+\theta_3 aM}] \\
&= e^{\theta_0+\theta_1 a+\theta_4' c}\mathbb{E}[e^{\nu\varepsilon}] e^{(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2}
\end{aligned}$$

where the first equality follows by the law of iterated expectations, the second by assumption (A4.4), the third by assumptions (A4.1)–(A4.3), the fourth by the accelerated failure time model, and the final one by the fact that M is normally distributed and has constant variance σ^2. Thus,

$$\log\{\mathbb{E}(T_{aM_{a^*}}|c)\} = \log(\mathbb{E}[e^{\nu\varepsilon}]) + \theta_0 + \theta_1 a + \theta_4' c + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2$$

and so

$$\log\{\mathbb{E}(T_{aM_a}|c)\} - \log\{\mathbb{E}(T_{aM_{a^*}}|c)\} = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$
$$\log\{\mathbb{E}(T_{aM_{a^*}}|c)\} - \log\{\mathbb{E}(T_{a^*M_{a^*}}|c)\} = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\} \times (a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})$$

These are the same algebraic formulas as in Proposition 2.4 and so the standard errors in that proposition apply here also. ■

We now show that for the accelerated failure time model with a continuous mediator the product and difference methods will coincide.

Proposition 4.2 (VanderWeele, 2011b):
Suppose that the model for the mediator is

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

and the outcome follows an accelerated failure time model:

$$\log(T) = \theta_0 + \theta_1 A + \theta_2 M + \theta_4' C + \nu\varepsilon$$

Suppose also a model is fit for the outcome with just the exposure, not the mediator:

$$\log(T) = \phi_0 + \phi_1 A + \phi_4' C + \varkappa\varepsilon$$

The difference method uses $\phi_1 - \theta_1$ as a measure of the indirect effect; the product method uses $\beta_1\theta_2$. If all of the models are correctly specified, then product and difference methods coincide, that is, $\phi_1 - \theta_1 = \beta_1\theta_2$.

Proof:
By the model for the outcome without the mediator we have

$$\mathbb{E}[\log(T)|a,c] = \phi_0 + \phi_1 a + \phi_4' c + \varkappa\mathbb{E}[\varepsilon]$$

and we also have

$$\begin{aligned}\mathbb{E}[\log(T)|a,c] &= \mathbb{E}[\mathbb{E}[\log(T)|a,M,c]] \\ &= \theta_0 + \theta_1 a + \theta_2\mathbb{E}[M|a,c] + \theta_4' c + \nu\mathbb{E}[\varepsilon] \\ &= \theta_0 + \theta_1 a + \theta_2\{\beta_0 + \beta_1 a + \beta_2' c\} + \theta_4' c + \nu\mathbb{E}[\varepsilon] \\ &= \{\theta_0 + \theta_2\beta_0\} + \{\theta_1 + \theta_2\beta_1\}a + \{\theta_4' + \theta\beta_2'\}c + \nu\mathbb{E}[\varepsilon]\end{aligned}$$

Because this holds for all a, we must have $\phi_1 = \{\theta_1 + \theta_2\beta_1\}$ and thus $\phi_1 - \theta_1 = \theta_2\beta_1$. ■

Proposition 4.3:
Suppose assumptions (A4.1)–(A4.4) hold and that the mediator M is binary and follows a logistic regression model

$$\text{logit}\{P(M = 1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

and suppose that T follows an accelerated failure time model

$$\log(T) = \theta_0 + \theta_1 A + \theta_2 M + \theta_3 AM + \theta_4' C + \nu\varepsilon$$

then natural direct and indirect effects on the mean survival time ratio scale are given by

$$\frac{\mathbb{E}(T_{aM_a}|c)}{\mathbb{E}(T_{aM_{a^*}}|c)} = \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2' c)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$
$$\frac{\mathbb{E}(T_{aM_{a^*}}|c)}{\mathbb{E}(T_{a^*M_{a^*}}|c)} = \frac{\exp(\theta_1 a)\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}{\exp(\theta_1 a^*)\{1+\exp(\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

with standard errors as given in Proposition 2.6.

Proof:
We have that

$$\begin{aligned}
\mathbb{E}(T_{aM_{a^*}}|c) &= \sum\nolimits_m \mathbb{E}[T_{am}|c, M_{a^*}=m]P(M_{a^*}=m|c) \\
&= \sum\nolimits_m \mathbb{E}[T_{am}|c]P(M_{a^*}=m|c) \\
&= \sum\nolimits_m \mathbb{E}[T|a,m,c]P(m|a^*,c) \\
&= \sum\nolimits_m \mathbb{E}[e^{\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c+\nu\varepsilon}]P(m|a^*,c) \\
&= e^{\theta_0+\theta_1 a+\theta_4' c}\mathbb{E}[e^{\nu\varepsilon}]\left(\frac{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)}{1+exp(\beta_0+\beta_1 a^*+\beta_2' C)}\right)
\end{aligned}$$

and thus

$$\frac{\mathbb{E}(T_{aM_a}|c)}{\mathbb{E}(T_{aM_{a^*}}|c)} = \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2' c)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$
$$\frac{\mathbb{E}(T_{aM_{a^*}}|c)}{\mathbb{E}(T_{a^*M_{a^*}}|c)} = \frac{\exp(\theta_1 a)\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}{\exp(\theta_1 a^*)\{1+\exp(\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

These are the same algebraic formulas as in Proposition 2.6 and so the standard errors in that proposition apply here also. ■

A.4.3. Mediation with Proportional Hazards Models

Proposition 4.4 (VanderWeele, 2011b):
Suppose assumptions (A4.1)–(A4.4) hold and that the mediator M is continuous and follows a linear regression model

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

with M conditionally normally distributed given A, C with conditional variance σ^2. Suppose that T follows a proportional hazards model

$$\lambda_T(t|a,m,c) = \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}$$

and the outcome is relatively rare, then natural direct and indirect effects on the log hazards ratio scale are given by

$$\begin{aligned}\log\{\lambda_{T_{aM_a}}(t|c)\} - \log\{\lambda_{T_{aM_{a^*}}}(t|c)\} &\approx (\theta_2\beta_1 + \theta_3\beta_1 a)(a-a^*)\\ \log\{\lambda_{T_{aM_{a^*}}}(t|c)\} - \log\{\lambda_{T_{a^*M_{a^*}}}(t|c)\} &\approx \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}\\ &\quad \times (a-a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\end{aligned}$$

with standard errors as given in Proposition 2.4.

Proof:
Under the proportional hazard model,

$$\lambda_{T_{aM_{a^*}}}(t|c) = \frac{f_{T_{aM_{a^*}}}(t|c)}{S_{T_{aM_{a^*}}}(t|c)}$$

where $f_{T_{aM_{a^*}}}(t|c)$ and $S_{T_{aM_{a^*}}}(t|c)$ denote the conditional density and survival functions respectively for $T_{aM_{a^*}}$. We have that

$$\begin{aligned}f_{T_{aM_{a^*}}}(t|c) &= \int f_{T_{am}}(t|c, M_{a^*} = m) dP_{M_{a^*}}(m|c)\\ &= \int f_{T_{am}}(t|c) dP_{M_{a^*}}(m|c) \quad \text{by assumption (A4.4)}\\ &= \int f_T(t|a,m,c) dP_M(m|a^*,c) \quad \text{by assumptions (A4.1)–(A4.3)}\\ &= \int \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}\\ &\quad \times \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}\right) dP_M(m|a^*,c)\end{aligned}$$

where $\Lambda_T(t|0,0,0) = \int_0^t \lambda_T(t|0,0,0)\, dt$. Likewise,

$$S_{T_{aM_{a^*}}}(t|c) = \int \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}\right) dP_M(m|a^*,c)$$

Thus,

$$\lambda_{T_{aM_{a^*}}}(t|c) = \lambda_T(t|0,0,0)\exp(\theta_1 a + \theta_4' c) r(t; a, a^*, c)$$

where

$$r(t; a, a^*, c) = \frac{\int e^{(\theta_2+\theta_3 a)m}\exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}\right) dP_M(m|a^*,c)}{\int \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}\right) dP_M(m|a^*,c)}$$

Since M is normally distributed, we have that

$$
\begin{aligned}
&r(t;a,a^*,c)\\
&\quad = e^{(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2}\\
&\qquad \times \frac{\int \exp(-\Lambda_T(t|0,0,0)e^{(\theta_2+\theta_3 a)^2+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}) \exp\left(-\frac{(m-(\beta_0+\beta_1 a^*+\beta_2' c))^2}{2}\right) dm}{\int \exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}) \exp\left(-\frac{(m-(\beta_0+\beta_1 a^*+\beta_2' c))^2}{2}\right) dm}
\end{aligned}
$$

which can be approximated by $e^{(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2}$ if $\Lambda_T(t|0,0,0)$ is small (i.e., if the outcome is relatively rare). Thus

$$\lambda_{T_{aM_{a^*}}}(t|c) \approx \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_4' c}e^{(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2}$$

and

$$
\begin{aligned}
\log\{\lambda_{T_{aM_{a^*}}}(t|c)\} &= \log(\lambda_T(t|0,0,0)) + \theta_1 a + \theta_4' c + (\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)\\
&\quad + \frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2
\end{aligned}
$$

From this it follows that

$$
\begin{aligned}
\log\{\lambda_{T_{aM_a}}(t|c)\} - \log\{\lambda_{T_{aM_{a^*}}}(t|c)\} &= (\theta_2\beta_1+\theta_3\beta_1 a)(a-a^*)\\
\log\{\lambda_{T_{aM_{a^*}}}(t|c)\} - \log\{\lambda_{T_{a^*M_{a^*}}}(t|c)\} &= \{\theta_1+\theta_3(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2\sigma^2)\}\\
&\quad \times(a-a^*) + 0.5\theta_3^2\sigma^2(a^2-a^{*2})
\end{aligned}
$$

These are the same algebraic formulaes as in Proposition 2.4 and so the standard errors in that proposition apply here also. ■

We now show that for the proportional hazards model with a continuous mediator and a rare outcome the product and difference methods will coincide approximately.

Proposition 4.5 (VanderWeele, 2011b):
Suppose that the model for the mediator is

$$\mathbb{E}[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

and the outcome follows a proportional hazards model

$$\lambda_T(t|a,m,c) = \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}$$

Suppose also that a proportional hazards model is also fit without the mediator:

$$\lambda_T(t|a,c) = \lambda_T(t|0,0)e^{\phi_1 a+\phi_4' c}$$

The difference method uses $\phi_1-\theta_1$ as a measure of the indirect effect; the product method uses $\beta_1\theta_2$. If all of the models are correctly specified and the outcome is rare, these two are approximately equal, that is, $\phi_1-\theta_1 \approx \beta_1\theta_2$.

Proof:
By the proportional hazards model without the mediator:

$$\lambda_T(t|a,c) = \lambda_T(t|0,0)e^{\phi_1 a + \phi_4' c}$$

and by the model with the mediator:

$$\begin{aligned}
\lambda_T(t|a,c) &= \frac{f_T(t|a,c)}{S_T(t|a,c)} \\
&= \frac{\int f_T(t|a,m,c)dP_M(m|a,c)}{\int S_T(t|a,m,c)\,dP_M(m|a,c)} \\
&= \frac{\int \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}\exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}\right)dP_M(m|a,c)}{\int \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}\right)dP_M(m|a,c)} \\
&= \lambda_T(t|0,0,0)\exp\left(\theta_1 a+\theta_4' c\right)r(t;a,c)
\end{aligned}$$

where

$$r(t;a,c) = \frac{\int e^{\theta_2 m}\exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}\right)dP_M(m|a,c)}{\int \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}\right)dP_M(m|a,c)}$$

As in the proof of Proposition 4.3, since M is normally distributed, we have that

$$\begin{aligned}
r(t;a,c) = {} & e^{\theta_2(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}\theta_2^2\sigma^2} \\
& \times \frac{\int \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_2^2+\theta_1 a+\theta_2 m+\theta_4' c}\right)\exp\left(-\frac{(m-(\beta_0+\beta_1 a+\beta_2' c))^2}{2}\right)dm}{\int \exp\left(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_4' c}\right)\exp\left(-\frac{(m-(\beta_0+\beta_1 a+\beta_2' c))^2}{2}\right)dm}
\end{aligned}$$

which can be approximated by $e^{\theta_2(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}\theta_2^2\sigma^2}$ if $\Lambda_T(t|0,0,0)$ is small (i.e., if the outcome is relatively rare). Thus

$$\lambda_T(t|a,c) \approx \{e^{\theta_2\beta_0+\frac{1}{2}\theta_2^2\sigma^2}\lambda_T(t|0,0,0)\}e^{(\theta_1+\theta_2\beta_1)a+(\theta_2\beta_2+\theta_4)'c}$$

Because this holds for all a, we must have $\phi_1 \approx \{\theta_1 + \theta_2\beta_1\}$ and thus $\phi_1 - \theta_1 \approx \theta_2\beta_1$. ■

Proposition 4.6:
Suppose assumptions (A4.1)–(A4.4) hold and that the mediator M is binary and follows a logistic regression model

$$\text{logit}\{P(M=1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

and suppose that T follows a proportional hazards model

$$\lambda_T(t|a,m,c) = \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}$$

and the outcome is relatively rare then natural direct and indirect effects on the hazards ratio scale are given by

$$\frac{\lambda_{T_{aM_a}}(t|c)}{\lambda_{T_{aM_{a^*}}}(t|c)} = \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2' c)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

$$\frac{\lambda_{T_{aM_{a^*}}}(t|c)}{\lambda_{T_{a^*M_{a^*}}}(t|c)} = \frac{\exp(\theta_1 a)\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}{\exp(\theta_1 a^*)\{1+\exp(\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

with standard errors as given in Proposition 2.6.

Proof:
Under the proportional hazard model,

$$\lambda_{T_{aM_{a^*}}}(t|c) = \frac{f_{T_{aM_{a^*}}}(t|c)}{S_{T_{aM_{a^*}}}(t|c)}$$

where $f_{T_{aM_{a^*}}}(t|c)$ and $S_{T_{aM_{a^*}}}(t|c)$ denote the conditional density and survival functions, respectively, for $T_{aM_{a^*}}$. We have that

$$\begin{aligned}
f_{T_{aM_{a^*}}}(t|c) &= \sum\nolimits_m f_{T_{am}}(t|c, M_{a^*}=m)P(M_{a^*}=m|c) \\
&= \sum\nolimits_m f_{T_{am}}(t|c)P(M_{a^*}=m|c) \text{ by assumption (A4.4)} \\
&= \sum\nolimits_m f_T(t|a,m,c)P(m|a^*,c) \text{ by assumptions (A4.1)–(A4.3)} \\
&= \sum\nolimits_m \lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c} \\
&\quad \exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c})P(m|a^*,c)
\end{aligned}$$

where $\Lambda_T(t|0,0,0) = \int_0^t \lambda_T(t|0,0,0)dt$. Likewise,

$$S_{T_{aM_{a^*}}}(t|c) = \sum\nolimits_m \exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c})P(m|a^*,c)$$

Thus,

$$\lambda_{T_{aM_{a^*}}}(t|c) = \lambda_T(t|0,0,0)\exp(\theta_1 a+\theta_4' c)r(t;a,a^*,c)$$

where

$$r(t;a,a^*,c) = \frac{\sum_m e^{(\theta_2+\theta_3 a)m}\exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c})P(m|a^*,c)}{\sum_m \exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c})P(m|a^*,c)}$$

If the outcome is rare so that $\Lambda_T(t|0,0,0)$ is close to zero, then $\exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c}) \approx 1$ and we obtain

$$\begin{aligned}
r(t;a,a^*,c) &= \frac{\sum_m e^{(\theta_2+\theta_3 a)m}\exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c})P(m|a^*,c)}{\sum_m \exp(-\Lambda_T(t|0,0,0)e^{\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c})P(m|a^*,c)} \\
&= \frac{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)}{1+\exp(\beta_0+\beta_1 a^*+\beta_2' C)}
\end{aligned}$$

and thus

$$\lambda_{T_{aM_{a^*}}}(t|c) \approx \lambda_T(t|0,0,0)\exp(\theta_1 a + \theta_4' c)\frac{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)}{1+\exp(\beta_0+\beta_1 a^*+\beta_2' C)}$$

From this it follows that

$$\frac{\lambda_{T_{aM_a}}(t|c)}{\lambda_{T_{aM_{a^*}}}(t|c)} = \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a+\beta_2' c)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

$$\frac{\lambda_{T_{aM_{a^*}}}(t|c)}{\lambda_{T_{a^*M_{a^*}}}(t|c)} = \frac{\exp(\theta_1 a)\{1+\exp(\theta_2+\theta_3 a+\beta_0+\beta_1 a^*+\beta_2' c)\}}{\exp(\theta_1 a^*)\{1+\exp(\theta_2+\theta_3 a^*+\beta_0+\beta_1 a^*+\beta_2' c)\}}$$

These are the same algebraic formulaes as in Proposition 2.6 and so the standard errors in that proposition apply here also. ■

A.4.4. Mediation with Additive Hazards Models

Lange and Hansen (2011) derive an analytic formula using natural direct and indirect effects on the hazard difference scale for the additive hazards model. Their approach also allows for multiple event types. Consider then the more general setting of K distinct event types. Define $\epsilon_t \in \{1,\ldots,K\}$ as the event type. Let T_{am} and ϵ_{am} denote the counterfactual event time and event type if A were set to a and M to m. Let $\lambda^k_{T_{aM_{a^*}}}(t)$ denote the counterfactual rate for event of type k if A were set to a and M to M_{a^*}. Assume that the rate for the event type k is follows an Aalen additive hazard model; that is, assume that the rate satisfies

$$\lim_{dt\to 0} P(T\in[t,t+dt],\epsilon=k \mid T\geq t, A=a, C=c, M=m)/dt$$
$$= \lambda_0^k(t)+\lambda_1^k(t)a+\lambda_2^k(t)m+\lambda_4^k(t)'c$$

where $\lambda_j^k(t)$ are potentially time dependent coefficient functions and A and C can take vector values. Note that by a simple conditioning argument the "any-event" rate satisfies

$$\lim_{dt\to 0} P(T\in[t,t+dt] \mid T\geq t, A=a, C=c, M=m)/dt$$
$$= g_0(t)+g_1(t)a+g_2(t)m+g_4(t)'c$$

for functions $g_j(t)=\sum_{k=1}^K \lambda_j(t)$

The following proposition holds under the analogues of no-unmeasured confounding-assumptions (A4.1)–(A4.4) for multiple event times.

Proposition 4.7 (Lange and Hansen, 2011):
Suppose $(T_{am},\epsilon_{am}) \perp\!\!\!\perp A|C$, $(T_{am},\epsilon_{am}) \perp\!\!\!\perp M|A,C$, $M_a \perp\!\!\!\perp A|C$, and $(T_{am},\epsilon_{am}) \perp\!\!\!\perp M_{a^*}|C$ and that the model for the mediator is

$$\mathbb{E}[M|a,c] = \beta_0+\beta_1 a+\beta_2' c$$

and the outcome follows an additive hazards model of the form above, then natural direct and indirect effects for an event of type k on the hazards difference scale are given by

$$\lambda^k{}_{T_{aM_a}}(t) - \lambda^k_{T_{aM_{a^*}}}(t) = \lambda^k_2(t)\beta_1(a-a^*)$$
$$\lambda^k_{T_{aM_{a^*}}}(t) - \lambda^k_{T_{a^*M_{a^*}}}(t) = \lambda^k_1(t)(a-a^*)$$

Proof:
Initially rewrite the probability of an event of type k within the time interval $]t, t+dt]$ for the counterfactual variable $(T_{aM_{a^*}}, \epsilon_{aM_{a^*}})$ conditional on being at risk at time t as

$$\begin{aligned}
&P(T_{aM_{a^*}} \in dt, \epsilon_{aM_{a^*}} = k \mid T^{a,M_{a^*}} \geq t) \\
&\quad = \sum\nolimits_c P(T_{aM_{a^*}} \in dt, \epsilon_{aM_{a^*}} = k \mid C = c, T_{aM_{a^*}} \geq t) P(C = c \mid T_{aM_{a^*}} \geq t)
\end{aligned}$$

The first of the probabilities in the summation can be rewritten as

$$\begin{aligned}
&P(T_{aM_{a^*}} \in dt, \epsilon_{aM_{a^*}} = k \mid C = c, T_{aM_{a^*}} \geq t) \\
&\quad = \int_{m\in\mathbb{R}} P(T_{aM_{a^*}} \in dt, \epsilon_{aM_{a^*}} = k \mid M_{a^*} = m, C = c, T_{aM_{a^*}} \geq t) \\
&\qquad \times P(M_{a^*} \in dm \mid C = c, T_{aM_{a^*}} \geq t) \\
&\quad \overset{(a)}{=} \int_{m\in\mathbb{R}} P(T_{am} \in dt, \epsilon_{am} = k \mid C = c, T_{am} \geq t) \\
&\qquad \times P(M_{a^*} \in dm \mid C = c, T_{aM_{a^*}} \geq t) \\
&\quad \overset{(b)}{=} \int_{m\in\mathbb{R}} P(T_{am} \in dt, \epsilon_{am} = k \mid C = c, A = a, T_{am} \geq t) \\
&\qquad \times \frac{P(M_{a^*} \in dm, T_{aM_{a^*}} \geq t \mid C = c)}{P(T_{aM_{a^*}} \geq t \mid C = c)} \\
&\quad = \int_{m\in\mathbb{R}} P(T_{am} \in dt, \epsilon_{am} = k \mid C = c, A = a, T_{am} \geq t) \\
&\qquad \times \frac{P(T_{aM_{a^*}} \geq t \mid M_{a^*} \in dm, C = c) P(M_{a^*} \in dm \mid C = c)}{P(T_{aM_{a^*}} \geq t \mid C = c)} \\
&\quad \overset{(c)}{=} \int_{m\in\mathbb{R}} P(T_{am} \in dt, \epsilon_{am} = k \mid C = c, M = m, A = a, T_{am} \geq t) \\
&\qquad \times \frac{P(T_{am} \geq t \mid C = c) P(M_{a^*} \in dm \mid C = c)}{P(T_{aM_{a^*}} \geq t \mid C = c)}, \\
&\quad \overset{(d)}{=} \int_{m\in\mathbb{R}} P(T \in dt, \epsilon = k \mid C = c, M = m, A = a, T \geq t) \\
&\qquad \times \frac{P(T_{am} \geq t \mid C = c) P(M \in dm \mid C = c, A = a^*)}{\int_{\tilde{m}\in\mathbb{R}} P(T_{am} \geq t \mid C = c) P(M \in d\tilde{m} \mid C = c, A = a^*)},
\end{aligned}$$

where equality (a) is due to assumption $(T_{am}, \epsilon_{am}) \perp\!\!\!\perp M_{a^*} C$, equality (b) is due to assumption $(T_{am}, \epsilon_{am}) \perp\!\!\!\perp A|C$, equality (c) is by assumption $(T_{am}, \epsilon_{am}) \perp\!\!\!\perp M|A, C$ and $(T_{am}, \epsilon_{am}) \perp\!\!\!\perp M_{a^*} C$, and equality (d) is due to assumption $M_a \perp\!\!\!\perp A|C$. Similar considerations combined with the expression for the survival function for the Aalen additive model (Martinussen and Scheike, 2006) give

$$\begin{aligned} P(T_{am} \geq t \mid C = c) &= P(T \geq t \mid C = c, A = a, M = m) \\ &= \exp\{-G_0(t) - G_1(t)a - G_2(t)m - G_4(t)'c\} \end{aligned}$$

where $G_j(t) = \int_0^t g_j(s)\, ds$. Hence by the additive hazard model and the bounded convergence theorem, it follows that

$$\begin{aligned} &\lim_{dt \to 0} P(T_{aM_{a^*}} \in dt, \epsilon_{aM_{a^*}} = k \mid T_{aM_{a^*}} \geq t)/dt \\ &\quad = \sum\nolimits_c p_c \{\mathbb{E}[\exp\{-G_0(t) - G_1(t)a - G_2(t)M - G_4(t)'c\} \mid A = a^*, C = c]\}^{-1} \\ &\qquad \times E\big[(\lambda_0^k(t) + \lambda_1^k(t)a + \lambda_2^k(t)M + \lambda_4^k(t)'c) \\ &\exp\{-G_0(t) - G_1(t)a - G_2(t)M - G_4(t)c\} \mid A = a^*, C = c] \\ &\quad = \lambda_0^k(t) + \lambda_1^k(t)a \\ &\qquad + \sum\nolimits_c \left\{ \lambda_4^k(t)'c p_c + \frac{\mathbb{E}[\lambda_2^k(t) M \exp\{-G_2(t)M\} \mid A = a^*, C = c]}{\mathbb{E}[\exp\{-G_2(t)M\} \mid A = a^*, C = c]} p_c \right\} \end{aligned}$$

where $p_c = P(C = c \mid T_{aM_{a^*}} \geq t)$. For a random variable $U \sim N(\mu, \omega^2)$ it follows by properties of the characteristic function of a normal random variable (Patel and Campbell, 1996) that

$$\frac{\mathbb{E}[U \exp\{aU\}]}{\mathbb{E}[\exp\{aU\}]} = \mu + a\omega^2$$

Since M is conditionally normal, this implies that we can rewrite the above expression as

$$\begin{aligned} &\lim_{dt \to 0} P(T_{aM_{a^*}} \in dt, \epsilon_{aM_{a^*}} = k \mid T_{aM_{a^*}} \geq t)/dt \\ &\quad = \lambda_0^k(t) + \lambda_1^k(t)a + \sum\nolimits_c \{\lambda_4^k(t)'c p_c + \lambda_2^k(t)(\beta_0 + \beta_1 a^* + \beta_2' c - G_2^k(t)\sigma^2) p_c\} \\ &\quad = \lambda_0^k(t) + \lambda_1^k(t)a + \lambda_2^k(t)(\beta_0 + \beta_1 a^* - G_2^k(t)\sigma^2) + \sum\nolimits_c \{\lambda_4^k(t)'c p_c + \lambda_2^k(t)\beta_2' c\} \end{aligned}$$

In summary, it has been established that the counterfactual rate for event type k can be expressed as

$$\begin{aligned} \lambda^k_{T_{aM_{a^*}}}(t) = \lambda_0^k(t) + \lambda_1^k(t)a + \lambda_2^k(t)(\beta_0 + \beta_1 a^* - G_2^k(t)\sigma^2) \\ + \sum\nolimits_c \{\lambda_4^k(t) + \lambda_2^k(t)\beta_2\} c p_c \end{aligned}$$

From this we have

$$\begin{aligned} \lambda^k{}_{T_{aM_a}}(t) - \lambda^k_{T_{aM_{a^*}}}(t) &= \lambda_2^k(t)\beta_1(a - a^*) \\ \lambda^k_{T_{aM_{a^*}}}(t) - \lambda^k_{T_{a^*M_{a^*}}}(t) &= \lambda_1^k(t)(a - a^*) \quad \blacksquare \end{aligned}$$

From Proposition 4.5 it also immediately follows that, under the same assumptions, natural direct and indirect effects on the cumulative hazards scale are given by $(a-a^*)\int_0^t \lambda_1^k(s)ds$ and $\beta_1(a-a^*)\int_0^t \lambda_2^k(s)ds$, respectively. If the data support that the effects in the Aalen model do not depend on time (an assumption that can be tested as discussed in Lange et al. (2012)), the time-dependent hazards in the expressions above can be replaced by simple time-invariant parameters and the results reduce to the expressions provided in the main text.

For standard errors for these estimators, let θ_1 denote the collection of parameters from the ordinary regression of the mediator on the exposure and baseline covariates and let $\theta_2(t)$ denote the collection of parameter functions of the Aalen model. In order to discuss estimation uncertainty, define the cumulative coefficient functions as $\Theta_2(t) = \int_0^t \theta_2(s)\,ds$. Under mild regularity conditions (see, e.g., Condition 5.1 of Martinussen and Scheike (2006)), it holds that $\hat{\theta}_1$ is asymptotically normally distributed and for any t it holds that $\hat{\Theta}_2(t)$ is also asymptotically normally distributed. In addition, the two vectors of estimates are uncorrelated. The covariance matrices of the two vectors of estimates are available as output from standard statistical software. Hence the cumulative direct effect at time t is asymptotically normal, while the cumulative indirect effect at time t is asymptotically distributed as the product of two uncorrelated normal random variables. Confidence bands and tests involving a cumulative indirect effect can be computed either by using the delta method or by simulating a large number of realizations of the two uncorrelated random variables (cf. Lange and Hansen, 2011).

A.4.5. A Weighting Approach to Mediation with Survival Data

A weighting approach to estimate natural direct and indirect effects can also be employed. This can be done for the proportional hazards model with a common outcome and can also be applied to continuous or binary outcomes as well. Here we give the derivation in Lange et al. (2012; cf. Hong, 2010) that applies for a binary or continuous outcome and provides the justification for the weighting approach.

Proposition 4.8 (Hong, 2010; Lange et al., 2012):
Under no-unmeasured-confounding assumptions (A2.1)–(A2.4) we have $\mathbb{E}[\,Y_{aM_{a^*}}\,] = E\left[YI(A=a)W\right]$ where

$$W = \frac{1}{P[A=a|C=c]}\,\frac{P[M=m|A=a^*,C=c]}{P[M=m|A=a,C=c]}$$

Proof:
Under assumptions (A2.1)–(A2.4) it was shown in Section A.2.1 that $\mathbb{E}[Y_{aM_{a^*}}] = \sum_m \sum_c \mathbb{E}[Y|M=m,A=a,C=c]\,P[M=m|A=a^*,C=c]\,P(C=c)$. We further have that this is equal to

$$= \sum\nolimits_{y,m,c} yP[Y=y|M=m,A=a,C=c]P[M=m|A=a^*,C=c]P(C=c)$$

$$= \sum_{y,m,a,c} yI(A=a)P[Y=y|M=m,A=a,C=c]P[M=m|A=a,C=c]$$
$$\times P[A=a|C=c]P(C=c)\frac{1}{P[A=a|C=c]}\frac{P[M=m|A=a^*,C=c]}{P[M=m|A=a,C=c]}$$
$$= \sum_{y,m,a,c} yI(A=a)WP[Y=y,M=m,A=a,C=c]$$
$$= E\left[YI(A=a)W\right]. \quad \blacksquare$$

Lange et al. (2012) consider fitting what they call natural effects marginal structural models of the form $\mathbb{E}(Y_{a,M_{a^*}}) = \kappa_0 + \kappa_1 a + \kappa_2 a^* + \kappa_3 a a^*$. The weighting procedure for generalized linear natural effect marginal structural models with canonical link amounts to solving an estimating equation of the form

$$\sum_{i=1}^{n}\sum_{a^*=0}^{1} d(A_i,a^*)(Y_i - \kappa_0 - \kappa_1 A_i - \kappa_2 a^* - \kappa_3 A_i a^*)\frac{1}{P[A_i|C_i]}\frac{P[M_i|A_i=a^*,C_i]}{P[M_i|A_i,C_i]}$$

with $d(A,a^*) = (1\ A\ a^*\ Aa^*)$. By the derivation in Proposition 4.6, this estimating equation can be shown to have mean zero under the natural effects marginal structural model for each choice of four-dimensional function $d(A,a^*)$, upon rewriting the above equation as

$$\sum_{i=1}^{n}\sum_{a=0}^{1}\sum_{a^*=0}^{1} d(a,a^*)I(A_i=a)[Y_i - \kappa_0 - \kappa_1 a - \kappa_2 a^* - \kappa_3 a a^*)$$
$$\times\frac{1}{P[A_i=a|C_i]}\frac{P[M_i|A_i=a^*,C_i]}{P[M_i|A_i=a,C_i]}$$

The fact that these estimating equations have mean zero at the true natural effect marginal structural model is key to the fact that the proposed estimators are asymptotically unbiased. The stabilized weights correspond with the choice $d(A,a^*) = P(A)(1\ A\ a^*\ Aa^*)$. A proof for other natural effects marginal structural models works along similar lines. Standard errors can be obtained by bootstrapping.

A.4.6. Sensitivity Analysis with Survival Data

We will first begin with sensitivity analysis for total effects on a hazard difference and hazard ratio scale. We will then discuss sensitivity analysis for direct and indirect effects on hazard difference and hazard ratio scales. Suppose we wish to compare the effects on the hazard difference scale of exposure levels $A = a$ and $A = a^*$ and that we have measured covariates C but that there is some unmeasured covariate U. As above, let T_a denote the counterfactual time-to-event outcome if exposure A had been set to a. The true hazard difference comparing exposure levels $A = a$ and $A = a^*$ is denoted by $\lambda_{T_a}(t|c) - \lambda_{T_{a^*}}(t|c)$. Suppose that the effect of exposure A on the time-to-event outcome is unconfounded conditional on (C,U) but not on C alone, that is, we assume $T_a \perp\!\!\!\perp A|(C,U)$ but that it is not the case that $T_a \perp\!\!\!\perp A|C$. On the hazard difference scale we define the bias factor as

$B_{hd}(c) = \lambda_T(t|a,c) - \lambda_T(t|a^*,c) - \{\lambda_{T_a}(t|c) - \lambda_{T_{a^*}}(t|c)\}$ that is, the difference between (i) the hazard difference comparing $A = a$ and $A = a^*$ conditional on covariates $C = c$ and (ii) the actual causal hazard difference, $\{\lambda_{T_a}(t|c) - \lambda_{T_{a^*}}(t|c)$, which we would have obtained had we been able to adjust for U as well.

Proposition 4.9 (VanderWeele, 2013c):
If $T_a \perp\!\!\!\perp A|(C,U)$ and if the outcome is rare and u' is any chosen reference value for the unmeasured confounder U, then

$$\begin{aligned} B_{hd}(c) \approx & \sum\nolimits_u \{\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')\}\{P(u|a,c) - P(u|c)\} \\ & - \sum\nolimits_u \{\lambda_T(t|a^*,c,u) - \lambda_T(t|a^*,c,u')\}\{P(u|a^*,c) \\ & - P(u|c)\}. \end{aligned}$$

Proof:
We have that

$$\lambda_{T_a}(t|c) = \frac{f_{T_a}(t|c)}{S_{T_a}(t|c)}$$

Since $T_a \perp\!\!\!\perp A|(C,U)$, we have

$$\begin{aligned} f_{T_a}(t|c) &= \sum\nolimits_u f_{T_a}(t|c,u)P(u|c) \\ &= \sum\nolimits_u f_{T_a}(t|a,c,u)P(u|c) \\ &= \sum\nolimits_u f_T(t|a,c,u)P(u|c) \\ &= \sum\nolimits_u \lambda_T(t|a,c,u)\exp\{-\Lambda_T(t|a,c,u)\}P(u|c) \end{aligned}$$

where $\Lambda_T(t|a,c,u) = \int_0^t \lambda_T(t|a,c,u)dt$. Likewise,

$$\begin{aligned} S_{T_a}(t|c) &= \sum\nolimits_u S_{T_a}(t|c,u)P(u|c) \\ &= \sum\nolimits_u S_{T_a}(t|a,c,u)P(u|c) \\ &= \sum\nolimits_u S_T(t|a,c,u)P(u|c) \\ &= \sum\nolimits_u \exp\{-\Lambda_T(t|a,c,u)\}P(u|c) \end{aligned}$$

Thus,

$$\lambda_{T_a}(t|c) = \frac{f_{T_a}(t|c)}{S_{T_a}(t|c)} = \frac{\sum_u \lambda_T(t|a,c,u)\exp\{-\Lambda_T(t|a,c,u)\}P(u|c)}{\sum_u \exp\{-\Lambda_T(t|a,c,u)\}P(u|c)}$$

If the outcome is rare so that $\Lambda_T(t|a,c,u)$ is close to zero, then $\exp\{-\Lambda_T(t|a,c,u)\}$ ≈ 1 and so we obtain:

$$\lambda_{T_a}(t|c) \approx \sum\nolimits_u \lambda_T(t|a,c,u)P(u|c)$$

Likewise with a rare outcome we have $\lambda_T(t|a,c) \approx \sum_u \lambda_T(t|a,c,u)P(u|a,c)$ and thus

$$\begin{aligned}
B_{hd}(c) &= \lambda_T(t|a,c) - \lambda_T(t|a^*,c) - \{\lambda_{T_a}(t|c) - \lambda_{T_{a^*}}(t|c)\} \\
&\approx \sum\nolimits_u \lambda_T(t|a,c,u)P(u|a,c) - \sum\nolimits_u \lambda_T(t|a^*,c,u)P(u|a^*,c) \\
&\quad - \sum\nolimits_u \lambda_T(t|a,c,u)P(u|c) + \sum\nolimits_u \lambda_T(t|a^*,c,u)P(u|c) \\
&= \sum\nolimits_u \lambda_T(t|a,c,u)\{P(u|a,c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \lambda_T(t|a^*,c,u)\{P(u|a^*,c) - P(u|c)\} \\
&= \sum\nolimits_u \{\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')\}\{P(u|a,c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \{\lambda_T(t|a^*,c,u) - \lambda_T(t|a^*,c,u')\}\{P(u|a^*,c) - P(u|c)\}. \quad \blacksquare
\end{aligned}$$

To use this bias formula would require specifying the effect of U on T on the hazard difference scale for the exposed and the unexposed, that is, $\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')$ and $\lambda_T(t|a^*,c,u) - \lambda_T(t|a^*,c,u')$, along with the distribution of the unmeasured confounder U among both the exposed and the unexposed, $P(u|a,c)$ and $P(u|a^*,c)$. Once the bias factor is computed, one can then take the estimate of the hazard difference for the effect of A on T that one had obtained using the observed data, controlling only for C, $\lambda_T(t|a,c) - \lambda_T(t|a^*,c)$, and obtain a corrected hazard difference (i.e., what one would have obtained had control been made for C and U) by subtracting the bias factor $B_{hd}(c)$ from the observed estimate. In practice this would require specifying a large number of sensitivity analysis parameters. An easier-to-use approach is possible under simplifying assumptions.

Corollary (VanderWeele, 2013c)
If $T_a \perp\!\!\!\perp A|C,U$ and the outcome is rare and if U is binary and $\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')$ does not vary with a, then

$$B_{hd}(c) = \delta(c)\gamma(c)$$

where $\delta(c) = \lambda_T(t|a,c,U=1) - \lambda_T(t|a,c,U=0)$ and $P(U=1|a,c) - P(U=0|a^*,c)$.

Proof:
If $\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')$ does not vary with a, then

$$\begin{aligned}
B_{hd}(c) &= \sum\nolimits_u \{\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')\}\{P(u|a,c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \{\lambda_T(t|a^*,c,u) - \lambda_T(t|a^*,c,u')\}\{P(u|a^*,c) - P(u|c)\} \\
&= \sum\nolimits_u \{\lambda_T(t|a,c,u) - \lambda_T(t|a,c,u')\}\{P(u|a,c) - P(u|a^*,c)\}
\end{aligned}$$

If U is binary and we take $u' = 0$ as the reference value, this becomes $\{\lambda_T(t|a,c,U=1) - \lambda_T(t|a,c,U=0)\}\{P(U=1|a,c) - P(U=1|a^*,c)\}$. $\blacksquare$

These results are analogous to those obtained for binary and continuous outcomes on the additive scale in Proposition 3.1 (VanderWeele and Arah, 2011). We see here that they are also applicable to time-to-event outcomes, which are rare.

The principle behind the proof of the proposition above is that with a rare outcome, the hazard marginalizes so that for an arbitrary time to event variable V and covariates Z and W, we have that $\lambda_V(t|z) \approx \sum_w \lambda_V(t|z,w)P(w|z)$. From this it follows that arguments for probabilities of a binary outcome that depend on marginalization will hold also approximately for the hazard of a time-to-event outcome, which is relatively rare. The proof of the proposition above essentially just used the fact that $\lambda_V(t|z) \approx \sum_w \lambda_V(t|z,w)P(w|z)$ for a rare time-to-event variable V and then replicated the argument for the additive scale for binary and continuous outcomes found in Proposition 3.1 (VanderWeele and Arah, 2011). Using this same insight, we can, for rare time-to-event outcomes, adapt other arguments in Propositions 3.1–3.4 to obtain sensitivity analysis results for time-to-event outcomes on the hazard ratio scale for total effects or on the hazard difference or hazard ratio scale for direct and indirect effects.

Consider now total effects on the hazard ratio scale and suppose once again that the effect of exposure A on the time-to-event outcome is unconfounded conditional on (C,U), that is, $T_a \perp\!\!\!\perp A|(C,U)$, but not on C alone. On the hazard ratio scale define the bias factor as

$$B_{hr}(c) = \frac{\lambda_T(t|a,c)}{\lambda_T(t|a^*,c)} \Big/ \frac{\lambda_{T_a}(t|c)}{\lambda_{T_{a^*}}(t|c)}$$

The quantity $B_{hr}(c)$ is defined as the ratio of (i) the hazard ratio comparing $A = a$ and $A = a^*$ conditional on covariates $C = c$ and (ii) the actual causal hazard ratio, $\frac{\lambda_{T_a}(t|c)}{\lambda_{T_{a^*}}(t|c)}$, which we would have obtained had we been able to adjust for U as well.

Provided that the outcome is rare, then using the argument in Proposition 3.2 for the risk ratio scale for probabilities, if u' is any chosen reference value for the unmeasured confounder U, we have

$$B_{hr}(c) \approx \frac{\sum_u \frac{\lambda_T(t|a,c,u)}{\lambda_T(t|a,c,u')} P(u|a,c)}{\sum_u \frac{\lambda_T(t|a,c,u)}{\lambda_T(t|a,c,u')} P(u|c)} \Big/ \frac{\sum_u \frac{\lambda_T(t|a^*,c,u)}{\lambda_T(t|a^*,c,u')} P(u|a^*,c)}{\sum_u \frac{\lambda_T(t|a^*,c,u)}{\lambda_T(t|a^*,c,u')} P(u|c)}$$

If, in addition, U is binary and $\gamma = \frac{\lambda_T(t|a,c,U=1)}{\lambda_T(t|a,c,U=0)}$ does not vary across strata of a, then

$$B_{hr}(c) \approx \frac{1 + (\gamma - 1)P(U = 1|a,c)}{1 + (\gamma - 1)P(U = 1|a^*,c)}$$

Suppose now we are interested in direct and indirect effects with respect to some mediator M. Let T_{am} denote the counterfactual time-to-event outcome if exposure A had been set to a and the mediator M to m. Let M_a denote the counterfactual mediator value if exposure A had been set to a. Controlled direct effects on the hazard difference scale are defined by $\lambda_{T_{am}}(t|c) - \lambda_{T_{a^*m}}(t|c)$. Controlled direct effects on the hazard ratio scale are defined by $\lambda_{T_{am}}(t|c)/\lambda_{T_{a^*m}}(t|c)$. Natural indirect and direct effects on the hazard difference scale are defined respectively

by $\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)$ and $\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)$. Natural indirect and direct effects on the hazard ratio scale are defined respectively by $\lambda_{T_{aM_a}}(t)/\lambda_{T_{aM_{a^*}}}(t)$ and $\lambda_{T_{aM_{a^*}}}(t)/\lambda_{T_{a^*M_{a^*}}}(t)$.

We first consider sensitivity analysis for controlled direct effects on the hazard difference scale. Suppose that the exposure–outcome relationship is unconfounded conditional on measured covariates C but there is an unmeasured confounder U of the mediator–outcome relationship, which, if controlled for in addition to C, would suffice for mediator–outcome confounding; that is, we assume $T_{am} \perp\!\!\!\perp A|C$ and $T_{am} \perp\!\!\!\perp M|\{A, C, U\}$. Let $B_{hd}^{CDE}(m|c)$ denote the difference between (i) the estimate of the controlled direct effect conditional on C, that is, $\lambda_T(t|a, m, c) - \lambda_T(t|a^*, m, c)$, and (ii) the true controlled direct effect:

$$B_{hd}^{CDE}(m|c) = \{\lambda_T(t|a, m, c) - \lambda_T(t|a^*, m, c)\} - \{\lambda_{T_{am}}(t|c) - \lambda_{T_{a^*m}}(t|c)\}$$

Provided that the outcome is rare, then using the argument in Proposition 3.3 above from VanderWeele (2010a) for the difference scale for probabilities, if u' is any chosen reference value for the unmeasured confounder U, then

$$B_{hd}^{CDE}(m|c) = \sum\nolimits_u \{\lambda_T(t|a, m, c, u) - \lambda_T(t|a, m, c, u')\}\{P(u|a, m, c) - P(u|a, c)\}$$
$$-\sum\nolimits_u \{\lambda_T(t|a^*, m, c, u) - \lambda_T(t|a^*, m, c, u')\}\{P(u|a^*, m, c) - P(u|a^*, c)\}$$

If U is binary, with $U \perp\!\!\!\perp A|C$ and for a particular value m, $\gamma_m = \lambda_T(t|a, c, m, U = 1) - \lambda_T(t|a, c, m, U = 0)$ is constant across strata of a, then

$$B_{hr}^{CDE}(m|c) = \delta_m \gamma_m$$

where $\delta_m = P(U = 1|a, m, c) - P(U = 1|a^*, m, c)$.

On the hazard ratio scale the bias factor compares the true controlled direct effect hazard ratio, $\lambda_{T_{am}}(t|c)/\lambda_{T_{a^*m}}(t|c)$, with its estimate, $\lambda_T(t|a, m, c)/\lambda_T(t|a^*, m, c)$; that is, the bias factor is defined by

$$B_{hr}^{CDE}(m|c) = \frac{\lambda_T(t|a, m, c)}{\lambda_T(t|a^*, m, c)} \Big/ \frac{\lambda_{T_{am}}(t|c)}{\lambda_{T_{a^*m}}(t|c)}$$

Suppose again that for all a and m, $T_{am} \perp\!\!\!\perp A|C$ and $T_{am} \perp\!\!\!\perp M|\{A, C, U\}$, and that the outcome is rare. It follows using the argument in Proposition 3.4 from VanderWeele (2010a) for the ratio scale for probabilities, that for any reference level u' of U we have that

$$B_{hr}^{CDE}(m|c) = \frac{\sum_u \frac{\lambda_T(t|a,m,c,u)}{\lambda_T(t|a,m,c,u')} P(u|a, m, c)}{\sum_u \frac{\lambda_T(t|a,m,c,u)}{\lambda_T(t|a,m,c,u')} P(u|a, c)} \Big/ \frac{\sum_u \frac{\lambda_T(t|a^*,m,c,u)}{\lambda_T(t|a^*,m,c,u')} P(u|a^*, m, c)}{\sum_u \frac{\lambda_T(t|a^*,m,c,u)}{\lambda_T(t|a^*,m,c,u')} P(u|a^*, c)}$$

If U is binary with $U \perp\!\!\!\perp A|C$ and if $\gamma_m = \frac{\lambda_T(t|a,m,c,U=1)}{\lambda_T(t|a,m,c,U=0)}$ is constant across strata of a, then

$$B_{hr}^{CDE}(m|c) = \frac{1 + (\gamma_m - 1)P(U = 1|a, m, c)}{1 + (\gamma_m - 1)P(U = 1|a^*, m, c)}$$

For natural direct and indirect effects on the hazard difference scale we define the bias factors by

$$B_{hd}^{NDE}(c) = \sum\nolimits_m \{\lambda_T(t|a,m,c) - \lambda_T(t|a^*,m,c)\}P(m|a^*,c) - \{\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)\}$$
$$B_{hd}^{NIE}(c) = \sum\nolimits_m \lambda_T(t|a,m,c)\{P(m|a,c) - P(m|a^*,c)\} - \{\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)\}$$

Suppose that U is an unmeasured mediator–outcome confounder and that control for measured C and unmeasured U together would suffice to control for confounding of the (i) exposure–outcome, (ii) mediator–outcome, and (iii) exposure–mediator relationships, and suppose also that (iv) there is no mediator–outcome confounder affected by the exposure; that is, we assume $T_{am} \perp\!\!\!\perp A|C$, $T_{am} \perp\!\!\!\perp M|\{A,C,U\}$, $M_a \perp\!\!\!\perp A|C$, and $T_{am} \perp\!\!\!\perp M_{a^*}|\{C,U\}$, and $U \perp\!\!\!\perp A|C$. Provided that the outcome is rare, it follows using the argument in Proposition 3.5 from VanderWeele (2010) for the ratio scale for probabilities, that for any reference level u' of U the bias formula for the natural direct effect is given by

$$B_{hd}^{NDE}(c) = \sum\nolimits_m \sum\nolimits_u \{\lambda_T(t|a,m,c,u) - \lambda_T(t|a,m,c,u')\}\{P(u|a,m,c) - P(u|a^*,m,c)\}P(m|a^*,c)$$

and the bias formula for the natural indirect effect is given by

$$B_{hd}^{NIE}(c) = -\sum\nolimits_m \sum\nolimits_u \{\lambda_T(t|a,m,c,u) - \lambda_T(t|a,m,c,u')\}\{P(u|a,m,c) - P(u|a^*,m,c)\}P(m|a^*,c).$$

A.5. MULTIPLE MEDIATORS

A.5.1. Notation

Suppose now that there are multiple mediators of interest, $\mathbf{M} = (M^{(1)},\ldots,M^{(K)})$, and that we are interested in the effects mediated through $(M^{(1)},\ldots,M^{(K)})$ jointly and the effects independent of $(M^{(1)},\ldots,M^{(K)})$. We can define controlled direct effects and natural direct and indirect effects in a similar way as before simply replacing our single mediator M with the entire vector of mediators $\mathbf{M} = (M^{(1)},\ldots,M^{(K)})$. Thus, let $\mathbf{M}_a$ be the counterfactual value of $\mathbf{M}$ if exposure A were set to the value a and let $Y_{a\mathbf{m}}$ denote the counterfactual value for Y if A were set to a and $\mathbf{M}$ were set to $\mathbf{m}$. The controlled direct effect is defined by $Y_{a\mathbf{m}} - Y_{a^*\mathbf{m}}$; the natural direct effect is defined as $Y_{a\mathbf{M}_{a^*}} - Y_{a^*\mathbf{M}_{a^*}}$; the natural indirect effect is defined as $Y_{a\mathbf{M}_a} - Y_{a\mathbf{M}_{a^*}}$; and once again the total effect can be decomposed into a natural direct and indirect effect: $Y_a - Y_a = Y_{a\mathbf{M}_a} - Y_{a^*\mathbf{M}_{a^*}} = (Y_{a\mathbf{M}_a} - Y_{a\mathbf{M}_{a^*}}) + (Y_{a\mathbf{M}_{a^*}} - Y_{a^*\mathbf{M}_{a^*}})$.

Suppose again that the four assumptions about confounding hold but now with respect to the whole set of mediators $\mathbf{M} = (M^{(1)},\ldots,M^{(K)})$. In other words, suppose we have (A5.1) $Y_{a\mathbf{m}} \perp\!\!\!\perp A|C$, (A5.2) $Y_{a\mathbf{m}} \perp\!\!\!\perp \mathbf{M}|\{A,C\}$, (A5.3) $\mathbf{M}_a \perp\!\!\!\perp A|C$,

and (A5.4) $Y_{a\mathbf{m}} \perp\!\!\!\perp \mathbf{M}_{a^*} | C$. We once again need to control for all exposure–outcome, mediator–outcome, and exposure–mediator confounders, but note must that now for assumptions (A5.2) and (A5.3) the mediator–outcome confounders that must be controlled for are for all of the mediators, not just one, and likewise the exposure–mediator confounders that must be controlled for are for all of the mediators, not just one. Assumption (A5.4) again requires that there be no effect of the exposure that confounds the mediator–outcome relationship for any of the mediators. If there were such a variable, then to proceed it would have to be included in the mediator vector $\mathbf{M}$ if assumption (A5.4) were not to be violated.

A.5.2. Regression-Based Approach

Proposition 5.1 (VanderWeele and Vansteelandt, 2013):
Suppose assumptions (A5.1)–(A5.4) hold and the following regression models are correctly specified:

$$\begin{aligned}
\mathbb{E}[Y|a,\mathbf{m},c] &= \theta_0 + \theta_1 a + \sum\nolimits_{i=1}^{K} \theta_2^{(i)} m^{(i)} + \sum\nolimits_{i=1}^{K} \theta_3^{(i)} a m^{(i)} \\
\mathbb{E}[M^{(i)}|a,c] &= \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c \quad \text{for } i \text{ continuous} \\
\text{logit}\{P[M^{(j)} = 1|a,c\} &= \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c \quad \text{for } i \text{ binary}
\end{aligned}$$

then the controlled direct effects and natural direct and indirect effects are given by

$$\begin{aligned}
\mathbb{E}[Y_{a\mathbf{m}} - Y_{a^*\mathbf{m}}|c] &= \theta_1 a + \sum\nolimits_{i=1}^{K} \theta_3^{(i)} a m^{(i)} \\
\mathbb{E}[Y_{a\mathbf{M}_{a^*}} - Y_{a^*\mathbf{M}_{a^*}}|c] &= \{\theta_1 + \sum\nolimits_{i=1}^{K} \theta_3^{(i)} \mathbb{E}[M^{(i)}|c,a^*]\}(a - a^*) \\
\mathbb{E}[Y_{a\mathbf{M}_a} - Y_{a\mathbf{M}_{a^*}}|c] &= \sum\nolimits_{i=1}^{K} \{\theta_2^{(i)} + \theta_3^{(i)} a\}\{\mathbb{E}[M^{(i)}|c,a] - \mathbb{E}[M^{(i)}|c,a^*]\}
\end{aligned}$$

where $\mathbb{E}[M^{(i)}|c,a] = \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c$ if $M^{(i)}$ is continuous and $\mathbb{E}[M^{(i)}|c,a] = \frac{\exp\{\beta_0^{(i)} + \beta_1^{(j)} a + \beta_2^{(j)\prime} c\}}{1 + \exp\{\beta_0^{(i)} + \beta_1^{(j)} a + \beta_2^{(j)\prime} c\}}$ if $M^{(i)}$ is binary.

Proof:
Consider the somewhat more general regression models:

$$\begin{aligned}
\mathbb{E}[Y|a,\mathbf{m},c] &= \theta_0 + \theta_1 a + \sum\nolimits_{i=1}^{K} \theta_2^{(i)} m^{(i)} + \sum\nolimits_{i=1}^{K} \theta_3^{(i)} a m^{(i)} \\
&\quad + \sum\nolimits_{i=1,j=1}^{K,K} \tau^{(ij)} m^{(i)} m^{(j)} + \theta_4' c \\
\mathbb{E}[M^{(i)}|a,c] &= \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c \quad \text{for } i \text{ continuous} \\
\text{logit}\{P[M^{(j)} = 1|a,c\} &= \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c \quad \text{for } i \text{ binary}
\end{aligned}$$

Under assumptions (A5.1) and (A5.2), we have for the controlled direct effect

$$\begin{aligned}\mathbb{E}[Y_{a\mathbf{m}} - Y_{a^*\mathbf{m}}|c] &= \mathbb{E}[Y|a,\mathbf{m},c] - \mathbb{E}[Y|a^*,\mathbf{m},c] \\ &= \theta_1 a + \sum_{i=1}^{K} \theta_3^{(i)} a m^{(i)}\end{aligned}$$

Under assumptions (A5.1)–(A5.4) we have, by the mediation formula,

$$\begin{aligned}\mathbb{E}[Y_{a\mathbf{M}_{a^*}}|c] &= \int_{\mathbf{m}} \mathbb{E}[Y|c,a,\mathbf{m}]dP(\mathbf{m}|c,a^*) \\ &= \int_{\mathbf{m}} \{\theta_0 + \theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} m^{(i)} + \sum_{i=1}^{K} \theta_3^{(i)} a m^{(i)} \\ &\quad + \sum_{i=1,j=1}^{K,K} \tau^{(ij)} m^{(i)} m^{(j)} + \theta_4' c\} dP(\mathbf{m}|c,a^*) \\ &= \theta_0 + \theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} \mathbb{E}[M^{(i)}|c,a^*] + \sum_{i=1}^{K} \theta_3^{(i)} a \mathbb{E}[M^{(i)}|c,a^*] \\ &\quad + \sum_{i=1,j=1}^{K,K} \tau^{(ij)} \mathbb{E}[M^{(i)}M^{(j)}|c,a^*] + \theta_4' c\end{aligned}$$

Thus the natural direct effect is given by

$$\begin{aligned}\mathbb{E}[Y_{a\mathbf{M}_{a^*}} - Y_{a^*\mathbf{M}_{a^*}}|c] &= \theta_0 + \theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} \mathbb{E}[M^{(i)}|c,a^*] \\ &\quad + \sum_{i=1}^{K} \theta_3^{(i)} a \mathbb{E}[M^{(i)}|c,a^*] \\ &\quad + \sum_{i=1,j=1}^{K,K} \tau^{(ij)} \mathbb{E}[M^{(i)}M^{(j)}|c,a^*] + \theta_4' c \\ &\quad - [\theta_0 + \theta_1 a^* + \sum_{i=1}^{K} \theta_2^{(i)} \mathbb{E}[M^{(i)}|c,a^*] \\ &\quad + \sum_{i=1}^{K} \theta_3^{(i)} a^* \mathbb{E}[M^{(i)}|c,a^*] \\ &\quad + \sum_{i=1,j=1}^{K,K} \tau^{(ij)} \mathbb{E}[M^{(i)}M^{(j)}|c,a^*] + \theta_4' c] \\ &= \{\theta_1 + \sum_{i=1}^{K} \theta_3^{(i)} \mathbb{E}[M^{(i)}|c,a^*]\}(a - a^*)\end{aligned}$$

If $M^{(i)}$ is continuous, $\mathbb{E}[M^{(i)}|c,a^*] = \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c$. If $M^{(i)}$ is binary, $\mathbb{E}[M^{(i)}|c,a^*] = \frac{\exp\{\beta_0^{(i)} + \beta_1^{(j)} a^* + \beta_2^{(j)\prime} c\}}{1+\exp\{\beta_0^{(i)} + \beta_1^{(j)} a^* + \beta_2^{(j)\prime} c\}}$. The natural indirect effect is given by

$$\begin{aligned}\mathbb{E}[Y_{a\mathbf{M}_a} - Y_{a\mathbf{M}_{a^*}}|c] &= \theta_0 + \theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} \mathbb{E}[M^{(i)}|c,a] \\ &\quad + \sum_{i=1}^{K} \theta_3^{(i)} a \mathbb{E}[M^{(i)}|c,a] \\ &\quad + \sum_{i=1,j=1}^{K,K} \tau^{(ij)} \mathbb{E}[M^{(i)}M^{(j)}|c,a] + \theta_4' c \\ &\quad - \{\theta_0 + \theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} \mathbb{E}[M^{(i)}|c,a^*]\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{K}\theta_3^{(i)}a\mathbb{E}[M^{(i)}|c,a^*]\\
&+\sum_{i=1,j=1}^{K,K}\tau^{(ij)}\mathbb{E}[M^{(i)}M^{(j)}|c,a^*]+\theta_4'c\}\\
=&\sum_{i=1}^{K}\{\theta_2^{(i)}+\theta_3^{(i)}a\}\{\mathbb{E}[M^{(i)}|c,a]-\mathbb{E}[M^{(i)}|c,a^*]\}\\
&+\sum_{i=1,j=1}^{K,K}\tau^{(ij)}\{\mathbb{E}[M^{(i)}M^{(j)}|c,a]-\mathbb{E}[M^{(i)}M^{(j)}|c,a^*]\}
\end{aligned}
$$

If $M^{(i)}$ is continuous, $\mathbb{E}[M^{(i)}|c,a]-\mathbb{E}[M^{(i)}|c,a^*]=\beta_1^{(i)}(a-a^*)$. If $M^{(i)}$ is binary, $\mathbb{E}[M^{(i)}|c,a]-\mathbb{E}[M^{(i)}|c,a^*]=\frac{\exp\{\beta_0^{(i)}+\beta_1^{(i)}a+\beta_2^{(i)\prime}c\}}{1+\exp\{\beta_0^{(i)}+\beta_1^{(i)}a+\beta_2^{(i)\prime}c\}}-\frac{\exp\{\beta_0^{(i)}+\beta_1^{(i)}a^*+\beta_2^{(i)\prime}c\}}{1+\exp\{\beta_0^{(i)}+\beta_1^{(i)}a^*+\beta_2^{(i)\prime}c\}}$. ■

For the mediator–mediator interaction terms, we could consider the following models:

$$\mathbb{E}[M^{(i)}M^{(j)}|a,c]=\beta_0^{(ij)}+\beta_1^{(ij)}a+\beta_2^{(ij)\prime}c \quad \text{for at least one of } M^{(i)},M^{(j)} \text{ continuous}$$

$$\text{logit}\{P[M^{(i)}M^{(j)}=1|a,c]\}=\beta_0^{(ij)}+\beta_1^{(ij)}a+\beta_2^{(ij)\prime}c \quad \text{for } M^{(i)},M^{(j)} \text{ both binary}$$

Under these models, if at least one of $M^{(i)}$ or $M^{(j)}$ is continuous, $\mathbb{E}[M^{(i)}M^{(j)}|c,a]-\mathbb{E}[M^{(i)}M^{(j)}|c,a^*]=\beta_1^{(ij)}(a-a^*)$. If both $M^{(i)}$ and $M^{(j)}$ are binary, then $\mathbb{E}[M^{(i)}M^{(j)}|c,a]-\mathbb{E}[M^{(i)}M^{(j)}|c,a^*]=\frac{\exp\{\beta_0^{(ij)}+\beta_1^{(ij)}a+\beta_2^{(ij)\prime}c\}}{1+\exp\{\beta_0^{(ij)}+\beta_1^{(ij)}a+\beta_2^{(ij)\prime}c\}}-\frac{\exp\{\beta_0^{(ij)}+\beta_1^{(ij)}a^*+\beta_2^{(ij)\prime}c\}}{1+\exp\{\beta_0^{(ij)}+\beta_1^{(ij)}a^*+\beta_2^{(ij)\prime}c\}}$. Note, however, if covariates C are included in the model and a mediator–mediator interaction term, $\tau^{(ij)}m^{(i)}m^{(j)}$, is also included, then this can lead to issues of model compatibility between the models for $M^{(i)}$ and $M^{(j)}$ and that for the product $M^{(i)}M^{(j)}$. For continuous mediators, a possible solution would be to assume a constant covariance matrix, in which case the average product follows from knowledge of the covariance and means. For instance, we would have that

$$\mathbb{E}[M^{(i)}M^{(j)}|c,a]=(\beta_0^{(i)}+\beta_1^{(i)}a+\beta_2^{(i)\prime}c)(\beta_0^{(j)}+\beta_1^{(j)}a+\beta_2^{(j)\prime}c)+\text{Cov}(\epsilon^{(i)},\epsilon^{(j)})$$

where $\epsilon^{(i)},\epsilon^{(j)}$ are the residuals in the models for both mediators. For dichotomous mediators, the Plackett copula could be used; that is, on top of the models for each mediator separately, one could postulate the model $\text{odds}(M^{(i)}=1|M^{(j)}=1,a,c)/\text{odds}(M^{(i)}=1|M^{(j)}=0,a,c)=\alpha$ with α unknown; α could be estimated using standard software for alternating logistic regression, which is, for instance, available via the option 'logor = exch' in proc genmod. With $p^{(i)}\equiv\mathbb{E}[M^{(i)}|c,a]$ and $p^{(j)}\equiv\mathbb{E}[M^{(j)}|c,a]$, we then have that

$$
\begin{aligned}
\mathbb{E}[M^{(i)}M^{(j)}|c,a]=&\frac{1}{2(\alpha-1)}[1-(p^{(i)}+p^{(j)})(1-\alpha)\\
&-\{(1-(p^{(i)}+p^{(j)})(1-\alpha)^2-4p^{(i)}p^{(j)}\alpha(\alpha-1)\}^{1/2}].
\end{aligned}
$$

Proposition 5.2. (VanderWeele and Vansteelandt, 2013):
Suppose that the outcome is binary and rare, that assumptions (A5.1)–(A5.4) hold, and that the following regression models are correctly specified:

$$\text{logit}[P\{Y=1|a,\mathbf{m},c\}] = \theta_0 + \theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} m^{(i)} + \sum_{i=1}^{K} \theta_3^{(i)} a m^{(i)} + \theta_4' c$$
$$\mathbb{E}[M^{(i)}=1|a,c] = \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c \quad \text{for } i=1,\ldots,K$$

with the vector of mediators $\mathbf{M}$ following a multivariate normal distribution conditional on A and C with conditional covariance matrix Σ, then

$$\log\{OR^{CDE}_{a,a^*|c}(\mathbf{m})\} = \theta_1 a + \sum_{i=1}^{K} \theta_3^{(i)} a m^{(i)}$$
$$\log\{OR^{NDE}_{a,a^*|c}(a^*)\} = \{\theta_1 + \sum_{i=1}^{K} \theta_3^{(i)} (\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c + \theta_2^{(i)} \sigma^{(i)2})\}(a-a^*)$$
$$+0.5(\theta_2+\theta_3 a)' \sum (\theta_2+\theta_3 a) - 0.5(\theta_2+\theta_3 a^*)' \sum (\theta_2+\theta_3 a^*)$$
$$\log\{OR^{NIE}_{a,a^*|c}(a)\} = \sum_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a)\beta_1^{(i)}(a-a^*)$$

Proof:
Under assumptions (A5.1)–(A5.4) we then have $\text{logit}\{P(Y_{a\mathbf{M}_{a^*}}=1|c)\}$

$$\approx \log\{P(Y_{a\mathbf{M}_{a^*}}=1|c)\}$$
$$= \log\left\{\int P(Y=1|a,\mathbf{m},c)dP(\mathbf{m}|a^*,c)\right\} \quad \text{by (A5.1)–(A5.4)}$$
$$\approx \log\left\{\int \exp(\theta_0+\theta_1 a + \sum_{i=1}^{K} \theta_2^{(i)} m^{(i)} + \sum_{i=1}^{K} \theta_3^{(i)} a m^{(i)} + \theta_4' c) dP(\mathbf{m}|a^*,c)\right\}$$
$$= \log\left\{\exp(\theta_0+\theta_1 a+\theta_4' c)\int \exp\left\{\sum_{i=1}^{K} \theta_2^{(i)} m^{(i)} + \theta_3^{(i)} a m^{(i)}\right\} dP(\mathbf{m}|a^*,c)\right\}$$
$$= \log\left\{\exp(\theta_0+\theta_1 a+\theta_4' c)\mathbb{E}[e^{\sum_{i=1}^{K}(\theta_2^{(i)}+\theta_3^{(i)}a)M^{(i)}}|a^*,c]\right\}$$

Note that conditional on a^* and c, $\sum_{i=1}^{K}(\theta_2^{(i)} + \theta_3^{(i)} a)M^{(i)}$ follows a normal distribution with mean $\sum_{i=1}^{K}(\theta_2^{(i)} + \theta_3^{(i)} a)(\beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c)$ and variance $(\theta_2+\theta_3 a)'\Sigma(\theta_2+\theta_3 a)$, where $\theta_2 = (\theta_2^{(1)},\ldots,\theta_2^{(K)})'$ and $\theta_3 = (\theta_3^{(1)},\ldots,\theta_3^{(K)})'$. It thus follows that

$$\text{logit}\{P(Y_{a\mathbf{M}_{a^*}}=1|c)\}$$
$$= \log\left\{\exp(\theta_0+\theta_1 a+\theta_4' c)\exp\left(\sum_{i=1}^{K}(\theta_2^{(i)}+\theta_3^{(i)}a)\left(\beta_0^{(i)}+\beta_1^{(i)}a^*+\beta_2^{(i)\prime}c\right)\right.\right.$$
$$\left.\left.+\frac{1}{2}(\theta_2+\theta_3 a)'\Sigma(\theta_2+\theta_3 a)\right)\right\}$$

$$= \theta_0 + \theta_1 a + \theta_4' c + \sum\nolimits_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a)$$
$$\times (\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c) + \frac{1}{2}(\theta_2 + \theta_3 a)' \sum (\theta_2 + \theta_3 a)$$

The log of natural direct effect odds ratio is then given by

$$\begin{aligned}
&\log\{OR^{NDE}_{a,a^*|c}(a^*)\} \\
&\quad = \text{logit}\{P(Y_{a\mathbf{M}_{a^*}} = 1|c)\} - \text{logit}\{P(Y_{a^*\mathbf{M}_{a^*}} = 1|c)\} \\
&\quad \approx \theta_0 + \theta_1 a + \theta_4' c + \sum\nolimits_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a)(\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c) \\
&\qquad + \tfrac{1}{2}(\theta_2 + \theta_3 a)' \sum (\theta_2 + \theta_3 a). \\
&\qquad - \{\theta_0 + \theta_1 a^* + \theta_4' c + \sum\nolimits_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a^*)(\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c) \\
&\qquad + \tfrac{1}{2}(\theta_2 + \theta_3 a^*)' \sum (\theta_2 + \theta_3 a^*)\} \\
&\quad = \{\theta_1 + \sum\nolimits_{i=1}^{K} \theta_3^{(i)} (\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c + \theta_2^{(i)} \sigma^{(i)2})\}(a - a^*) \\
&\qquad + 0.5(\theta_2 + \theta_3 a)' \sum (\theta_2 + \theta_3 a) - 0.5(\theta_2 + \theta_3 a^*)' \sum (\theta_2 + \theta_3 a^*)
\end{aligned}$$

The log of natural indirect effect odds ratio is then given by

$$\begin{aligned}
&\log\{OR^{NIE}_{a,a^*|c}(a)\} \\
&\quad = \text{logit}\{P(Y_{a\mathbf{M}_a} = 1|c)\} - \text{logit}\{P(Y_{a\mathbf{M}_{a^*}} = 1|c)\} \\
&\quad \approx \theta_0 + \theta_1 a + \theta_4' c + \sum\nolimits_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a)(\beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c) \\
&\qquad + \tfrac{1}{2}(\theta_2 + \theta_3 a)' \sum (\theta_2 + \theta_3 a) \\
&\qquad - [\theta_0 + \theta_1 a + \theta_4' c + \sum\nolimits_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a)(\beta_0^{(i)} + \beta_1^{(i)} a^* + \beta_2^{(i)\prime} c) \\
&\qquad + \tfrac{1}{2}(\theta_2 + \theta_3 a)' \sum (\theta_2 + \theta_3 a)] \\
&\quad = \sum\nolimits_{i=1}^{K} (\theta_2^{(i)} + \theta_3^{(i)} a)\beta_1^{(i)} (a - a^*). \quad \blacksquare
\end{aligned}$$

Suppose we were to apply the regression-based approach sequentially. Let $\overline{\mathbf{M}}^{(k)}$ be the subset of mediators $(M^{(1)}, \ldots, M^{(k)})$. Consider the regression models

$$\mathbb{E}[Y|a, \overline{\mathbf{m}}^{(k)}, c] = \theta_0 + \theta_1 a + \sum\nolimits_{i=1}^{k} \theta_2^{(i)} m^{(i)} + \sum\nolimits_{i=1}^{k} \theta_3^{(i)} a m^{(i)} + \theta_4' c$$
$$\mathbb{E}[M^{(i)}|a, c] = \beta_0^{(i)} + \beta_1^{(i)} a + \beta_2^{(i)\prime} c$$

for $i = 1, \ldots, k$. Under assumptions (A5.1)–(A5.4), we then have by Pearl's mediation formula that the exposure effect that is mediated by the first k mediators

equals

$$\mathbb{E}[Y_{a\overline{\mathbf{M}}_a^{(k)}} - Y_{a\overline{\mathbf{M}}_{a^*}^{(k)}}|c] = \sum_{i=1}^{k}\{\theta_2^{(i)} + \theta_3^{(i)}a\}\{\mathbb{E}[M^{(i)}|c,a] - \mathbb{E}[M^{(i)}|c,a^*]\}$$
$$\sum_{i=1}^{k}\{\theta_2^{(i)} + \theta_3^{(i)}a\}\beta_1^{(i)}(a-a^*)$$

While this approach is valid for each fixed k, a concern is that the models for $\mathbb{E}[Y|a,\overline{\mathbf{m}}^{(k)},c]$ and $\mathbb{E}[M^{(k)}|a,c]$ may not be compatible across time points k. For instance, suppose that

$$\mathbb{E}[M^{(i)}|\overline{\mathbf{m}}^{(i-1)},a,c] = \gamma_0^{(i)} + \gamma_1^{(i)}a + \gamma_2^{(i)\prime}c + \gamma_3^{(i)\prime}\overline{\mathbf{m}}^{(i-1)}$$

for $i = 1,\ldots,k$, which is compatible with the aforementioned models for $\mathbb{E}[M^{(i)}|a,c]$. Then the model for $\mathbb{E}[Y|a,\overline{\mathbf{m}}^{(k)},c]$ implies that

$$\mathbb{E}[Y|a,\overline{\mathbf{m}}^{(k-1)},c] = \theta_0 + \theta_1 a + \sum_{i=1}^{k-1}\theta_2^{(i)}m^{(i)} + \sum_{i=1}^{k-1}\theta_3^{(i)}am^{(i)} + \theta_4'c$$
$$+(\theta_2^{(k)} + \theta_3^{(k)}a)(\gamma_0^{(i)} + \gamma_1^{(i)}a + \gamma_2^{(i)\prime}c + \gamma_3^{(i)\prime}\overline{\mathbf{m}}^{(i-1)})$$

This model is no longer of the same form because it includes interactions between a and c, as well as squared terms a^2 that were not previously allowed for. This can be remedied by extending the outcome regression model to include such terms:

$$\mathbb{E}[Y|a,\overline{\mathbf{m}}^{(k)},c] = \theta_0 + \theta_1 a + \sum_{i=1}^{k}\theta_2^{(i)}m^{(i)} + \sum_{i=1}^{k}\theta_3^{(i)}am^{(i)} + \theta_4'c + \theta_5'ac + \theta_6'a^2.$$

With this extension one still has

$$\mathbb{E}[Y_{a\overline{\mathbf{M}}_a^{(k)}} - Y_{a\overline{\mathbf{M}}_{a^*}^{(k)}}|c] = \sum_{i=1}^{k}\{\theta_2^{(i)} + \theta_3^{(i)}a\}\beta_1^{(i)}(a-a^*).$$

If the exposure is binary and we allow for exposure–covariate interaction in the outcome model then this problem of model incompatibility is thus remedied. Alternatively, if there are no exposure–mediator interactions in the outcome model, then the models will remain compatible with each other.

A.5.3. Weighting Approach

The following proposition justifies the weighting approach for multiple mediators described in the text. It is given in VanderWeele and Vansteelandt (2013) and is a simple extension of the approach in Albert (2012).

Proposition 5.3. (VanderWeele and Vansteelandt, 2013; cf. Albert, 2012):
Under assumptions (A5.1)–(A5.4), $\mathbb{E}[Y_{a\mathbf{M}_{a^*}}] = E\left[\frac{I(A=a^*)}{P(A=a^*|c)}\mathbb{E}(Y|A=a,\mathbf{M},c)\right]$.

Proof:
Under (A5.1)–(A5.4),

$$\begin{aligned}\mathbb{E}[Y_{a\mathbf{M}_{a^*}}] &= \int \mathbb{E}(Y|A=a,\mathbf{M}=\mathbf{m},c)f(\mathbf{M}=\mathbf{m}|A=a^*,c)f(c)\,d\mathbf{m}dc \\ &= \int I(A=a^*)\mathbb{E}(Y|A=a,\mathbf{M},c)f(\mathbf{M}|A,c)f(c)d\mathbf{M}\,dAdc \\ &= \int \frac{I(A=a^*)}{P(A=a^*|c)}\mathbb{E}(Y|A=a,\mathbf{M},c)f(\mathbf{M},A,c)\,d\mathbf{M}dAdc \\ &= E\left[\frac{I(A=a^*)}{P(A=a^*|c)}\mathbb{E}(Y|A=a,\mathbf{M},c)\right]. \quad \blacksquare\end{aligned}$$

A.5.4. Sum of Individual Mediated Effects Versus Joint Mediated Effects

Consider two mediators, $M^{(1)}$ and $M^{(2)}$, and suppose that neither affects the other. For simplicity assume a binary exposure. The natural indirect effect through $M^{(1)}$ is, by definition, $Y_{1M_1^{(1)}} - Y_{1M_0^{(1)}}$. The natural indirect effect through $M^{(2)}$ is, by definition, $Y_{1M_1^{(2)}} - Y_{1M_0^{(2)}}$. The natural indirect effect through $(M^{(1)}M^{(2)})$ is, by definition, $Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_0^{(2)}}$. If the mediators do not affect each other, then natural indirect effect through $M^{(1)}$ is equal to $Y_{1M_1^{(1)}} - Y_{1M_0^{(1)}} = Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}}$ and the natural indirect effect through $M^{(2)}$ is equal to $Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}}$. The sum of the two natural indirect effects for $M^{(1)}$ and $M^{(2)}$ considered separately is thus $\{Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}}\} + \{Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}}\}$. The difference between the sum of the two natural indirect effects for $M^{(1)}$ and $M^{(2)}$ considered separately and the natural indirect effect through $(M^{(1)}M^{(2)})$ jointly is then given by

$$\begin{aligned}&\{Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}}\} + \{Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}}\} - \{Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_0^{(2)}}\} \\ &\quad = Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}} + Y_{1M_0^{(1)}M_0^{(2)}}\end{aligned}$$

This difference in some sense captures the effect mediated by the interaction between $M^{(1)}$ and $M^{(2)}$.

We now show that if the mediators do not affect each other and if there is no interaction between $M^{(1)}$ and $M^{(2)}$ at the individual counterfactual level, then the sum of the two natural indirect effects for $M^{(1)}$ and $M^{(2)}$ considered separately and the natural indirect effect through $(M^{(1)}M^{(2)})$ jointly must be equal. We will say that there is no interaction between $M^{(1)}$ and $M^{(2)}$ at the individual counterfactual level if for any two values $m^{(1)}, m^{(1)\dagger}$ of $M^{(1)}$, $Y_{am^{(1)}m^{(2)}} - Y_{am^{(1)\dagger}m^{(2)}}$ is constant across $m^{(2)}$ or, equivalently, for any two values $m^{(2)}, m^{(2)\dagger}$ of $M^{(2)}$, $Y_{am^{(1)}m^{(2)}} - Y_{am^{(1)}m^{(2)\dagger}}$ is constant across $m^{(1)}$. If this is the case, then $Y_{1M_1^{(1)}m^{(2)}} - Y_{1M_0^{(1)}m^{(2)}}$ must be constant across $m^{(2)}$ and thus $Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}}$ must be equal to

$Y_{1M_1^{(1)}M_0^{(2)}} - Y_{1M_0^{(1)}M_0^{(2)}}$ and so $Y_{1M_1^{(1)}M_1^{(2)}} - Y_{1M_0^{(1)}M_1^{(2)}} - Y_{1M_1^{(1)}M_0^{(2)}} + Y_{1M_0^{(1)}M_0^{(2)}}$ must be equal to 0.

A.5.5. Marginal Structural Models for Controlled Direct Effects in the Presence of Exposure-Induced Confounding

Suppose that there is one or more mediator–outcome confounders L affected by exposure A as in Figure 5.4. Suppose also that the effect of A on Y is unconfounded conditional on baseline covariates C, in counterfactual notation (A2.1) $Y_{am} \perp\!\!\!\perp A|C$, and that (A5.5) the effect of M on Y is unconfounded conditional on (L, A, C), in counterfactual notation $Y_{am} \perp\!\!\!\perp M|\{A, C, L\}$, as would be the case in Figures 5.4 and 5.5.

The following proposition justifies the weighting approach to controlled direct effects in the presence of exposure-induced confounding described in the text. It is a simple example of the marginal structural model of Robins (1999a) with the treatment at period 1 taken as the exposure and the treatment at period 2 taken as the mediator (cf. VanderWeele, 2009a).

Proposition 5.4 (Robins, 1986, 1999a):
Under assumptions (A2.1) $Y_{am} \perp\!\!\!\perp A|C$ and (A5.5) $Y_{am} \perp\!\!\!\perp M|\{A, C, L\}$:

$$\begin{aligned}\mathbb{E}[Y_{am}] &= \sum\nolimits_{l,c} \mathbb{E}[Y|a, l, m, c]P(l|a, c)P(c) \\ &= E\left[\frac{Y}{P(A|C)P(M|A, C, L)}\right]\end{aligned}$$

Proof:
Under (A2.1) and (A5.5),

$$\begin{aligned}\mathbb{E}[Y_{am}] &= \sum\nolimits_{c} \mathbb{E}[Y_{am}|c]P(c) \\ &= \sum\nolimits_{c} \mathbb{E}[Y_{am}|a, c]P(c) \\ &= \sum\nolimits_{l,c} \mathbb{E}[Y_{am}|a, c, l]P(l|a, c)P(c) \\ &= \sum\nolimits_{l,c} \mathbb{E}[Y_{am}|a, m, c, l]P(l|a, c)P(c) \\ &= \sum\nolimits_{l,c} \mathbb{E}[Y|a, m, c, l]P(l|a, c)P(c)\end{aligned}$$

where the second equality follows by (A2.1) and the fourth equality by (A5.5). We moreover have

$$\begin{aligned}&\sum\nolimits_{l,c} \mathbb{E}[Y|a, m, l, c]P(l|a, c)P(c) = \sum\nolimits_{l,c,y} yP(y|a, m, l, c)P(l|a, c)P(c) \\ &= \sum\nolimits_{l,c,y} yP(y|a, m, l, c)\frac{P(m|l, a, c)}{P(m|l, a, c)}P(l|a, c)\frac{P(a|c)}{P(a|c)}P(c)\end{aligned}$$

$$= \sum_{l,c,y} yP(y,a,m,l,c)\frac{1}{P(m|l,a,c)}\frac{1}{P(a|c)}$$
$$= \mathbb{E}\left[\frac{Y}{P(A|C)P(M|A,C,L)}\right]. \quad \blacksquare$$

See Robins (1999a) and Robins et al. (2000) for more on statistical inference with marginal structural models.

A.5.6. Structural Mean Models for Controlled Direct Effects in the Presence of Exposure-Induced Confounding

The following proposition justifies the structural mean model approach to estimating controlled direct effects in the presence of exposure–induced mediator–outcome confounding.

Proposition 5.5. (Vansteelandt, 2009; Joffe and Greene, 2009):
Suppose we have the model $\mathbb{E}[Y_{am} - Y_{0m}] = \psi_1 a + \psi_2 am$ and we fit

$$\mathbb{E}[Y|a,m,l,c] = \kappa_0 + \kappa_1 a + \kappa_2 m + \kappa_3 am + \kappa_4' c + \kappa_5' l.$$
$$\mathbb{E}[\widehat{Y}|a,c] = \gamma_0 + \gamma_1 a + \gamma_2' c$$

where $\widehat{Y} = Y - \widehat{\kappa}_2 m + \widehat{\kappa}_3 am$. Under assumptions (A2.1) and (A5.5) we have

$$\psi_1 = \gamma_1$$
$$\psi_2 = \kappa_3$$

Proof:
We have $\mathbb{E}[Y_{am+1} - Y_{0m+1}] - \mathbb{E}[Y_{am} - Y_{0m}] = [\gamma_1 a + \gamma_2 a(m+1)] - [\gamma_1 a + \gamma_2 am] = \psi_2 a$. By Proposition 5.4, we have

$$\begin{aligned}
&\mathbb{E}[Y_{am+1} - Y_{0m+1}] - \mathbb{E}[Y_{am} - Y_{0m}] \\
&\quad = \sum_{l,c}\{\mathbb{E}[Y|a,l,m+1,c] - \mathbb{E}[Y|a,l,m+1,c]\}P(l|a,c)P(c) \\
&\qquad - \sum_{l,c}\{\mathbb{E}[Y|A=0,l,m+1,c] - \mathbb{E}[Y|A=0,l,m+1,c]\}P(l|a,c)P(c) \\
&\quad = \kappa_3 a
\end{aligned}$$

Thus $\kappa_3 = \psi_2$. Now consider the general model

$$\mathbb{E}(Y|A=a,M=m,L,c) = E\{Y_{am}|A=a,M=m,L,c\}$$
$$= q_a(a,L,c;\gamma) + q_m(m,a,L,c;\gamma)$$

where $q_a(a,L,c;\gamma)$ and $q_m(m,a,L,c;\gamma)$ are arbitrary known functions of an unknown finite-dimensional parameter γ, satisfying $q_m(0,a,L,c;\gamma) = 0$; for example, the model above corresponds with $q_a(a,l,c;\gamma) = \kappa_0 + \kappa_1 a + \kappa_4' c + \kappa_5' l$ and $q_m(m,a,L,c;\gamma) = \kappa_2 m + \kappa_3 am$. We then have $\mathbb{E}(Y|A=a,M=0,L,c) = q_a(a,L,c;\gamma) = \mathbb{E}(Y|A=a,M=m,L,c) - q_m(m,a,L,c;\gamma)$.

Under the assumption (A5.5), namely $Y_{am} \perp\!\!\!\perp M|A = a, L, C = c$, we have that

$$\begin{aligned} E\{Y_{a0}|A = a, L, c\} &= E\{Y_{a0}|A = a, M = 0, L, c\} \\ &= E\{Y - q_m(m, a, L, c; \gamma)|A = a, M = m, L, c\} \end{aligned}$$

It follows under the model $\mathbb{E}[Y_{am} - Y_{0m}] = \psi_1 a + \psi_2 am$ that

$$\begin{aligned} E\{\widehat{Y}|A, c\} &= E\{Y - q_m(M, A, L, c; \gamma)|A, c\} \\ &= E\left[E\{Y - q_m(M, A, L, c; \gamma)|M, A, L\}|A, c\right] \\ &= E\{Y_{a0}|A, c\} \\ &= E\{Y_{00} - \psi_1 A|A, c\} \\ &= E\{Y_{00}|A, c\} + \gamma_1 A \\ &= E\{Y_{00}|c\} + \gamma_1 A \end{aligned}$$

Since $\mathbb{E}[\widehat{Y}|a, c] = \gamma_0 + \gamma_1 a + \gamma_2' c$, we have $\psi_1 = \gamma_1$. ■

A.5.7. Effect Decomposition and Exposure-Induced Confounding

Let $G_{a^*|C}$ denote a random draw from the distribution of the mediator when setting the exposure to a^* amongst those with covariates C. We then have that $NIE^R = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})$, our randomized interventional analogue of the natural indirect effect, is the effect on the outcome of randomly assigning an individual who is given the exposure to a value of the mediator from the distribution of the mediator amongst those given exposure versus no exposure (conditional on covariates); this is an effect through the mediator; $NDE^R = \mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})$, our randomized interventional analogue of the natural direct effect, is a direct effect comparing exposure versus no exposure with the mediator in both cases randomly drawn from the distribution of the population when given no exposure (conditional on covariates). The effect $TE^R = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})$, our randomized interventional analogue of the total effect, compares the expected outcome when having the exposure with the mediator randomly drawn from the distribution of the population when given the exposure (conditional on covariates) to the expected outcome when not having the exposure with the mediator randomly drawn from the distribution of the population when not exposed (conditional on covariates). With effects thus defined we have the decomposition $\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}}) = \{\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})\} + \{\mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})\}$ so that the overall effect decomposes into the sum of the effect through the mediator and the direct effect.

Proposition 5.6 (VanderWeele et al., 2014c):
If (A2.1) $Y_{am} \perp\!\!\!\perp A|C$, (A2.3) $M_a \perp\!\!\!\perp A|C$, and (A5.5) $Y_{am} \perp\!\!\!\perp M|\{A, C, L\}$, then

$$\begin{aligned} \mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}}) = \sum\nolimits_{c,l,m} &\{\mathbb{E}[Y|a, l, m, c]P(l|a, c) \\ &- \mathbb{E}[Y|a^*, l, m, c]P(l|a^*, c)\}P(m|a^*, c)P(c) \end{aligned}$$

$$\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}}) = \sum\nolimits_{c,l,m} \mathbb{E}[Y|a,l,m,c]P(l|a,c) \times \{P(m|a,c) - P(m|a^*,c)\}P(c)$$

Proof:
We have that

$$\begin{aligned}
\mathbb{E}(Y_{aG_{a^*|c}}|c) &= \sum\nolimits_m \mathbb{E}[Y_{am}|c, G_{a^*|c} = m]P(G_{a^*|c} = m|c) \\
&= \sum\nolimits_m \mathbb{E}[Y_{am}|c]P(M_{a^*} = m|c) \\
&= \sum\nolimits_m \mathbb{E}[Y_{am}|a,c]P(M_{a^*} = m|a^*,c) \text{ by (A2.1) and (A2.3)} \\
&= \sum\nolimits_{l,m} \mathbb{E}[Y_{am}|a,l,c]P(l|a,c)P(M_{a^*} = m|a^*,c) \\
&= \sum\nolimits_{l,m} \mathbb{E}[Y_{am}|a,l,m,c]P(l|a,c)P(M_{a^*} = m|a^*,c) \text{ by (A5.5)} \\
&= \sum\nolimits_{l,m} \mathbb{E}[Y|a,l,m,c]P(l|a,c)P(m|a^*,c)
\end{aligned}$$

Similarly, $\mathbb{E}(Y_{aG_{a|c}}|c) = \sum_{l,m} \mathbb{E}[Y|a,l,m,c]P(l|a,c)P(m|a,c)$ and $\mathbb{E}(Y_{a^*G_{a^*|c}}|c) = \sum_{l,m} \mathbb{E}[Y|a^*,l,m,c]P(l|a^*,c)P(m|a,c)$. Subtracting gives the expressions for $\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)$ and $\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{aG_{a^*|c}}|c)$. ■

Proposition 5.7. (VanderWeele et al., 2014c):
A weighting-based estimator for the conditional natural direct and indirect effect can be obtained upon duplicating the dataset and adding an exposure variable A^* which is 0 for the first replication and 1 for the second. For each individual, a weight is obtained by taking calculating

$$\frac{\sum_l P(m|l,a^*,c)P(l|a^*,c)}{P(m|l,a,c)}$$

If a model for the outcome is fit conditional on the two exposures A and A^* and covariates on the duplicated dataset using weighted regression, the randomized interventional analogues of the natural direct and indirect effects were obtained as the coefficients of A and A^*, respectively.

Proof:
The validity of this proposed weighting estimator can be shown noting that it is obtained under a marginal structural model for the composite counterfactual $Y_{aG_{a^*|c}}$, for example,

$$\mathbb{E}[Y_{aG_{a^*|c}}|c] = \beta_0 + \beta_1 a + \beta_2 a^* + \beta_3 c$$

where $\beta_1(a - a^*) = \mathbb{E}[Y_{aG_{a^*|c}}|c] - \mathbb{E}[Y_{a^*G_{a^*|c}}|c]$ and $\beta_2(a - a^*) = \mathbb{E}[Y_{aG_{a|c}}|c] - \mathbb{E}[Y_{aG_{a^*|c}}|c]$. Upon letting a^* take all possible values over the support of A and

noting that the expression for $\mathbb{E}[Y_{aG_{a^*|c}}|c]$ can be equivalently rewritten as

$$\sum_{y,l,m} yP(y,l,m|c,a)\frac{P(l|a,c)P(m|a^*,c)}{P(l,m|c,a)} = E\left(Y\frac{P(M|c,a^*)}{P(M|l,c,a)}|c,a\right)$$

the proposed weighting estimator is obtained. For marginal natural direct and indirect effects, we use that $\mathbb{E}[Y_{aG_{a^*|c}}]$ can be equivalently rewritten as

$$\sum_{y,l,m} yP(y,l,m|c,a)P(c)\frac{P(l|a,c)P(m|a^*,c)}{P(l,m|c,a)} = E\left(YI(A=a)\frac{P(M|c,a^*)}{P(a|c)P(M|l,c,a)}|c,a\right)$$

from which the proposed estimators are obtained. ■

A.5.8. Path-Specific Effects

Let M_{al} denote the value of M that would be observed if A were set to a and L to l and let Y_{alm} be the value of Y that would be observed if A were set to a, L to l, and M to m. Although we cannot identify the effects mediated through pathways involving M (i.e., the combination of $A \to L \to M \to Y$ and $A \to M \to Y$) and the effects through pathways not involving M (i.e., the combination of $A \to Y$ and $A \to L \to Y$), Avin et al. (2005) showed that we can identify the effects (i) through pathways involving neither L nor M (i.e. $A \to Y$) (ii) through the additional pathways not involving L (i.e., $A \to M \to Y$), and (iii) through the pathways involving L (i.e., the combination of $A \to L \to M \to Y$ and $A \to L \to Y$). For simplicity, let us refer to these effects as $E_{A\to Y}$, $E_{A\to M\to Y}$, and $E_{A\to LY}$. Formally, the conditional effects $E_{A\to Y}(c)$, $E_{A\to M\to Y}(c)$, and $E_{A\to LY}(c)$ can be defined as follows: $E_{A\to Y}(c) = \mathbb{E}[Y_{aL_{a^*}M_{a^*}} - Y_{a^*L_{a^*}M_{a^*}}|c]$, $E_{A\to M\to Y}(c) = \mathbb{E}[Y_{aL_{a^*}M_{aL_{a^*}}} - Y_{aL_{a^*}M_{a^*}}|c]$, and $E_{A\to LY}(c) = \mathbb{E}[Y_{aL_aM_a} - Y_{aL_{a^*}M_{aL_{a^*}}}|c]$. We then have the following effect decomposition:

$$\begin{aligned} Y_a - Y_{a^*} &= Y_{aL_aM_a} - Y_{a^*L_{a^*}M_{a^*}} \\ &= (Y_{aL_aM_a} - Y_{aL_{a^*}M_{aL_{a^*}}}) + (Y_{aL_{a^*}M_{aL_{a^*}}} - Y_{aL_{a^*}M_{a^*}}) \\ &\quad + (Y_{aL_{a^*}M_{a^*}} - Y_{a^*L_{a^*}M_{a^*}}) \end{aligned}$$

and thus $\mathbb{E}[Y_a - Y_{a^*}|c] = E_{A\to LY}(c) + E_{A\to M\to Y}(c) + E_{A\to Y}(c)$. Marginal effects $E_{A\to LY}(c)$, $E_{A\to M\to Y}(c)$, and $E_{A\to Y}(c)$ are then given by $\sum_c E_{A\to LY}(c)P(c)$, $\sum_c E_{A\to M\to Y}(c)P(c)$, and $\sum_c E_{A\to Y}(c)P(c)$, respectively. Avin et al. (2005) showed the that these effects were identified under the causal diagram in Figure 5.4. VanderWeele et al. (2014c) gave identifying expressions and discussed weighting estimators for these three effects.

Proposition 5.8. (Avin et al., 2005; cf. VanderWeele et al., 2014c):
Suppose Figure 5.4 is a causal diagram, then

$$E_{A\to Y}(c) = \sum_{l,m}\{\mathbb{E}[Y|c,a,l,m] - \mathbb{E}[Y|c,a^*,l,m]\}P(l,m|a^*,c)$$

$$E_{A\to M\to Y}(c)=\sum\nolimits_{l,m}\mathbb{E}[Y|c,a,l,m]\{P(m|c,a,l)-P(m|c,a^*,l)\}P(l|c,a^*)$$

$$E_{A\to LY}(c)=\sum\nolimits_{l,m}\mathbb{E}[Y|c,a,l,m]P(m|c,a,l)\{P(l|c,a)-P(l|c,a^*)\}$$

Proof:
If Figure 5.4 is a causal diagram, then it follows that $(\mathrm{i}^\dagger)$ $Y_{alm}\perp\!\!\!\perp A|C$, $(\mathrm{ii}^\dagger)$ $Y_{alm}\perp\!\!\!\perp(L,M)|\{A,C\}$, $(\mathrm{iii}^\dagger)$ $(L_a,M_a)\perp\!\!\!\perp A|C$, $(\mathrm{iv}^\dagger)$ $Y_{alm}\perp\!\!\!\perp(L_{a^*},M_{a^*})|C$, $(\mathrm{v}^\dagger)$ $Y_{alm}\perp\!\!\!\perp(L_{a^*},M_{al})|C$ $(\mathrm{vi}^\dagger)$ $M_{al}\perp\!\!\!\perp L_{a^*}|C$, and $(\mathrm{vii}^\dagger)$ $M_{al}\perp\!\!\!\perp(A,L)|C$ (Avin et al., 2005; Pearl, 2009). We have already shown that under $(\mathrm{i}^\dagger)$–$(\mathrm{iv}^\dagger)$ we have

$$\begin{aligned}\mathbb{E}[Y_{aL_{a^*}M_{a^*}}|c]&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c,L_{a^*}=l,M_{a^*}=m]P(L_{a^*}=l,M_{a^*}=m|c)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c]P(L_{a^*}=l,M_{a^*}=m|c)\text{ by }(\mathrm{iv}^\dagger)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c,a]P(L_{a^*}=l,M_{a^*}=m|c,a^*)\text{ by }(\mathrm{i}^\dagger)\text{ and }(\mathrm{iii}^\dagger)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c,a,l,m]P(L_{a^*}=l,M_{a^*}=m|c,a^*)\text{ by }(\mathrm{ii}^\dagger)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y|c,a,l,m]P(l,m|c,a^*)\end{aligned}$$

Under $(\mathrm{i}^\dagger)$–$(\mathrm{vii}^\dagger)$, we have

$$\begin{aligned}\mathbb{E}[Y_{aL_{a^*}M_{aL_{a^*}}}|c]&=\sum\nolimits_{l}\mathbb{E}[Y_{alM_{al}}|c,L_{a^*}=l]P(L_{a^*}=l|c)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c,L_{a^*}=l,M_{al}=m]P(M_{al}=m|c,L_{a^*}=l)\\&\quad P(L_{a^*}=l|c)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c]P(M_{al}=m|c)P(L_{a^*}=l|c)\text{ by }(\mathrm{v}^\dagger)\text{ and }(\mathrm{vi}^\dagger)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c,a]P(M_{al}=m|c,a,l)P(L_{a^*}=l|c,a^*)\\&\quad\text{by }(\mathrm{i}^\dagger)\text{ and }(\mathrm{iii}^\dagger)\text{ and }(\mathrm{vii}^\dagger)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y_{alm}|c,a,l,m]P(M_{al}=m|c,a,l)P(L_{a^*}=l|c,a^*)\text{ by }(\mathrm{ii}^\dagger)\\&=\sum\nolimits_{l,m}\mathbb{E}[Y|c,a,l,m]P(m|c,a,l)P(l|c,a^*)\end{aligned}$$

Since $\mathbb{E}[Y_{aL_aM_a}|c]=\sum_{l,m}\mathbb{E}[Y|c,a,l,m]P(l,m|c,a)$ and $\mathbb{E}[Y_{a^*L_{a^*}M_{a^*}}|c]=\sum_{l,m}\mathbb{E}[Y|c,a^*,l,m]P(l,m|c,a^*)$, we thus have that

$$E_{A\to Y}(c)=\sum\nolimits_{l,m}\{\mathbb{E}[Y|a,l,m,c]-\mathbb{E}[Y|a^*,l,m,c]\}P(l,m|a^*,c)$$

$$E_{A\to M\to Y}(c)=\sum\nolimits_{l,m}\mathbb{E}[Y|c,a,l,m]\{P(m|c,a,l)-P(m|c,a^*,l)\}P(l|c,a^*)$$

$$E_{A\to LY}(c)=\sum\nolimits_{l,m}\mathbb{E}[Y|c,a,l,m]P(m|c,a,l)\{P(l|c,a)-P(l|c,a^*)\}$$

Marginal effects are similarly obtained, but require additional averaging over the distribution $P(c)$ of C. ■

Proposition 5.9 (VanderWeele et al., 2014c):
A weighting-based estimator for the conditional effects $E_{A\rightarrow LY}(c), E_{A\rightarrow M\rightarrow Y}(c)$ and $E_{A\rightarrow Y}(c)$ can be obtained upon merging three copies of the dataset and adding exposure variables A^* and A^{**}, where A^* equals the observed exposure for the first replication and $1-A$ for the next two replications and where A^{**} equals the observed exposure for the first two replications and $1-A$ for the third replication. For each individual, a weight is now obtained by

$$\frac{P(l|a^*,c)P(m|l,a^{**},c)}{P(l|a,c)P(m|l,a,c)}$$

If a model for the outcome is now fitted conditional on the three exposures A, A^*, and A^{**} and covariates on the obtained dataset using weighted regression, the effects $E_{A\rightarrow Y}$, $E_{A\rightarrow LY}$, and $E_{A\rightarrow M\rightarrow Y}$ of interest are obtained as the coefficients of A, A^*, and A^{**}, respectively.

Proof:
The validity of this approach can be shown by making reference to a marginal structural model for the composite counterfactual $Y_{aL_{a^*}M_{a^{**}L_{a^*}}}$, for example,

$$\mathbb{E}[Y_{aL_{a^*}M_{a^{**}L_{a^*}}}|c] = \beta_0 + \beta_1 a + \beta_2 a^* + \beta_3 a^{**} + \beta_4 c,$$

from which it is easily verified that $\beta_1(a-a^*) = E_{A\rightarrow Y}(c), \beta_2(a-a^*) = E_{A\rightarrow LY}(c)$ and $\beta_3(a-a^*) = E_{A\rightarrow M\rightarrow Y}(c)$. Upon letting a^* and a^{**} take all possible values over the support of A and noting that

$$\begin{aligned}\mathbb{E}[Y_{aL_{a^*}M_{a^{**}L_{a^*}}}|c] &= \sum\nolimits_{y,l,m} yP(y|c,a,l,m)P(m|c,a^{**},l)P(l|c,a^*)\\ &= E\left(Y\frac{P(M|c,a^{**},l)P(L|c,a^*)}{P(M|c,a,l)P(L|c,a)}|a,c\right)\end{aligned}$$

the proposed weighting estimator is obtained. Marginal effects are similarly obtained, but require additional weighting by the reciprocal of $P(a|c)$. ■

A.5.9. Sensitivity Analysis for Exposure-Induced Confounding

Suppose now that exposure is randomized and that we proceed to attempt to estimate natural direct and indirect effects using methods such as those described in Chapter 2. These methods estimate

$$Q_{NIE} = \sum_m \mathbb{E}[Y|A=1,m,c]\{P(m|A=1,c) - P(m|A=0,c)\}$$
$$Q_{NDE} = \sum_m \{\mathbb{E}[Y|A=1,m,c] - \mathbb{E}[Y|A=0,m,c]\}P(m|A=0,c)$$

These expressions Q_{NIE} and Q_{NDE} will be consistent for the natural indirect and direct effects, respectively, if assumptions (A2.1)–(A2.4) hold. Suppose that exposure is randomized but that one of assumptions (A2.2) or (A2.4) do not hold; that is, either there is an unmeasured mediator–outcome confounder or there is a mediator–outcome confounder affected by the exposure.

Define the bias factor for natural indirect effect, B_c^{NIE}, as the difference between Q_{NIE} and the true natural indirect effect; and define the bias factor for natural direct effect, B_c^{NDE}, as the difference between Q_{NDE} and the true natural direct effect; that is,

$$B_c^{NIE} = Q_{NIE} - \mathbb{E}[Y_{1M_1} - Y_{1M_0}|c]$$
$$B_c^{NDE} = Q_{NDE} - \mathbb{E}[Y_{1M_0} - Y_{0M_0}|c]$$

Proposition 5.10. (VanderWeele and Chiba, 2014):
Suppose that exposure A is randomized. Let $\gamma_{mc} = \mathbb{E}[Y_{1m}|A=1,m,c] - \mathbb{E}[Y_{1m}|A=0,m,c]$ and let $\Gamma_c = \sum_m \gamma_{mc}P(m|A=0,c)$, then

$$B_c^{NIE} = -\Gamma_c$$
$$B_c^{NDE} = \Gamma_c$$

Proof:
Under randomization of A, we have

$$\begin{aligned}\mathbb{E}[Y_{aM_a}|c] &= \mathbb{E}[Y_a|c] \\ &= \mathbb{E}[Y_a|A=a,c] \\ &= \mathbb{E}[Y|A=a,c] \\ &= \sum_m \mathbb{E}[Y|A=a,m,c]P(m|A=a,c)\end{aligned}$$

where the first equality follows by composition, the second by randomization, the third by consistency, and the fourth by iterated expectations. We also have

$$\begin{aligned}\mathbb{E}[Y_{1M_0}|c] &= \sum_m \mathbb{E}[Y_{1m}|M_0=m,c]P(M_0=m|c) \\ &= \sum_m \mathbb{E}[Y_{1m}|A=0,M_0=m,c]P(M_0=m|A=0,c) \\ &= \sum_m \mathbb{E}[Y_{1m}|A=0,M=m,c]P(M=m|A=0,c) \\ &= \sum_m \{\mathbb{E}[Y_{1m}|A=1,M=m,c] - \gamma_{mc}\}P(M=m|A=0,c)\end{aligned}$$

$$
\begin{aligned}
&= \sum\nolimits_m \mathbb{E}[Y|A=1,M=m,c]P(M=m|A=0,c) \\
&\quad - \sum\nolimits_m \gamma_{mc} P(M=m|A=0,c) \\
&= \sum\nolimits_m \mathbb{E}[Y|A=1,M=m,c]P(M=m|A=0,c) - \Gamma_c
\end{aligned}
$$

where the first equality follows by iterated expectations, the second by randomization, the third by consistency, the fourth by definition of γ_{mc}, and the fifth by consistency. From this it follows that

$$
\begin{aligned}
B_c^{NIE} &= \sum\nolimits_m \mathbb{E}[Y|A=1,m,c]\{P(m|A=1,c) - P(m|A=0,c)\} \\
&\quad - \mathbb{E}[Y_{1M_1} - Y_{1M_0}|c] \\
&= \mathbb{E}[Y_{1M_1}|c] - \{\mathbb{E}[Y_{1M_0}|c] + \Gamma_c\} - \mathbb{E}[Y_{1M_1} - Y_{1M_0}|c] \\
&= -\Gamma_c
\end{aligned}
$$

and likewise

$$
\begin{aligned}
B_c^{NDE} &= \sum\nolimits_m \{\mathbb{E}[Y|A=1,m,c] \\
&\quad - \mathbb{E}[Y|A=0,m,c]\}P(m|A=0,c) - \mathbb{E}[Y_{1M_0} - Y_{0M_0}|c] \\
&= \mathbb{E}[Y_{1M_0}|c] + \Gamma_c - \mathbb{E}[Y_{0M_0}|c] - \mathbb{E}[Y_{1M_0} - Y_{0M_0}|c] \\
&= \Gamma_c
\end{aligned}
$$

This completes the proof. ■

Suppose again that exposure is randomized but that no further assumptions are made about confounding. Under assumptions (A2.1)–(A2.4) above, the natural indirect and direct effects on the risk ratio scale would be identified by (VanderWeele and Vansteelandt, 2010)

$$
Q_{NIE} = \frac{\sum_m \mathbb{E}[Y|A=1,m,c]P(m|A=1,c)}{\sum_m \mathbb{E}[Y|A=1,m,c]P(m|A=0,c)}
$$

$$
Q_{NDE} = \frac{\sum_m \mathbb{E}[Y|A=1,m,c]P(m|A=0,c)}{\sum_m \mathbb{E}[Y|A=0,m,c]P(m|A=0,c)}
$$

However, these expressions will be biased for the true natural indirect and direct effects if there is an unmeasured mediator–outcome confounder or a mediator–outcome confounder affected by the exposure. Define the following bias factors:

$$
B_c^i = \frac{1}{Q_{NIE}} - \frac{1}{\frac{P(Y_{1M_1}=1|c)}{P(Y_{1M_0}=1|c)}}
$$

$$
B_c^d = Q_{NDE} - \frac{P(Y_{1M_0}=1|c)}{P(Y_{0M_0}=1|c)}
$$

We then have the following result.

Proposition 5.11 (VanderWeele and Chiba, 2014):
Suppose that exposure A is randomized. Let $\gamma_{mc} = \mathbb{E}[Y_{1m}|A = 1, m, c] - \mathbb{E}[Y_{1m}|A = 0, m, c]$ and let $\Gamma_c = \sum_m \gamma_{mc} P(m|A = 0, c)$, then

$$B_c^i = \frac{\Gamma_c}{\mathbb{E}[Y|A = 1, c]}$$
$$B_c^d = \frac{\Gamma_c}{\mathbb{E}[Y|A = 0, c]}$$

and thus

$$\frac{P(Y_{1M_1} = 1|c)}{P(Y_{1M_0} = 1|c)} = \frac{Q_{NIE}}{1 - Q_{NIE} \times B_c^i}$$
$$\frac{P(Y_{1M_0} = 1|c)}{P(Y_{0M_0} = 1|c)} = Q_{NDE} - B_c^d$$

Proof:
We have $P(Y_{0M_0} = 1|c) = \sum_m \mathbb{E}[Y|A = 0, m, c] P(m|A = 0, c) = \mathbb{E}[Y|A = 0, c]$ and $P(Y_{1M_0} = 1|c) = \sum_m \mathbb{E}[Y|A = 1, M = m, c] P(M = m|A = 0, c) - \Gamma_c$ and thus we have

$$\begin{aligned} B_c^d &= Q_{NDE} - \frac{P(Y_{1M_0} = 1|c)}{P(Y_{0M_0} = 1|c)} \\ &= \frac{\sum_m \mathbb{E}[Y|A = 1, m, c] P(m|A = 0, c)}{\sum_m \mathbb{E}[Y|A = 0, m, c] P(m|A = 0, c)} - \frac{P(Y_{1M_0} = 1|c)}{P(Y_{0M_0} = 1|c)} \\ &= \frac{\sum_m \mathbb{E}[Y|A = 1, m, c] P(m|A = 0, c)}{\mathbb{E}[Y|A = 0, c]} \\ &\quad - \frac{\sum_m \mathbb{E}[Y|A = 1, M = m, c] P(M = m|A = 0, c) - \Gamma_c}{\mathbb{E}[Y|A = 0, c]} \\ &= \frac{\Gamma_c}{\mathbb{E}[Y|A = 0, c]} \end{aligned}$$

Also, $P(Y_{1M_1} = 1|c) = \sum_m \mathbb{E}[Y|A = 1, m, c] P(m|A = 1, c) = \mathbb{E}[Y|A = 1, c]$ and $P(Y_{1M_0} = 1|c) = \sum_m \mathbb{E}[Y|A = 1, M = m, c] P(M = m|A = 0, c) - \Gamma_c$ and thus we have

$$\begin{aligned} B_c^i &= \frac{1}{Q_{NIE}} - \frac{1}{\frac{P(Y_{1M_1} = 1|c)}{P(Y_{1M_0} = 1|c)}} \\ &= \frac{\sum_m \mathbb{E}[Y|A = 1, m, c] P(m|A = 0, c)}{\sum_m \mathbb{E}[Y|A = 1, m, c] P(m|A = 1, c)} - \frac{P(Y_{1M_0} = 1|c)}{P(Y_{1M_1} = 1|c)} \\ &= \frac{\sum_m \mathbb{E}[Y|A = 1, m, c] P(m|A = 0, c)}{\mathbb{E}[Y|A = 1, c]} \\ &\quad - \frac{\sum_m \mathbb{E}[Y|A = 1, M = m, c] P(M = m|A = 0, c) - \Gamma_c}{\mathbb{E}[Y|A = 1, c]} \\ &= \frac{\Gamma_c}{\mathbb{E}[Y|A = 1, c]} \end{aligned}$$

Since

$$B_c^i = \frac{1}{Q_{NIE}} - \frac{1}{\frac{P(Y_{1M_1}=1|c)}{P(Y_{1M_0}=1|c)}},$$

solving for $\frac{P(Y_{1M_1}=1|c)}{P(Y_{1M_0}=1|c)}$ gives

$$\frac{P(Y_{1M_1}=1|c)}{P(Y_{1M_0}=1|c)} = \frac{Q_{NIE}}{1 - Q_{NIE} \times B_c^i}$$

This completes the proof. ■

A.6. MEDIATION ANALYSIS WITH TIME-VARYING EXPOSURES AND MEDIATORS

A.6.1. Notation and Definitions for Time-Varying Exposures and Mediators

Suppose now that the exposure, mediator, and possibly confounding variables vary over time. Let $(A(1),\ldots,A(T))$, $(M(1),\ldots,M(T))$, and $(L(1),\ldots,L(T))$ denote values of the exposures, mediator, and time-varying confounders at periods $1,\ldots,T$, with initial baseline covariates C, and subsequent temporal ordering $A(t)$, $M(t)$, $L(t)$. We will revisit this question of temporal ordering again below. The relationships among the variables are given in Figure 6.1. For any variable W, let $\overline{W}(t) = (W(1),\ldots,W(t))$ and let $\overline{W} = \overline{W}(T) = (W(1),\ldots,W(T))$. Let $\underline{W}(t) = (W(t),\ldots,W(T))$. By convention, we let $W(t)$ denote the empty set for $t \leq 0$. Let $Y_{\overline{a}\overline{m}}$ be the counterfactual outcome if $\overline{A}$ were set to $\overline{a}$ and if $\overline{M}$ were set to $\overline{m}$. Let $M_{\overline{a}}(t)$ be the counterfactual value of $M(t)$ if $\overline{A}$ were set to $\overline{a}$. We assume consistency that when $\overline{A} = \overline{a}$ we have $M_{\overline{a}}(t) = M(t)$ and $Y_{\overline{a}}(t) = Y(t)$ and when $\overline{A} = \overline{a}$ and $\overline{M} = \overline{m}$ we have $Y_{\overline{a}\overline{m}} = Y$. We assume composition where $Y_{\overline{a}} = Y_{\overline{a}M_{\overline{a}}}$. Let $\overline{a}$ and $\overline{a}^*$ be two distinct exposure histories. Controlled direct effects are defined as $Y_{\overline{a}\overline{m}} - Y_{\overline{a}^*\overline{m}}$. The natural direct effect can be defined as $Y_{\overline{a}M_{\overline{a}^*}} - Y_{\overline{a}^*M_{\overline{a}^*}}$ and the natural indirect effect as $Y_{\overline{a}M_{\overline{a}}} - Y_{\overline{a}M_{\overline{a}^*}}$. We have the decomposition of a total effect into natural direct and indirect effects $Y_{\overline{a}} - Y_{\overline{a}^*} = (Y_{\overline{a}M_{\overline{a}}} - Y_{\overline{a}M_{\overline{a}^*}}) + (Y_{\overline{a}M_{\overline{a}^*}} - Y_{\overline{a}^*M_{\overline{a}^*}})$.

Note that if the entire vector $A = (A(1),\ldots,A(T))$ is taken as the exposure and $M = (M(1),\ldots,M(T))$ is taken as the mediator, then the variable $L(1)$ is itself affected by the exposure (namely, by $A(1)$) and, in turn, confounds the mediator–outcome relationship between $M(2)$ and Y. From this it follows that natural direct and indirect effects are not identified in this setting (Avin et al., 2005). However, identification of randomized interventional analogues may once again be possible.

Let $\overline{G}_{\overline{a}|c}(t)$ denote a random draw from the distribution of the mediator $\overline{M}(t)$ that would have been observed in the population with baseline covariates $C = c$ if exposure status $\overline{A}$ had been fixed to $\overline{a}$. As in Chapter 5, we can define, now in the longitudinal setting, randomized interventional analogues of the natural direct

effect as $\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|c}}|c)$ and randomized interventional analogues of the natural indirect effect as $\mathbb{E}(Y_{\overline{a}G_{\overline{a}|c}}|c) - \mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c)$. We once again have a decomposition, even with time-varying exposures and mediators: $\mathbb{E}(Y_{\overline{a}\overline{G}_{\overline{a}|c}(t)}|c) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|c}}|c) = \{\mathbb{E}(Y_{\overline{a}G_{\overline{a}|c}}|c) - \mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c)\} + \{\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|c}}|c)\}$.

A.6.2. Controlled Direct Effects with Time-Varying Exposures and Mediators

The following proposition justifies the weighting approach to controlled direct effects in the presence of exposure-induced confounding described in the text. It is a simple example of the marginal structural model of Robins (1999a) with the alternating treatment periods being taken as the exposure and the mediator over time (cf. van der Laan and Petersen, 2008; VanderWeele, 2009a).

Proposition 6.1 (Robins, 1986, 1999):
Under the assumptions that (A6.1) $Y_{\overline{am}} \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),\overline{L}(t-1),C$ and (A6.2) $Y_{\overline{am}} \perp\!\!\!\perp M(t)|\overline{A}(t),\overline{M}(t-1),\overline{L}(t-1),C$ we have

$$\begin{aligned}\mathbb{E}[Y_{\overline{am}}] &= \sum_{c,\bar{l}} \mathbb{E}[Y|\overline{a},\overline{m},\bar{l},c] \prod_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}P(c) \\ &= E\left[\frac{Y}{\prod_{t=1}^{T} P\{A(t)|\overline{a}(t-1),\overline{m}(t-1),\bar{l}(t-1),c) \prod_{t=1}^{T} P\{M(t)|\overline{a}(t),\overline{m}(t-1),\bar{l}(t-1),c)}\right]\end{aligned}$$

Proof:
Under (A6.1) and (A6.2),

$$\begin{aligned}\mathbb{E}[Y_{\overline{am}}] &= \sum\nolimits_{c} \mathbb{E}[Y_{\overline{am}}|c]P(c) \\ &= \sum\nolimits_{c} \mathbb{E}[Y_{\overline{am}}|a(1),m(1),c]P(c) \\ &= \sum\nolimits_{c,l(1)} \mathbb{E}[Y_{\overline{am}}|a(1),m(1),l(1),c]P\{l(1)|a(1),m(1),c\}P(c) \\ &= \sum\nolimits_{c,l(1)} \mathbb{E}[Y_{\overline{am}}|\overline{a}(2),\overline{m}(2),l(1),c]P\{l(1)|a(1),m(1),c\}P(c)\end{aligned}$$

and iteratively continuing this argument we obtain

$$\begin{aligned}\mathbb{E}[Y_{\overline{am}}] &= \sum\nolimits_{c,\bar{l}} \mathbb{E}[Y_{\overline{am}}|\overline{a},\overline{m},\bar{l},c] \prod\nolimits_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}P(c) \\ &= \sum\nolimits_{c,\bar{l}} \mathbb{E}[Y|\overline{a},\overline{m},\bar{l},c] \prod\nolimits_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}P(c)\end{aligned}$$

Moreover, we have

$$\begin{aligned}&\sum\nolimits_{c,\bar{l}} \mathbb{E}[Y|\overline{a},\overline{m},\bar{l},c] \prod\nolimits_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}P(c) \\ &= \sum\nolimits_{c,\bar{l},y} yP(y|\overline{a},\overline{m},\bar{l},c) \prod\nolimits_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}P(c)\end{aligned}$$

$$= \sum_{c,\bar{l},y} yP(y|\bar{a},\bar{m},\bar{l},c) \prod_{t=1}^{T-1} P\{\bar{l}(t)|\bar{a}(t),\bar{m}(t),\bar{l}(t-1),c)\}P(c)$$
$$\times \frac{\prod_{t=1}^{T} P\{A(t)|\bar{a}(t-1),\bar{m}(t-1),\bar{l}(t-1),c) \prod_{t=1}^{T} P\{M(t)|\bar{a}(t),\bar{m}(t-1),\bar{l}(t-1),c)}{\prod_{t=1}^{T} P\{A(t)|\bar{a}(t-1),\bar{m}(t-1),\bar{l}(t-1),c) \prod_{t=1}^{T} P\{M(t)|\bar{a}(t),\bar{m}(t-1),\bar{l}(t-1),c)}$$
$$= \sum_{c,\bar{l},y} yP(y,\bar{a},\bar{m},\bar{l},c)$$
$$\times \frac{1}{\prod_{t=1}^{T} P\{A(t)|\bar{a}(t-1),\bar{m}(t-1),\bar{l}(t-1),c) \prod_{t=1}^{T} P\{M(t)|\bar{a}(t),\bar{m}(t-1),\bar{l}(t-1),c)}$$
$$= \mathbb{E}\left[\frac{Y}{\prod_{t=1}^{T} P\{A(t)|\bar{a}(t-1),\bar{m}(t-1),\bar{l}(t-1),c) \prod_{t=1}^{T} P\{M(t)|\bar{a}(t),\bar{m}(t-1),\bar{l}(t-1),c)}\right]$$

This completes the proof. ■

See Robins (1999a) and Robins et al. (2000) for more on statistical inference with marginal structural models.

A.6.3. Natural Direct and Indirect Effect and their Randomized Interventional Analogues with Time-Varying Exposures and Mediators

Suppose now that at each time, conditional on the past, the exposure–outcome, mediator–outcome, and exposure–mediator relationships are unconfounded. Formally, analogous to (A2.1)–(A2.3): for all t, (A6.1) $Y_{\overline{am}} \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),\overline{L}(t-1),C$, (A6.2) $Y_{\overline{am}} \perp\!\!\!\perp M(t)|\overline{A}(t),\overline{M}(t-1),\overline{L}(t-1),C$, and (A6.3) $M_{\overline{a}}(t) \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),\overline{L}(t-1),C$. It can be shown that although natural direct and indirect effects are not in general identified in this setting, the randomized interventional analogues, $\{\mathbb{E}(Y_{\overline{a}G_{\overline{a}|c}}|c) - \mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c)\}$ and $\{\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|c}}|c)\}$, are identified.

Proposition 6.2 (VanderWeele and Tchetgen Tchetgen, 2014):
Under assumptions (A6.1)–(A6.3), we have

$$\mathbb{E}[Y_{\overline{a}G_{\overline{a}^*|c}}|c] = Q(\overline{a},\overline{a}^*) := \sum_{\overline{m}} \sum_{\bar{l}(T-1)} \mathbb{E}[Y|\overline{a},\overline{m},\bar{l},c] \prod_{t=1}^{T-1}$$
$$\times P\{\bar{l}(t)|\bar{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}$$
$$\times \sum_{\bar{l}^\dagger(T-1)} \prod_{t=1}^{T} P\{M(t)|\bar{a}^*(t),\overline{m}(t-1),\bar{l}^\dagger(t-1),c\}$$
$$\times P\{\bar{l}^\dagger(t-1)|\bar{a}^*(t-1),\overline{m}(t-1),\bar{l}^\dagger(t-2),c)\}$$

and thus the randomized interventional analogues of natural direct and indirect effects are given by

$$\mathbb{E}(Y_{\overline{a}G_{\overline{a}|c}}|c) - \mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c) = Q(\overline{a},\overline{a}) - Q(\overline{a},\overline{a}^*)$$
$$\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|c}}|c) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|c}}|c) = Q(\overline{a},\overline{a}^*) - Q(\overline{a}^*,\overline{a}^*)$$

Proof:
We have that

$$
\begin{aligned}
&\mathbb{E}[Y_{\overline{a}G_{\overline{a}^*|c}}|c] \\
&= \sum\nolimits_{m(1)} \mathbb{E}[Y_{\overline{a}m(1)\underline{G}_{\overline{a}^*|c}(2)}|G_{\overline{a}^*|c}(1)=m(1),c]P\{G_{\overline{a}^*|c}(1)=m(1)|c\} \\
&= \sum\nolimits_{m(1)} \mathbb{E}[Y_{\overline{a}m(1)\underline{G}_{\overline{a}^*|c}(2)}|G_{\overline{a}^*|c}(1)=m(1),a(1),m(1),c]P\{M_{\overline{a}^*}(1)=m(1)|a^*(1),c\} \\
&= \sum\nolimits_{m(1)} \mathbb{E}[Y_{\overline{a}m(1)\underline{G}_{\overline{a}^*|c}(2)}|G_{\overline{a}^*|c}(1)=m(1),a(1),m(1),c]P\{m(1)|a^*(1),c\} \\
&= \sum\nolimits_{\overline{m}(2)} \mathbb{E}[Y_{\overline{am}(2)\underline{G}_{\overline{a}^*|c}(3)}|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),a(1),m(1),c] \\
&\quad \times P\{G_{\overline{a}^*|c}(2)=m(2)|G_{\overline{a}^*|c}(1)=m(1),a(1),m(1),c\}P\{m(1)|a(1),c\} \\
&= \sum\nolimits_{\overline{m}(2)} \mathbb{E}[Y_{\overline{am}(2)\underline{G}_{\overline{a}^*|c}(3)}|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),a(1),m(1),c] \\
&\quad \times P\{G_{\overline{a}^*|c}(2)=m(2)|G_{\overline{a}^*|c}(1)=m(1),a^*(1),m(1),c\}P\{m(1)|a(1),c\} \\
&= \sum\nolimits_{\overline{m}(2)} \sum\nolimits_{\overline{l}(1)} \mathbb{E}[Y_{\overline{am}(2)\underline{G}_{\overline{a}^*|c}(3)}|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),a(1),m(1),l(1),c] \\
&\quad \times P\{l(1)|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),a(1),m(1),c\} \\
&\quad \times P\{M_{\overline{a}^*}(2)=m(2)|\overline{a}^*(1),m(1),c\}P\{m(1)|a(1),c\} \\
&= \sum\nolimits_{\overline{m}(2)} \sum\nolimits_{\overline{l}(1)} \mathbb{E}[Y_{\overline{am}(2)\underline{G}_{\overline{a}^*|c}(3)}|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),\overline{a}(2),\overline{m}(2),l(1),c]P\{l(1)|a(1),m(1),c\} \\
&\qquad \times \sum\nolimits_{\overline{l}^\dagger(1)} P\{M_{\overline{a}^*}(2)=m(2)|a^*(1),m(1),l^{\dagger}(1),c\}P(l^{\dagger}(1)|a^*(1),m(1),c\}P\{m(1)|a(1),c\} \\
&= \sum\nolimits_{\overline{m}(2)} \sum\nolimits_{\overline{l}(1)} \mathbb{E}[Y_{\overline{am}(2)\underline{G}_{\overline{a}^*|c}(3)}|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),\overline{a}(2),\overline{m}(2),l(1),c]P\{l(1)|a(1),m(1),c\} \\
&\quad \times \sum\nolimits_{\overline{l}^\dagger(1)} P\{M_{\overline{a}^*}(2)=m(2)|\overline{a}^*(2),m(1),l^{\dagger}(1),c\}P(l^{\dagger}(1)|a^*(1),m(1),c\}P\{m(1)|a(1),c\} \\
&= \sum\nolimits_{\overline{m}(2)} \sum\nolimits_{\overline{l}(1)} \mathbb{E}[Y_{\overline{am}(2)\underline{G}_{\overline{a}^*|c}(3)}|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),\overline{a}(2),\overline{m}(2),l(1),c]P\{l(1)|a(1),m(1),c\} \\
&\quad \times \sum\nolimits_{\overline{l}^\dagger(1)} P\{m(2)|\overline{a}^*(2),m(1),l^{\dagger}(1),c\}P(l^{\dagger}(1)|a^*(1),m(1),c\}P\{m(1)|a(1),c\}
\end{aligned}
$$

Note that in the expectation in the second and subsequent equalities we cannot remove $G_{\overline{a}^*|c}(1)$ from the conditioning set because it will be associated with $\underline{G}_{\overline{a}^*|c}(2)$. In the fifth inequality we can make the substitution $P\{G_{\overline{a}^*|c}(2)=m(2)|G_{\overline{a}^*|c}(1)=m(1),a(1),m(1),c\}=P\{G_{\overline{a}^*|c}(2)=m(2)|G_{\overline{a}^*|c}(1)=m(1),a^*(1),m(1),c\}$ because the first expression is equal to $\frac{P\{G_{\overline{a}^*|c}(2)=m(2),G_{\overline{a}^*|c}(1)=m(1)|a(1),m(1),c\}}{P\{G_{\overline{a}^*|c}(1)=m(1)|a(1),m(1),c\}}$ and the second is equal to $\frac{P\{G_{\overline{a}^*|c}(2)=m(2),G_{\overline{a}^*|c}(1)=m(1)|a^*(1),m(1),c\}}{P\{G_{\overline{a}^*|c}(1)=m(1)|a^*(1),m(1),c\}}$, and these latter expressions are equal to each other since $(G_{\overline{a}^*|c}(2),G_{\overline{a}^*|c}(1))$, being random draws, will be independent of any actual observed variables. Likewise in the seventh equality, we can remove $\overline{G}_{\overline{a}^*|c}(2)$ from the conditioning set in $P\{l(1)|\overline{G}_{\overline{a}^*|c}(2)=\overline{m}(2),a(1),m(1),c\}$ because $\overline{G}_{\overline{a}^*|c}(2)$ will be independent of all actual observed variables. If we carry on with this argument iteratively, we obtain

$$
= \sum\nolimits_{\overline{m}} \sum\nolimits_{\overline{l}(T-1)} \mathbb{E}[Y_{\overline{am}}|\overline{G}_{\overline{a}^*|c}=\overline{m},\overline{a},\overline{m},\overline{l},c] \prod\nolimits_{t=1}^{T-1} P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t),\overline{l}(t-1),c)\}
$$

$$
\begin{aligned}
&\times \sum_{\bar{l}^{\dagger}(T-1)} \prod_{t=1}^{T} P\{M(t)|\overline{a}(t),\overline{m}(t-1),\bar{l}^{\dagger}(t-1),c\} P\{\bar{l}^{\dagger}(t-1)|\overline{a}(t-1),\\
&\overline{m}(t-1),\bar{l}^{\dagger}(t-2),c)\}\\
&= \sum_{\overline{m}} \sum_{\bar{l}(T-1)} \mathbb{E}[Y_{\overline{am}}|\overline{a},\overline{m},\bar{l},c] \prod_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}\\
&\times \sum_{\bar{l}^{\dagger}(T-1)} \prod_{t=1}^{T} P\{M(t)|\overline{a}(t),\overline{m}(t-1),\bar{l}^{\dagger}(t-1),c\} P\{\bar{l}^{\dagger}(t-1)|\overline{a}(t-1),\\
&\overline{m}(t-1),\bar{l}^{\dagger}(t-2),c)\}\\
&= \sum_{\overline{m}} \sum_{\bar{l}(T-1)} \mathbb{E}[Y|\overline{a},\overline{m},\bar{l},c] \prod_{t=1}^{T-1} P\{\bar{l}(t)|\overline{a}(t),\overline{m}(t),\bar{l}(t-1),c)\}\\
&\times \sum_{\bar{l}^{\dagger}(T-1)} \prod_{t=1}^{T} P\{M(t)|\overline{a}^{*}(t),\overline{m}(t-1),\bar{l}^{\dagger}(t-1),c\} P\{\bar{l}^{\dagger}(t-1)|\overline{a}^{*}(t-1),\\
&\overline{m}(t-1),\bar{l}^{\dagger}(t-2),c)\}
\end{aligned}
$$

This completes the proof. ■

We refer to this final expression $Q(\overline{a},\overline{a}^{*})$ as the mediational g-formula. If $\overline{L}$ is empty so that there is no exposure-induced mediator–outcome confounder, then this reduces to

$$Q(\overline{a},\overline{a}^{*}) = \sum_{\overline{m}} \mathbb{E}[Y|\overline{a},\overline{m},c] \prod_{t=1}^{T} P\{M(t)|\overline{a}^{*}(t),\overline{m}(t-1),c\}$$

and this expression can in fact be used to identify natural direct and indirect effects (not just the randomized interventional analogues). To see this, consider the causal diagram in Figure 6.3 in which $A(t)$ and $M(t)$ are not time-varying and suppose this were a nonparametric structural equation model (Shpitser and Pearl, 2008; Pearl, 2009). The following assumptions would then hold: (A6.1*) $Y_{\overline{am}} \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),C$, (A6.2*) $Y_{\overline{am}} \perp\!\!\!\perp M(t)|\overline{A}(t),\overline{M}(t-1),C$, (A6.3*) $M_{\overline{a}}(t) \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),C$, and (A6.4*) $Y_{\overline{am}} \perp\!\!\!\perp \overline{M}_{\overline{a}^{*}}(t)|C$. It can be shown that assumption (A6.4*) follows from the nonparametric structural equation model using a twin network diagram (Shpitser and Pearl, 2008). Note also that assumptions (A6.1*)–(A6.4*) would also hold if there were a variable U_A in Figure 6.3 with edges into $A(t)$ for any or all t [but no edges into any $M(t)$] and/or if there were a variable U_M in Figure 6.3 with edges into $M(t)$ for any or all t [but no edges into any $A(t)$].

Under assumptions (i*)–(iv*), average natural direct and indirect effects conditional on $C = c$ are identified since

$$
\begin{aligned}
\mathbb{E}[Y_{\overline{a}M_{\overline{a}^{*}}}|c] &= \sum_{\overline{m}} \mathbb{E}[Y_{\overline{am}}|M_{\overline{a}^{*}} = \overline{m},c] P(M_{\overline{a}^{*}} = \overline{m}|c)\\
&= \sum_{\overline{m}} \mathbb{E}[Y_{\overline{am}}|c] P(M_{\overline{a}^{*}} = \overline{m}|c)\\
&= \sum_{\overline{m}} \mathbb{E}[Y|\overline{a},\overline{m},c] \prod_{t=1}^{T} P\{M(t)|\overline{a}^{*}(t),\overline{m}(t-1),c\}
\end{aligned}
$$

where the final equality follows by application of Robins' g-formula (Robins, 1986). The average natural direct effect conditional on $C = c$ is thus given by

$$\mathbb{E}[Y_{\overline{a}M_{\overline{a}^*}}|c] - \mathbb{E}[Y_{\overline{a}^*M_{\overline{a}^*}}|c] = \sum\nolimits_{\overline{m}} \{\mathbb{E}[Y|\overline{a},\overline{m},c] - \mathbb{E}[Y|\overline{a}^*,\overline{m},c]\} \times \prod\nolimits_{t=1}^{T} P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),c\}$$

The average natural indirect effect conditional on $C = c$ is given by

$$\mathbb{E}[Y_{\overline{a}M_{\overline{a}}}|c] - \mathbb{E}[Y_{\overline{a}M_{\overline{a}^*}}|c] = \sum\nolimits_{\overline{m}} \mathbb{E}[Y|\overline{a},\overline{m},c] \times \prod\nolimits_{t=1}^{T} [P\{M(t)|\overline{a}(t),\overline{m}(t-1),c\} - P\{M(t)|\overline{a}^*(t), \overline{m}(t-1),c\}]$$

This final expression is a generalization of Pearl's mediation formula (Pearl, 2012) for time-varying exposures and mediators.

In other words if $\overline{L}$ is empty, then the empirical expressions that suffice to identify the randomized interventional analogues of natural direct and indirect effects under assumptions (A6.1)–(A6.4) in fact also in this setting identify the natural direct and indirect effects as well by a time-varying analogue of Pearl's "mediation formula" (Pearl, 2012). However, even when $\overline{L}$ is not empty so we cannot identify the natural direct and indirect effects themselves, we still can, under assumptions (A6.1)–(A6.3) identify the randomized interventional analogues of the natural direct and indirect effects.

Note also that if M were empty, then the expression for $Q(\overline{a},\overline{a}^*)$ simply reduces to

$$= \sum\nolimits_{\overline{m}} \sum\nolimits_{\overline{l}(T-1)} \mathbb{E}[Y|\overline{a},\overline{l},c] \prod\nolimits_{t=1}^{T-1} P\{\overline{l}(t)|\overline{a}(t),\overline{l}(t-1),c)\}$$

because, with M empty, $\sum_{\overline{l}^\dagger(T-1)} \prod_{t=1}^{T} P\{\overline{l}^\dagger(t-1)|\overline{a}^*(t-1),\overline{l}^\dagger(t-2),c)\} = 1$. Thus with M empty, the formula for $Q(\overline{a},\overline{a}^*)$ simply reduces to the regular g-formula of Robins (1986). We see then that, on the one hand, if there is no-time-varying confounding the "mediational g-formula" $Q(\overline{a},\overline{a}^*)$ reduces to the time-varying analogue of Pearl's mediation formula. And if, on the other hand, there is no mediation, then the "mediational g-formula" reduces to the regular g-formula.

Suppose instead that after the initial baseline covariates C, the subsequent temporal ordering of the variables were $A(t)$, $L(t)$, $M(t)$, as in Figure 6.2, and that analogous to (A6.1†)–(A6.3†) we have that for all t, (A6.1‡) $Y_{\overline{am}} \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),\overline{L}(t-1),C$, (A6.2‡) $Y_{\overline{am}} \perp\!\!\!\perp M(t)|\overline{A}(t),\overline{M}(t-1),\overline{L}(t),C$, and (A6.3‡) $M_{\overline{a}}(t) \perp\!\!\!\perp A(t)|\overline{A}(t-1),\overline{M}(t-1),\overline{L}(t-1),C$. Under assumptions (A6.1‡)–(A6.3‡) we would have by similar arguments (cf. VanderWeele and Tchetgen Tchetgen, 2014) that

$$\mathbb{E}[Y_{\overline{a}G_{\overline{a}^*|c}}|c] = \sum\nolimits_{\overline{m}} \sum\nolimits_{\overline{l}(T-1)} \mathbb{E}[Y|\overline{a},\overline{m},\overline{l},c] \prod\nolimits_{t=1}^{T-1} P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t-1),\overline{l}(t-1),c)\} \times \sum\nolimits_{\overline{l}^\dagger(T-1)} \prod\nolimits_{t=1}^{T} P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),\overline{l}^\dagger(t),c\} P\{\overline{l}^\dagger(t)|\overline{a}^*(t),\overline{m}(t-1),\overline{l}^\dagger(t-1),c)\}$$

As another variation instead of considering randomized interventions that fix the mediator $\overline{M}$ for each individual to a value randomly drawn from the distribution in the subpopulation with baseline covariates $C = c$, if $\overline{A}$ had been fixed to $\overline{a}^*$, we could instead consider randomizing the mediator $\overline{M}$ for each individual to the value randomly drawn the distribution in the entire population if $\overline{A}$ had been fixed to $\overline{a}^*$. We then let Let $\overline{G}_{\overline{a}}(t)$ denote a random draw from the distribution of the mediator $\overline{M}(t)$ that would have been observed in the population if exposure $\overline{A}$ had been fixed to $\overline{a}$ and we have the decomposition $\mathbb{E}(Y_{\overline{a}\overline{G}_{\overline{a}}(t)}) - \mathbb{E}(Y_{\overline{a}^* G_{\overline{a}^*}}) = \{\mathbb{E}(Y_{\overline{a}G_{\overline{a}}}) - \mathbb{E}(Y_{\overline{a}G_{\overline{a}^*}})\} + \{\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*}}|c) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*}})\}$. Under assumptions (A6.1†)–(A6.3†), we have

$$\begin{aligned}\mathbb{E}[Y_{\overline{a}G_{\overline{a}^*}}] = &\sum\nolimits_{\overline{m}}\sum\nolimits_{\overline{l}(T-1)} \mathbb{E}[Y|\overline{a},\overline{m},\overline{l},c]\prod\nolimits_{t=1}^{T-1} P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t),\overline{l}(t-1),c)\}P(c)\\ &\times\sum\nolimits_{\overline{l}^\dagger(T-1)}\prod\nolimits_{t=1}^{T} P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),\\ &\overline{l}^\dagger(t-1),c\}P\{\overline{l}^\dagger(t-1)|\overline{a}^*(t-1),\overline{m}(t-1),\overline{l}^\dagger(t-2),c)\}P(c).\end{aligned}$$

and under assumptions (A6.1‡)–(A6.3‡), we would then have

$$\begin{aligned}\mathbb{E}[Y_{\overline{a}G_{\overline{a}^*}}] = &\sum\nolimits_{\overline{m}}\sum\nolimits_{\overline{l}(T-1)} \mathbb{E}[Y|\overline{a},\overline{m},\overline{l},c]\prod\nolimits_{t=1}^{T-1} P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t-1),\overline{l}(t-1),c)\}P(c)\\ &\times\sum\nolimits_{\overline{l}^\dagger(T-1)}\prod\nolimits_{t=1}^{T} P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),\overline{l}^\dagger(t),c\}P\{\overline{l}^\dagger(t)|\overline{a}^*(t),\overline{m}(t-1),\\ &\overline{l}^\dagger(t-1),c)\}P(c).\end{aligned}$$

A.6.4. Counterfactual Analysis of MacKinnon's Three-Wave Mediation Model

MacKinnon (2008) considered a three-wave mediation model with linear structural equations as depicted in Figure 6.4. We relabel indices somewhat to correspond to the notation of this chapter and also add a set of baseline covariates C, but otherwise the model considered here is MacKinnon's model (MacKinnon, 2008, pp. 204–206, Autoregressive Model III). We let $A(0)$, $M(0)$, and $Y(0)$ denote baseline values of A, M, and Y that could be included in the baseline covariates C but are given here to make clearer the relation with MacKinnon (2008).

Proposition 6.3 (VanderWeele and Tchetgen Tchetgen, 2014):
Consider then the following regression models:

$$\begin{aligned}\mathbb{E}[M(1)|m(0),y(0),\overline{a}(1),c] &= \beta_{10} + \beta_{11}a(0) + \beta_{12}a(1) + \beta_{13}m(0) + \beta_{14}y(0) + \beta'_{15}c\\ \mathbb{E}[M(2)|\overline{m}(1),\overline{y}(1),\overline{a}(2),c] &= \beta_{20} + \beta_{21}a(1) + \beta_{22}a(2) + \beta_{23}m(1) + \beta_{24}y(1) + \beta'_{25}c\\ \mathbb{E}[Y(1)|\overline{m}(1),y(0),\overline{a}(1),c] &= \theta_{10} + \theta_{11}a(0) + \theta_{12}a(1) + \theta_{13}m(0)\\ &\quad + \theta_{14}m(1) + \theta_{15}y(0) + \theta'_{16}c\\ \mathbb{E}[Y(2)|\overline{m}(2),\overline{y}(1),\overline{a}(2),c] &= \theta_{20} + \theta_{21}a(1) + \theta_{22}a(2) + \theta_{23}m(1) + \theta_{24}m(2) + \theta_{25}y(1)\\ &\quad + \theta'_{26}c\end{aligned}$$

Under assumptions (A6.1†)–(A6.3†) with $V=(C,A(0),M(0),Y(0))$, and $L(1)=Y(1)$, with two intervention periods, $A(1)$ and $A(2)$, the randomized interventional analogues of the natural direct and indirect effects are given by

$$\begin{aligned}\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|\nu}}|\nu)-\mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|\nu}}|\nu)&=(\theta_{21}+\theta_{12}\theta_{25})[a(1)-a^*(1)]\\&\quad+\theta_{22}[a(2)-a^*(2)]\\\{\mathbb{E}(Y_{\overline{a}G_{\overline{a}|\nu}}|\nu)-\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|\nu}}|\nu)\}&=\{\theta_{23}\beta_{12}+\theta_{25}\theta_{14}\beta_{12}+\beta_{21}\theta_{24}+\beta_{24}\theta_{12}\theta_{24}\}\\&\quad\times[a(1)-a^*(1)]+\beta_{22}\theta_{24}[a(2)-a^*(2)]\end{aligned}$$

Proof:
By the mediational g-formula we have

$$\begin{aligned}\mathbb{E}[Y_{\overline{a}G_{\overline{a}^*|c}}|c]&=\int_{\overline{m}}\int_{\overline{l}(T-1)}\mathbb{E}[Y|\overline{a},\overline{m},\overline{l},c]\prod_{t=1}^{T-1}P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t),\overline{l}(t-1),c)\}\\&\quad\times\int_{\overline{l}^\dagger(T-1)}\prod_{t=1}^{T}P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),\\&\quad\overline{l}^\dagger(t-1),c\}P\{\overline{l}^\dagger(t-1)|\overline{a}^*(t-1),\overline{m}(t-1),\overline{l}^\dagger(t-2),c)\}\end{aligned}$$

We have that

$$\begin{aligned}&\int_{\overline{l}(T-1)}\mathbb{E}[Y|\overline{a},\overline{m},\overline{l},c]\prod_{t=1}^{T-1}P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t),\overline{l}(t-1),c)\}\\&=\int_{y(1)}\mathbb{E}[Y(2)|\overline{m}(2),\overline{y}(1),\overline{a}(2),c]P\{y(1)|\overline{m}(1),y(0),\overline{a}(1),c\}\\&=\theta_{20}+\theta_{21}a(1)+\theta_{22}a(2)+\theta_{23}m(1)+\theta_{24}m(2)+\theta_{25}\mathbb{E}[Y(1)|\overline{m}(1),y(0),\overline{a}(1),x]+\theta'_{26}c\\&=\theta_{20}+\theta_{21}a(1)+\theta_{22}a(2)+\theta_{23}m(1)+\theta_{24}m(2)\\&\quad+\theta_{25}\{\theta_{10}+\theta_{11}a(0)+\theta_{12}a(1)+\theta_{13}m(0)+\theta_{14}m(1)+\theta_{15}y(0)+\theta'_{16}c\}+\theta'_{26}c\end{aligned}$$

Thus,

$$\begin{aligned}&\int_{\overline{m}}\int_{\overline{l}(T-1)}\mathbb{E}[Y|\overline{a},\overline{m},\overline{l},c]\prod_{t=1}^{T-1}P\{\overline{l}(t)|\overline{a}(t),\overline{m}(t),\overline{l}(t-1),c)\}\\&\quad\times\int_{\overline{l}^\dagger(T-1)}\prod_{t=1}^{T}P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),\overline{l}^\dagger(t-1),c\}P\{\overline{l}^\dagger(t-1)|\overline{a}^*(t-1),\\&\quad\overline{m}(t-1),\overline{l}^\dagger(t-2),c)\}\\&=\theta_{20}+\theta_{21}a(1)+\theta_{22}a(2)+\theta_{25}\{\theta_{10}+\theta_{11}a(0)+\theta_{12}a(1)+\theta_{13}m(0)+\theta_{15}y(0)+\theta'_{16}c\}\\&\quad+\theta'_{26}c+\int_{\overline{m}}\{\theta_{23}m(1)+\theta_{24}m(2)+\theta_{25}\theta_{14}m(1)\}\\&\quad\times\int_{\overline{l}^\dagger(T-1)}\prod_{t=1}^{T}P\{M(t)|\overline{a}^*(t),\overline{m}(t-1),\overline{l}^\dagger(t-1),c\}P\{\overline{l}^\dagger(t-1)|\overline{a}^*(t-1),\\&\quad\overline{m}(t-1),\overline{l}^\dagger(t-2),c)\}\\&=\theta_{20}+\theta_{21}a(1)+\theta_{22}a(2)+\theta_{25}\{\theta_{10}+\theta_{11}a(0)+\theta_{12}a(1)+\theta_{13}m(0)\\&\quad+\theta_{15}y(0)+\theta'_{16}c\}+\theta'_{26}c+\{\theta_{23}+\theta_{25}\theta_{14}\}\mathbb{E}[M(1)|m(0),y(0),\overline{a}^*(1),c]\end{aligned}$$

$$
\begin{aligned}
&+ \theta_{24} \int_{y(1)} \mathbb{E}[M(2)|\overline{m}(1),\overline{y}(1),\overline{a}^*(2),c]P\{\overline{y}(1)|\overline{m}(1),y(0),\overline{a}^*(1),c\} \\
&= \theta_{20} + \theta_{21}a(1) + \theta_{22}a(2) + \theta_{25}\{\theta_{10} + \theta_{11}a(0) + \theta_{12}a(1) + \theta_{13}m(0) + \theta_{15}y(0) + \theta'_{16}c\} \\
&\quad + \theta'_{26}c + \{\theta_{23} + \theta_{25}\theta_{14}\}\{\beta_{10} + \beta_{11}a(0) + \beta_{12}a^*(1) + \beta_{13}m(0) + \beta_{14}y(0) + \beta'_{15}c\} \\
&\quad + \theta_{24} \int_{y(1)} \{\beta_{20} + \beta_{21}a^*(1) + \beta_{22}a^*(2) + \beta_{23}m(1) + \beta_{24}y(1) + \beta'_{25}c\}P\{\overline{y}(1)|\overline{m}(1), \\
&\quad y(0),\overline{a}^*(1),c\} \\
&= \theta_{20} + \theta_{21}a(1) + \theta_{22}a(2) + \theta_{25}\{\theta_{10} + \theta_{11}a(0) + \theta_{12}a(1) + \theta_{13}m(0) + \theta_{15}y(0) \\
&\quad + \theta'_{16}c\} + \theta'_{26}c + \{\theta_{23} + \theta_{25}\theta_{14}\}\{\beta_{10} + \beta_{11}a(0) + \beta_{12}a^*(1) + \beta_{13}m(0) + \beta_{14}y(0) \\
&\quad + \beta'_{15}c\} + \theta_{24}\{\beta_{20} + \beta_{21}a^*(1) + \beta_{22}a^*(2) + \beta_{23}m(1) + \beta_{24}\mathbb{E}[Y(1)|\overline{m}(1),y(0), \\
&\quad \overline{a}^*(1),c] + \beta'_{25}c\} \\
&= \theta_{20} + \theta_{21}a(1) + \theta_{22}a(2) + \theta_{25}\{\theta_{10} + \theta_{11}a(0) + \theta_{12}a(1) + \theta_{13}m(0) + \theta_{15}y(0) + \theta'_{16}c\} \\
&\quad + \theta'_{26}c + \{\theta_{23} + \theta_{25}\theta_{14}\}\{\beta_{10} + \beta_{11}a(0) + \beta_{12}a^*(1) + \beta_{13}m(0) + \beta_{14}y(0) + \beta'_{15}c\} \\
&\quad + \theta_{24}[\beta_{20} + \beta_{21}a^*(1) + \beta_{22}a^*(2) + \beta_{23}m(1) \\
&\quad + \beta_{24}\{\theta_{10} + \theta_{11}a(0) + \theta_{12}a^*(1) + \theta_{13}m(0) + \theta_{14}m(1) + \theta_{15}y(0) + \theta'_{16}c\} + \beta'_{25}c]
\end{aligned}
$$

Thus the randomized interventional analogue of the natural direct effect is given by

$$\mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|v}}|v) - \mathbb{E}(Y_{\overline{a}^*G_{\overline{a}^*|v}}|v) = (\theta_{21} + \theta_{12}\theta_{25})[a(1) - a^*(1)] + \theta_{22}[a(2) - a^*(2)]$$

and the randomized interventional analogue of the natural direct effect is given by

$$
\begin{aligned}
\{\mathbb{E}(Y_{\overline{a}G_{\overline{a}|v}}|v) - \mathbb{E}(Y_{\overline{a}G_{\overline{a}^*|v}}|v)\} &= \{\theta_{23}\beta_{12} + \theta_{25}\theta_{14}\beta_{12} + \beta_{21}\theta_{24} + \beta_{24}\theta_{12}\theta_{24}\} \\
&\quad \times [a(1) - a^*(1)] + \beta_{22}\theta_{24}[a(2) - a^*(2)]. \quad \blacksquare
\end{aligned}
$$

A.7. SELECTED TOPICS IN MEDIATION ANALYSIS

A.7.1. Multiple Versions of the Mediator and Ill-Defined Mediators

In this Section we will consider the consequences of ill-defined mediators and possibly having multiple versions of the mediator as pertains to the interpretation of estimators of direct and indirect effects. First, however, we will review analogous results for multiple versions of treatment. As noted in Section 1, under the potential outcomes framework, Y_a might be used denote the potential outcome Y for each individual if treatment or exposure A were set, possibly contrary to fact, to the value a. Articulating the potential outcomes framework in this way requires what Rubin called the "Stable Unit Treatment Value Assumption" or "SUTVA." Rubin (1980) points out that to be well-defined, notation such as Y_a effectively presupposes (i) that if individual j is given treatment a, then individual j's outcome under treatment a does not depend on which treatment individual $j' \neq j$ received and (ii) that there do not exist multiple versions of treatment a which might give rise to different outcomes depending on which version is administered.

The first of these assumptions is sometimes referred to as "no-interference," which Rubin (1980) attributes to Cox (1958); the second assumption is a "no-versions-of-treatment assumption" which Rubin attributes to Neyman (1935). Included also within SUTVA is an assumption that in other literature is sometimes referred to as consistency. As discussed in Section 1, the consistency assumption (Robins, 1986) states that $Y_a = Y$ when $A = a$—that is, that the value of Y which would have been observed if A had been set to what it in fact was is equal to the value of Y which was in fact observed. The consistency assumption ties the potential outcomes (or counterfactual data) to the observed data. Under Rubin's articulation of SUTVA, if there is only one version of treatment, then if $A = a$, the manner in which treatment A was in fact set to a is irrelevant, so Y_a is well-defined and is equal to Y when $A = a$. Rubin's SUTVA thus includes a no-multiple-versions-of-treatment assumption and this no-multiple-versions-of-treatment assumption itself includes the consistency assumption.

Consider now the context in which we might in fact have multiple versions of treatment. Suppose that there is some underlying version of treatment variable K but that the analyst has data on a coarsened variable A where each value of A corresponds to one or more values of K (i.e., the mapping from K to A is a many-to-one map). Let Y be the outcome and Y_k be the potential outcome for an individual if K had been k. Suppose that in an observational study for a set of covariates C we had $Y_k \perp\!\!\!\perp K|C$—that is, no confounding of the effect of K on Y conditional on covariates C. Suppose also the consistency assumption held for K such that $Y_k = Y$ when $K = k$. An analyst who had used the treatment A might then compute the "causal effect" comparing treatment levels $A = a$ and $A = a^*$ by calculating $\sum_c \mathbb{E}[Y|A = a, c]P(c) - \sum_c \mathbb{E}[Y|A = a^*, c]P(c)$. VanderWeele and Hernán (2013) showed that under the assumption of no unmeasured confounding for the effect of the versions variable K on the outcome Y, $Y_k \perp\!\!\!\perp K|C$, the following equality holds:

$$\begin{aligned}
&\sum\nolimits_c \mathbb{E}[Y|A = a, c]P(c) - \sum\nolimits_c \mathbb{E}[Y|A = a^*, c]P(c) \\
&\qquad = \sum\nolimits_{c,k} \mathbb{E}[Y_k|l]P(K = k|A = a, c)P(c) \\
&\qquad\quad - \sum\nolimits_{c,k} \mathbb{E}[Y_k|l]P(K = k|A = a^*, c)P(c)
\end{aligned}$$

This latter expression can itself be interpreted as a comparison in a randomized trial in which, within strata of covariates $C = c$, one arm is randomly assigned a "version of treatment" K from the observed distribution of K in the population amongst with $A = a$ and $C = c$ and the other arm is randomly assigned a "version of treatment" K from the observed distribution of K in the population amongst with $A = a^*$ and $C = c$.

Note that nothing in the analysis above required that K itself be continuous; the variable K might indicate a complex set of potential interventions that are coarsened in treatment variable A by partitioning the support of K. Thus, the ordinary estimator for the causal effect adjusting for C, namely, $\sum_c \mathbb{E}[Y|A = a, c]P(c) - \sum_c \mathbb{E}[Y|A = a^*, c]P(c)$, even in the presence of multiple versions of treatment (and

even when the treatment variable A does not unambiguously correspond to a single intervention), has the interpretation of a causal effect comparing two randomized interventions. Several points merit attention, however. First, the ordinary estimate only merits the interpretation of the effect comparing two randomized interventions if we have adequately controlled for confounding for the underlying version of treatment variable K. Second, even if we can potentially interpret the effect of treatment in this manner, an intervention corresponding to the treatment effect we are supposedly estimating will often not be possible to realistically implement in practice. Finally, if we do not know what the underlying version of treatment K is, it may be difficult to assess whether we have indeed controlled for all relevant confounders. Further discussion of these and related points is given elsewhere (Hernán and VanderWeele, 2011; VanderWeele and Hernán (2013).

Let us now turn to the somewhat analogous setting in which the treatment or exposure is well-defined and there is only one version of each exposure level but in which we are interested in mediation and are in a setting in which there are multiple versions of the mediator. In our mediation setting of Chapter 2, under the no-confounding assumptions (A2.1)–(A2.4) in the text, the natural indirect and direct effects and controlled direct effect are identified, respectively, by

$$Q_1 = \sum_m \mathbb{E}[Y|A=1,m,c]\{P(m|A=1,c) - P(m|A=0,c)\}$$
$$Q_2 = \sum_m \{\mathbb{E}[Y|A=1,m,c] - \mathbb{E}[Y|A=0,m,c]\}P(m|A=0,c)$$
$$Q_3 = \mathbb{E}[Y|A=1,m,c] - \mathbb{E}[Y|A=0,m,c]$$

Now consider the setting in which there are multiple versions of the mediator. For each possible value m of M, there will be some set of possible interventions—some set of versions of the mediator—that would fix M to m; we might denote this set by $\mathcal{K}^m$ (and perhaps denote the various versions by $1,2,3,\ldots$,etc.). For an individual with $M = m$ let K^m be the version that actually occurred so that M was at level m; for notational simplicity, for individuals with M not at level m we define $K^m = 0$ to indicate that there was not any version of the mediator that set M to m, since M was not level m. Following VanderWeele and Hernán (2013) for multiple versions of treatment for total effects, we let K denote a vector $(K^m : m \in \mathcal{M})$ where $\mathcal{M}$ is the support of M. The vector K will thus be all zeros except at one place, namely the place corresponding to the value of the mediator that actually occurred for that individual. The variable M is effectively a coarsening of K and thus we will have that Y and A are independent of M conditional on K as in Figure 7.1. We can then let Y_{ak} be the counterfactual outcome that would have occurred had A been set to a, K to k, and K_a to the value of K that would have occurred had A been set to a. The no-unmeasured-confounding assumptions for A and K with respect to Y can then be given as (A7.1) $Y_{ak} \perp\!\!\!\perp A|C$, (A7.2) $Y_{ak} \perp\!\!\!\perp K|\{A,C\}$, (A7.3) $K_a \perp\!\!\!\perp A|C$, and (A7.4) $Y_{ak} \perp\!\!\!\perp K_{a^*}|C$.

Let G_a be a random draw from the distribution of the version of the mediator that involves first randomly selecting a value of M from amongst those with $A = a$ and $C = c$ and then randomly selecting a version of the mediator from among those with $M = m, A = 1$, and $C = c$. We will then let Y_{1G_a} be the value of the

outcome that would arise if the exposure is set to 1 and the version of the mediator is randomly selected from the distribution G_a. Likewise we will let Y_{0G_a} be the value of the outcome that would arise if the exposure is set to 0 and the version of the mediator is randomly selected from the distribution G_a. We then have the following results.

Proposition 7.1 (VanderWeele, 2012a):
Under assumptions (A7.1)–(A7.4), the estimate, Q_1, of the natural indirect effect using data on only the mediator measurement M is equal to

$$Q_1 = \mathbb{E}[Y_{1G_1}|c] - \mathbb{E}[Y_{1G_0}|c]$$

and the estimate, Q_2, of the natural direct effect using data on only the mediator measurement M is equal to

$$Q_2 = \big(\mathbb{E}[Y_{1G_0}|c] - \mathbb{E}[Y_{0G_0}|c]\big) + \big(\mathbb{E}[Y_{0G_0}|c] - \mathbb{E}[Y_{0H_0}|c]\big)$$

where Y_{0H_0} is the value of the outcome that would arise if the exposure is set to 0 and the version of the mediator is randomly selected from a distribution H_0 that first randomly selects a value of M from among those with $A = 0$ and $C = c$ and then randomly selects a version of the mediator from among those with $M = m, A = 0$ and $C = c$.

Proof:
Under assumptions (A7.1)–(A7.4) we have that

$$\begin{aligned}\sum\nolimits_m \mathbb{E}[Y|a,m,c]P(m|a^*,c) &= \sum\nolimits_{k,m} \mathbb{E}[Y|a,k,m,c]P(k|a,m,c)P(m|a^*,c)\\ &= \sum\nolimits_k \mathbb{E}[Y|A=a,k,c]\sum\nolimits_m P(k|a,m,c)P(m|a^*,c)\\ &= \sum\nolimits_k \mathbb{E}[Y_{ak}|c]\sum\nolimits_m P(k|a,m,c)P(m|a^*,c)\end{aligned}$$

where the first equality follows by iterated expectations, the third by assumptions (A7.1) and (A7.2). When $a = 1$, then, by the definition of G_{a^*}, this is also equal to

$$\begin{aligned}&= \sum\nolimits_k \mathbb{E}[Y_{1k}|c]\sum\nolimits_m P(k|A=1,m,c)P(m|a^*,c)\\ &= \sum\nolimits_k \mathbb{E}[Y_{1k}|c]P(G_{a^*}=k|c)\\ &= \mathbb{E}[Y_{1G_{a^*}}|c].\end{aligned}$$

If data were only available on A, M, Y, and C, but not on version, then the estimators used for the natural indirect effect would be consistent for

$$\begin{aligned}Q_1 &= \sum\nolimits_m \mathbb{E}[Y|A=1,m,c]\{P(m|A=1,c) - P(m|A=0,c)\}\\ &= \mathbb{E}[Y_{1G_1}|c] - \mathbb{E}[Y_{1G_0}|c]\end{aligned}$$

thus establishing the result for natural indirect effects. If data were only available on A, M, Y, and C but not on version, then the estimators used for the natural direct effect would be consistent for

$$\begin{aligned}
Q_2 &= \sum\nolimits_m \{\mathbb{E}[Y|A=1,m,c] - \mathbb{E}[Y|A=0,m,c]\}P(m|A=0,c) \\
&= \sum\nolimits_m \mathbb{E}[Y|A=1,m,c]P(m|A=0,c) \\
&\quad - \sum\nolimits_{m,k} \mathbb{E}[Y|A=0,k,m,c]P(k|A=0,m,c)P(m|A=0,c) \\
&= \mathbb{E}[Y_{1G_0}|c] - \sum\nolimits_k \mathbb{E}[Y_{0k}|c] \sum\nolimits_m P(k|A=0,m,c)P(m|A=0,c) \\
&= \mathbb{E}[Y_{1G_0}|c] - \mathbb{E}[Y_{0H_0}|c] \\
&= (\mathbb{E}[Y_{1G_0}|c] - \mathbb{E}[Y_{0G_0}|c]) + (\mathbb{E}[Y_{0G_0}|c] - \mathbb{E}[Y_{0H_0}|c])
\end{aligned}$$

where the final lines follows by adding and subtracting $\mathbb{E}[Y_{0G_0}|c]$, thus establishing the result for natural direct effects. ■

Note that assumption (A7.3) is not strictly necessary for the proof of the result but is necessary for interpreting an expression such as $\mathbb{E}[Y_{1G_1}|c] - \mathbb{E}[Y_{1G_0}|c]$ as the effect of the exposure on the outcome mediated by version conditional on M and for interpreting $\mathbb{E}[Y_{0G_0}|c] - \mathbb{E}[Y_{0H_0}|c]$ as the effect of the exposure on the outcome mediated by version not captured by mediator measurement M. Moreover, although we did not make use of assumption (A7.4), if there were an effect of the exposure that confounded the mediator–outcome relationship, then assumption (A7.2) would not hold conditional on a set of baseline covariate C and thus the derivations above would not follow.

Proposition 7.2 (VanderWeele, 2012a):
Under assumptions (A7.1)–(A7.4), the estimate, Q_3, of the controlled direct effect using data on only the mediator measurement M is equal to

$$Q_3 = \mathbb{E}[Y_{1F_1} - Y_{0F_1}|c] + \mathbb{E}[Y_{0F_1} - Y_{0F_0}|c]$$

where F_1 is a random draw of a version from the distribution $P(k|A=1,m,c)$ and F_0 is a random draw of a version from the distribution $P(k|A=0,m,c)$.

Proof:
If data were only available on A, M, Y, and C but not on version, then the estimators used for the controlled direct effect would be consistent for

$$\begin{aligned}
Q_3 &= \mathbb{E}[Y|A=1,m,c] - \mathbb{E}[Y|A=0,m,c] \\
&= \sum\nolimits_k \mathbb{E}[Y|A=1,k,m,c]P(k|A=1,m,c) \\
&\quad - \sum\nolimits_k \mathbb{E}[Y|A=0,k,m,c]P(k|A=0,m,c) \\
&= \sum\nolimits_k \mathbb{E}[Y|A=1,k,c]P(k|A=1,m,c) - \sum\nolimits_k \mathbb{E}[Y|A=0,k,c]P(k|A=0,m,c)
\end{aligned}$$

$$
\begin{aligned}
&= \sum\nolimits_k \mathbb{E}[Y_{1k}|c]P(k|A=1,m,c) - \sum\nolimits_k \mathbb{E}[Y_{0k}|c]P(k|A=0,m,c) \\
&= \sum\nolimits_k \mathbb{E}[Y_{1k} - Y_{0k}|c]P(k|A=1,m,c) \\
&\quad + \sum\nolimits_k \mathbb{E}[Y_{0k}|c]\{P(k|A=1,m,c) - P(k|A=0,m,c)\}
\end{aligned}
$$

where the second to last equality follows by assumptions (A7.1) and (A7.2) and the final equality follows by adding and subtracting $\sum_k \mathbb{E}[Y_{0k}|c]P(k|A=1,m,c)$. From the definition of F_a we then also have that this is equal to

$$
\begin{aligned}
&= \sum\nolimits_k \mathbb{E}[Y_{1k} - Y_{0k}|c]P(F_1 = k|c) \\
&\quad + \sum\nolimits_k \mathbb{E}[Y_{0k}|c]P(F_1 = k|c) - \sum\nolimits_k \mathbb{E}[Y_{0k}|c]P(F_0 = k|c) \\
&= \mathbb{E}[Y_{1F_1} - Y_{0F_1}|c] + \mathbb{E}[Y_{0F_1} - Y_{0F_0}|c]
\end{aligned}
$$

thus establishing the result. ■

Another way to view the controlled direct effect estimand is thus as two parts where the first part is the controlled direct effect for the exposure on the outcome not through version, standardized by the distribution of the versions of the mediator amongst those with $A = 1, M = m, C = c$. The second part is a comparison of the counterfactual $\mathbb{E}[Y_{0k}|c]$ standardized by the distribution of versions amongst those with $A = 1, M = m, C = c$ versus amongst those with $A = 0, M = m, C = c$. This once again picks up an effect of the exposure on the version of the mediator not through the mediator measure M.

A.7.2. Interpretation of Direct and Indirect Effect Estimates Without the Cross-World Independence Assumption

Let $G_{a^*|c}$ denote a random draw from the distribution of the mediator when setting the exposure to a^* amongst those with covariates $C = c$. As in Section A.5.6, $NIE^R = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})$ is the randomized interventional analogue of the natural indirect effect; $NDE^R = \mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})$ is the randomized interventional analogue of the natural direct effect and $TE^R = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})$ is the randomized interventional analogue of the total effect. With effects thus defined we have the decomposition $\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}}) = \{\mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})\} + \{\mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})\}$ so that the overall effect decomposes into the sum of the effect through the mediator and the direct effect. The next proposition states that under assumptions (A2.1)–(A2.3)—that is, without assuming the cross-world independence assumption (A2.4)—the usual natural direct and indirect effect empirical estimands have the interpretation of these randomized interventional analogues of the natural direct and indirect effects.

Proposition 7.3 (Didelez et al., 2006):
If (A2.1) $Y_{am} \perp\!\!\!\perp A|C$, (A2.2) $Y_{am} \perp\!\!\!\perp M|\{A, C\}$, and (A2.3) $M_a \perp\!\!\!\perp A|C$ hold, then

$$\sum\nolimits_m \mathbb{E}[Y|A=1,m,c]\{P(m|A=1,c)-P(m|A=0,c)\} = \mathbb{E}(Y_{aG_{a|c}}) - \mathbb{E}(Y_{aG_{a^*|c}})$$

$$\sum\nolimits_m \{\mathbb{E}[Y|A=1,m,c]-\mathbb{E}[Y|A=0,m,c]\}P(m|A=0,c) = \mathbb{E}(Y_{aG_{a^*|c}}) - \mathbb{E}(Y_{a^*G_{a^*|C}})$$

Proof:
This follows from Proposition 5.6 taking $L = \varnothing$. ■

A.7.3. Direct and Indirect Effects in Health Disparities Research

Let R denote the race/ethnicity variable used in the regression. Let $R = 1$ indicate black and let $R = 0$ indicate white. Let M denote adult SES. Let Y denote the health outcome. Let C denote baseline covariate at the time of conception or early in life (e.g., family and neighborhood SES). Let $H_c(0)$ be a random draw from the adult SES distribution of the white population with baseline covariates c. Let Y_m denote an individual's counterfactual outcome if his or her adult SES were set to m. Then $\mathbb{E}[Y_{H_c(0)}|R=1,c]$ denotes the expected outcome for a black individual with baseline covariates c if their adult SES were set to a random draw from that of the white population with baseline covariates c. Note that $P(H_c(0)=m|c,r) = P(H_c(0)=m) = P(m|R=0,c)$. The associations between adult SES and the outcome reflect the actual effects of adult SES, that is, $\mathbb{E}[Y_m|R=1,c] = \mathbb{E}[Y|R=1,m,c]$. Methods from the mediation analysis literature for the natural direct effect conditional on C with R as the exposure, M as the mediator, and Y as the outcome effectively estimate

$$\begin{aligned}
&\sum\nolimits_m \mathbb{E}[Y|R=1,m,c]P(m|R=0,c) - \sum\nolimits_m \mathbb{E}[Y|R=0,m,c]P(m|R=0,c)\\
&\quad = \sum\nolimits_m \mathbb{E}[Y_m|R=1,H_c(0)=m,c]P(H_c(0)=m|R=1,c) - \mathbb{E}[Y|R=0,c]\\
&\quad = \mathbb{E}[Y_{H_c(0)}|R=1,H_c(0),c] - \mathbb{E}[Y|R=0,c]
\end{aligned}$$

Thus the "direct effect" that is obtained for race not through adult SES (when also controlling for baseline covariates c—for example, family SES and neighborhood SES at conception or early in life) could be interpreted as the health disparity that would remain for individuals with baseline covariates c if, within this population, the adult SES distribution of the black population were set equal to that of the white population.

Methods from the mediation analysis literature for the natural indirect effect conditional on C with R as the exposure, M as the mediator, and Y as the outcome effectively estimate

$$\sum\nolimits_m \mathbb{E}[Y|R=1,m,c]P(m|R=1,c) - \sum\nolimits_m \mathbb{E}[Y|R=1,m,c]P(m|R=0,c)$$

Similarly, as above, let $H_c(1)$ be a random draw from the adult SES distribution of the black population with baseline covariates c so that $\mathbb{E}[Y_{H_c(1)}|R=1,c]$ denotes

the expected outcome for a black individual with baseline covariates c if their adult SES were set to a random draw from that of the black population with baseline covariates c. Note that $P(H_c(1) = m) = P(H_c(1) = m|c,r) = P(m|R = 1,c)$. If the associations between adult SES, and the outcome reflect the actual effects of adult SES, then $\mathbb{E}[Y_m|R = 1,c] = \mathbb{E}[Y|R = 1,m,c]$. We thus have

$$\begin{aligned}
&\sum\nolimits_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 1,c) - \sum\nolimits_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 0,c) \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_m|R = 1,H_c(1) = m,c]P(H_c(1) = m|R = 1,c) \\
&\qquad - \sum\nolimits_m \mathbb{E}[Y_m|R = 1,H_c(0) = m,c]P(H_c(0) = m|R = 1,c) \\
&\quad = \mathbb{E}[Y_{H_c(1)}|R = 1,c] - \mathbb{E}[Y_{H_c(0)}|R = 1,c]
\end{aligned}$$

The "mediated effect" can thus be interpreted as how the health outcomes for the black population with baseline covariates c would change if the adult SES distribution of this black population were set equal to that of the black population versus that of the white population.

The overall disparity measure for those with early family and neighborhood SES of c is given by

$$\begin{aligned}
&\mathbb{E}[Y|R = 1,c] - \mathbb{E}[Y|R = 1,c] \\
&\quad = \sum\nolimits_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 1,c) - \sum\nolimits_m \mathbb{E}[Y|R = 0,m,c]P(m|R = 0,c) \\
&\quad = \sum\nolimits_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 1,c) - \sum\nolimits_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 0,c) \\
&\qquad + \sum\nolimits_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 0,c) - \sum\nolimits_m \mathbb{E}[Y|R = 0,m,c]P(m|R = 0,c) \\
&\quad = \{\mathbb{E}[Y_{H_c(1)}|R = 1,c] - \mathbb{E}[Y_{H_c(0)}|R = 1,c]\} + \mathbb{E}[Y_{H_c(0)}|R = 1,c] - \mathbb{E}[Y|R = 0,m,c]
\end{aligned}$$

where the second equality is obtained by adding and subtracting $\sum_m \mathbb{E}[Y|R = 1,m,c]P(m|R = 0,c)$ and, in the third equality, the two expressions are simply the "direct effect" and "mediated effect" disparities measures given above.

A similar interpretation would hold for binary outcomes on an odds ratio scale, provided that the outcome is rare (cf. VanderWeele and Vansteelandt, 2010). If the outcome is continuous and there are no statistical interactions between R and M, then the coefficient for R in the model that includes M (and C) will give the empirical quantity used to estimate the direct effect, and the difference in the coefficients for race in the models without versus with adult SES will give the empirical quantity used to estimate the mediated effect (VanderWeele and Vansteelandt, 2009). For a binary outcome with logistic regression, provided that the outcome is rare (or if a log-linear model is used with a common outcome) and if there are no statistical interactions between R and M, then once again the coefficient for R in the model that includes M (and C) will give the empirical quantity used to estimate the direct effect, and the difference in the coefficients for race in the models without versus with adult SES will give the empirical quantity used to estimate the mediated effect (VanderWeele and Vansteelandt, 2010).

A.7.4. A Three-Way Decomposition into Direct, Indirect, and Interactive Effects

Proposition 7.4 (VanderWeele, 2013b):
We have the decomposition: $Y_1 - Y_0 = (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$.

Proof:
We have that $Y_1 - Y_0 = Y_{1M_1} - Y_{0M_0} = (Y_{1M_1} - Y_{1M_0}) + (Y_{1M_0} - Y_{0M_0})$. By adding and subtracting the pure indirect effect, $(Y_{0M_1} - Y_{0M_0})$, we obtain

$$Y_1 - Y_0 = (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) + \{(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})\}$$

The third quantity in this decomposition is the difference between the total indirect effect and the pure indirect effect as described in Section A.2.1. This quantity is also equal to the difference between the total direct effect and the pure direct effect, $(Y_{1M_1} - Y_{0M_1}) - (Y_{1M_0} - Y_{0M_0})\}$. We will consider the value that this difference between the total indirect and the pure indirect effect, $(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})$, might take under several different scenarios. If $M_0 = M_1$, then both indirect effects are 0 and so the difference is 0. If $M_1 = 1$ and $M_0 = 0$, then $(M_1 - M_0) = 1$ and the difference will be $(Y_{11} - Y_{10} - Y_{01} + Y_{00}) = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$. If $M_1 = 0$ and $M_0 = 1$, then $(M_1 - M_0) = -1$ and the difference will be $(-Y_{11} + Y_{10} + Y_{01} - Y_{00}) = (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$. Thus, the difference $(Y_{1M_1} - Y_{1M_0}) - (Y_{0M_1} - Y_{0M_0})$ is always equal to $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$ and we have $Y_1 - Y_0 = (Y_{1M_0} - Y_{0M_0}) + (Y_{0M_1} - Y_{0M_0}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)$. ■

Proposition 7.5 (VanderWeele, 2013b):
If (A2.4) $Y_{am} \perp\!\!\!\perp M_{a^*} | C$ holds, then

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|c] = {} & \mathbb{E}[Y_{1M_0} - Y_{0M_0}|c] + \mathbb{E}[Y_{0M_1} - Y_{0M_0}|c] + \mathbb{E}[Y_{11} - Y_{10} - Y_{01} \\ & + Y_{00}|c]\mathbb{E}[M_1 - M_0|c]\end{aligned}$$

Proof:
We have that $\mathbb{E}[Y_a - Y_{a^*}|c] =$

$$\begin{aligned}& \mathbb{E}[Y_{aM_a} - Y_{a^*M_a}|c] + \mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\ & = \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] + \mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] + \{\mathbb{E}[Y_{aM_a} - Y_{a^*M_a}|c] \\ & \quad - \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]\}\end{aligned}$$

where the first quantity is the conditional pure direct effect, the second is the conditional pure indirect effect and the third is the difference between the conditional total direct effect and the conditional pure direct effect. Under assumption (A2.4), namely $Y_{am} \perp\!\!\!\perp M_{a^*} | C$, we have that this difference is

$$
\begin{aligned}
&\{\mathbb{E}[Y_{aM_a} - Y_{a^*M_a}|c] - \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]\} \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|M_a = m, c]P(M_a = m|c) \\
&\qquad - \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|M_{a^*} = m, c]P(M_{a^*} = m|c) \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]P(M_a = m|c) - \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]P(M_{a^*} = m|c) \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\}
\end{aligned}
$$

where m^* is an arbitrary value of M and where the first equality follows by iterated expectations, the second by assumption (A2.4), and the fourth because for some fixed level of m^*, $\sum_m \mathbb{E}[Y_{a^*m}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} = 0$ and $\sum_m \mathbb{E}[Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} = 0$. Thus, for arbitrary exposure and mediator, under (A2.4) we have the decomposition of the conditional effect:

$$
\begin{aligned}
\mathbb{E}[Y_a - Y_{a^*}|c] = {} & \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] + \mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] \\
& + \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) \\
& - P(M_{a^*} = m|c)\}
\end{aligned}
$$

where the first term is the pure direct effect, the second is the pure indirect effect, and the third is a mediated interactive effect. If we consider binary exposure and mediator with $a = 1, a^* = 0, m^* = 0$ we have

$$
\begin{aligned}
&\sum\nolimits_m \mathbb{E}[Y_{1m} - Y_{0m} - Y_{10} + Y_{00}|c]\{P(M_1 = m|c) - P(M_0 = m|c)\} \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{1m} - Y_{0m}|c]\{P(M_1 = m|c) - P(M_0 = m|c)\} \\
&\quad = \mathbb{E}[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
&\qquad + \mathbb{E}[Y_{10} - Y_{00}|c]\{P(M_1 = 0|c) - P(M_0 = 0|c)\} \\
&\quad = \mathbb{E}[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
&\qquad + \mathbb{E}[Y_{10} - Y_{00}|c][1 - P(M_1 = 1|c) - \{1 + P(M_0 = 1|c)\}] \\
&\quad = \mathbb{E}[Y_{11} - Y_{01}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
&\qquad - \mathbb{E}[Y_{10} - Y_{00}|c]\{P(M_1 = 1|c) - P(M_0 = 1|c)\} \\
&\quad = \mathbb{E}[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\{\mathbb{E}[M_1|c] - \mathbb{E}[M_0|c]\}
\end{aligned}
$$

and so we have

$$
\begin{aligned}
\mathbb{E}[Y_1 - Y_0|c] = {} & \mathbb{E}[Y_{1M_0} - Y_{0M_0}|c] + \mathbb{E}[Y_{0M_1} - Y_{0M_0}|c] \\
& + \mathbb{E}[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\mathbb{E}[M_1 - M_0|c]. \quad \blacksquare
\end{aligned}
$$

Proposition 7.6 (VanderWeele, 2013b):
Under assumptions (A2.1)–(A2.4), $\mathbb{E}[Y_{1M_0} - Y_{0M_0}|c]$, $\mathbb{E}[Y_{0M_1} - Y_{0M_0}|c]$, and $\mathbb{E}[(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0)|c]$ are all identified.

Proof:
In Section A.2 it was established that under assumptions (A2.1)–(A2.4) we have

$$\mathbb{E}[Y_{1M_0} - Y_{0M_0}|c] = \sum\nolimits_m \{\mathbb{E}[Y|A=1,m,c] - \mathbb{E}[Y|A=0,m,c]\}P(m|A=0,c)$$
$$\mathbb{E}[Y_{0M_1} - Y_{0M_0}|c] = \sum\nolimits_m \mathbb{E}[Y|A=0,m,c]\{P(m|A=1,c) - P(m|A=0,c)\}$$

We have shown in Proposition 7.5 that under (A2.4) we have

$$\mathbb{E}[(Y_{11}-Y_{10}-Y_{01}+Y_{00})(M_1-M_0)|c] = \mathbb{E}[Y_{11}-Y_{10}-Y_{01}+Y_{00}|c]\mathbb{E}[M_1-M_0|c]$$

Under (A2.1) and (A2.2) the first term in this product is equal to $\{\mathbb{E}[Y|A=1,M=1,c] - \mathbb{E}[Y|A=1,M=0,c] - \mathbb{E}[Y|A=0,M=1,c] + \mathbb{E}[Y|A=0,M=0,c]\}$, and under (A2.3) the second term in this product is equal to $\mathbb{E}[M|A=1,c] - \mathbb{E}[M|A=0,c]$. ■

On a risk ratio scale, the conditional total effect risk ratio is defined by $RR_c^{TE} = \mathbb{E}[Y_a|c]/\mathbb{E}[Y_{a^*}|c]$; the pure direct effect risk ratio is defined by $RR_c^{DE} = \mathbb{E}[Y_{aM_{a^*}}|c]/\mathbb{E}[Y_{a^*M_{a^*}}|c]$ and the pure indirect effect risk ratio is defined by $RR_c^{IE} = \mathbb{E}[Y_{a^*M_a}|c]/\mathbb{E}[Y_{a^*M_{a^*}}|c]$.

Proposition 7.7 (VanderWeele, 2013b):
We have the decomposition $(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}$ where $RERI_{mediated} = \left(\frac{\mathbb{E}[Y_{aM_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{aM_{a^*}}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{a^*M_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} + 1\right)$.

Proof:
We have that

$$\mathbb{E}[Y_a - Y_{a^*}|c] = \mathbb{E}[Y_{aM_a} - Y_{a^*M_a}|c] + \mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c]$$
$$\mathbb{E}[Y_a - Y_{a^*}|c] = \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] + \mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] + \{\mathbb{E}[Y_{aM_a} - Y_{a^*M_a}|c] - \mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]\}$$

Dividing both sides of the equation by $\mathbb{E}[Y_{a^*M_{a^*}}|c]$ gives

$$(RR_c^{TE} - 1) = (RR_c^{DE} - 1) + (RR_c^{IE} - 1) + \left(\frac{\mathbb{E}[Y_{aM_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{aM_{a^*}}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{a^*M_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} + 1\right)$$
$$(RR_c^{DE} - 1) + (RR_c^{IE} - 1) + RERI_{mediated}. \quad ■$$

If we define the total indirect effect risk ratio as $RR_c^{TIE} = \mathbb{E}[Y_{aM_a}|c]/\mathbb{E}[Y_{aM_{a^*}}|c]$, then the total effect risk ratio decomposes as $RR_c^{TE} = RR_c^{TIE} \times RR_c^{DE}$. VanderWeele and Vansteelandt (2010) proposed as a measure of the proportion mediated on the

risk difference scale the measure $\frac{RR_c^{DE}(RR_c^{TIE}-1)}{(RR_c^{TE}-1)}$. The numerator in this quantity is in fact equal to

$$\begin{aligned}
RR_c^{DE}(RR_c^{TIE}-1) &= RR_c^{TE} - RR_c^{DE} \\
&= \frac{\mathbb{E}[Y_{aM_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{aM_{a^*}}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} \\
&= \left(\frac{\mathbb{E}[Y_{a^*M_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - 1\right) \\
&\quad + \left(\frac{\mathbb{E}[Y_{aM_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{aM_{a^*}}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{a^*M_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} + 1\right) \\
&= (RR_c^{IE} - 1) + RERI_{mediated}
\end{aligned}$$

that is, the sum of the excess relative risk for the pure indirect effect plus the mediated relative excess risk due to interaction.

By a similar argument as in Proposition 7.7, if we let T denote a time-to-event outcome and let T_a denote the counterfactual event time if A had been set to a and likewise if we let T_{am} denote the counterfactual event time if A had been set to a and M had been set to m and use $\lambda_V(t)$ and $\lambda_V(t|c)$ to denote the hazard and conditional hazard, respectively, for a time-to-event variable V, then we have the decomposition on a hazard ratio scale of

$$\begin{aligned}
\left(\frac{\lambda_{T_a}(t)}{\lambda_{T_{a^*}}(t)} - 1\right) &= \left(\frac{\lambda_{T_{aM_{a^*}}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - 1\right) + \left(\frac{\lambda_{T_{a^*M_a}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - 1\right) \\
&\quad + \left(\frac{\lambda_{T_{aM_a}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - \frac{\lambda_{T_{aM_{a^*}}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} - \frac{\lambda_{T_{a^*M_a}}(t)}{\lambda_{T_{a^*M_{a^*}}}(t)} + 1\right)
\end{aligned}$$

Similarly, in a setting in which there is a variable L that is affected by exposure A and in turn affects both M and Y as in Figure 7.2, we have a three-way decomposition using the randomized interventional analogues of natural direct and indirect effects considered in Section 5.4. Let $G_{a|c}$ denote a random draw from the distribution of the mediator amongst those with exposure status a conditional on $C = c$. As noted in Section A.5, we have the decomposition $\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c) = \{\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{aG_{a^*|c}}|c)\} + \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\}$ so that the total effect decomposes into the sum of the effect through the mediator and the direct effect. These effects arise from randomly choosing for each individual a value of the mediator from the distribution of the mediator amongst all of those with a particular exposure.

We might further decompose this as follows:

$$\begin{aligned}
&\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c) \\
&= \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} + \{\mathbb{E}(Y_{a^*G_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} \\
&\quad + [\{\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a|c}}|c)\} - \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\}]
\end{aligned}$$

where the first term in the decomposition is the randomized interventional analogue of the pure direct effect, the second is the randomized interventional analogue of the pure indirect effect, and the third is the difference between the randomized interventional analogue of the total direct effect and the pure direct effect. We now show that this third term in fact has the interpretation of a mediated interaction. We have that

$$\begin{aligned}
&\{\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a|c}}|c)\} - \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|G_{a|c} = m, c]P(G_{a|c} = m|c) \\
&\qquad - \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|G_{a^*|c} = m, c]P(G_{a^*|c} = m|c) \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]P(M_a = m|c) - \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]P(M_{a^*} = m|c) \\
&\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\}
\end{aligned}$$

where m^* is an arbitrary value of M. This final expression can be interpreted as a measure of interaction. If we set $a = 1, a^* = 0$, and $m^* = 0$, then for binary exposure and mediator, even in the presence of an exposure induced mediator—outcome confounder, we would have the three-way effect decomposition:

$$\begin{aligned}
\mathbb{E}(Y_{1G_{1|c}}|c) - \mathbb{E}(Y_{0G_{0|c}}|c) = &\{\mathbb{E}(Y_{1G_{0|c}}|c) - \mathbb{E}(Y_{0G_{0|c}}|c)\} + \{\mathbb{E}(Y_{0G_{1|c}}|c) \\
&- \mathbb{E}(Y_{0G_{0|c}}|c)\} \\
&+ \mathbb{E}[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c]\{\mathbb{E}[M_1|c] - \mathbb{E}[M_0|c]\}
\end{aligned}$$

As in Chapter 5, these effects would be identified under assumptions (A5.1) $Y_{am} \perp\!\!\!\perp A|C$, (A5.2) $Y_{am} \perp\!\!\!\perp M|\{A, C, L\}$, and (A5.3) $M_a \perp\!\!\!\perp A|C$ above, that is, that conditional on C there is no unmeasured exposure–outcome or exposure–mediator confounding, along with an assumption (A5.2), that is, that conditional on (A, C, L), there is no unmeasured confounding of the mediator–outcome relationship. These three assumptions would hold in the causal diagram in Figure 7.2.

For Y and M continuous, under assumptions (A2.1)–(A2.4) and correct specification of the regression models for Y and M:

$$\begin{aligned}
\mathbb{E}[Y|a, m, c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\
\mathbb{E}[M|a, c] &= \beta_0 + \beta_1 a + \beta_2' c
\end{aligned}$$

we have from Section A.2.2 that the pure direct effect is given by

$$\mathbb{E}[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$$

and that the total indirect effect was given by

$$\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

and likewise the pure indirect effect was given by

$$\mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] = (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)$$

The mediated interactive effect is given by the difference between the total indirect effect and the pure indirect effect and is thus equal to

$$\begin{aligned}\mathbb{E}[Y_{aM_a} - Y_{aM_{a^*}}|c] - \mathbb{E}[Y_{a^*M_a} - Y_{a^*M_{a^*}}|c] &= (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*) \\ &\quad -(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*) \\ &= \theta_3\beta_1(a - a^*)(a - a^*)\end{aligned}$$

Suppose now instead that Y were binary and M continuous, that assumptions (A2.1)–(A2.4) held, that the outcome was rare, and that the following regressions were correctly specified:

$$\begin{aligned}\text{logit}(P(Y = 1|a, m, c)) &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ \mathbb{E}[M|a, c] &= \beta_0 + \beta_1 a + \beta_2' c\end{aligned}$$

In Section A.2.2, we showed that the pure direct effect risk ratio and the pure indirect effect risk ratio were given by

$$\begin{aligned}RR_c^{DE} &= \exp\left[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right] \\ RR_c^{IE} &= \exp\left[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)\right]\end{aligned}$$

where σ^2 is the variance of the error term in the linear regression model for M. The total effect is given by

$$\begin{aligned}RR_c^{TE} = \exp\big[&\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) \\ &+ 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\big]\end{aligned}$$

and from this it follows that $RERI_{mediated}$ is equal to

$$\begin{aligned}&\left(\frac{\mathbb{E}[Y_{1M_1}|c]}{\mathbb{E}[Y_{0M_0}|c]} - \frac{\mathbb{E}[Y_{1M_0}|c]}{\mathbb{E}[Y_{0M_0}|c]} - \frac{\mathbb{E}[Y_{0M_1}|c]}{\mathbb{E}[Y_{0M_0}|c]} + 1\right) \\ &= \exp\left[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*)\right. \\ &\quad \left. + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right] \\ &\quad - \exp\left[\{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\right] \\ &\quad - \exp\left[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)\right] + 1.\end{aligned}$$

A.8. OTHER TOPICS RELATED TO INTERMEDIATES

A.8.1. Principal Stratification

Let A denote some binary treatment or exposure variable and Y some outcome, and let S denote a variable that occurs between A and Y. We will use S rather than M as our intermediate variable in this section because the approaches that are considered do not really correspond to mediation. We let S_a denote the potential outcome

or counterfactual outcome for each individual that we would have observed had A, possibly contrary to fact, been a. A principal stratum is simply a subgroup of individuals homogeneous in their joint potential outcomes (S_0, S_1). If S is also binary, then we have four principal strata: $(S_0 = 0, S_1 = 0)$, sometimes called "never-takers"; $(S_0 = 0, S_1 = 1)$, sometimes called "compliers"; $(S_0 = 1, S_1 = 0)$, sometimes called "defiers"; and $(S_0 = 1, S_1 = 1)$, sometimes called "always takers."

Suppose now that the outcome Y is measured after some individuals die where $S = 1$ denotes survival. For those who die $(S = 0)$ the outcome is not simply missing but undefined. We could attempt to simply compare outcomes amongst those who actually survived $(S = 1)$: We could examine the contrast $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$. However, survival is a post-treatment variable and it may be affected by treatment; conditioning on it would essentially break randomization and could induce bias. An alternative comparison that would make sense in this setting is to compare the quality-of-life outcomes for the group that would have survived irrespective of which treatment they were given. In the notation given above, this is $\mathbb{E}[Y_1 - Y_0|S_0 = 1, S_1 = 1]$. This is a principal strata causal effect, sometimes referred to as the survivor average causal effect (SACE). In this context in which outcomes are effectively censored or truncated due to death, this is really the only comparison that is fair. The principal stratification approach is thus of considerable importance in addressing these questions, and a number of papers have provided methods to try to assess this survivor average causal effect when outcomes are truncated due to death (Robins, 1986; Zhang and Rubin, 2003; Hayden et al., 2005; Rubin, 2006; Frangakis et al., 2007; Imai, 2008; Egleston et al., 2009; Chiba and VanderWeele, 2011). Here we will present one particularly simple sensitivity analysis approach (Chiba and VanderWeele, 2011). Assessing the survivor average causal effect is made considerably easier by an assumption sometimes referred to as "monotonicity." Stated formally, the monotonicity assumption is (A8.1) $S_0 \leq S_1$. The assumption states that, for all individuals, survival under the treatment is always at least as good as survival under the control condition.

Proposition 8.1 (Chiba and VanderWeele, 2011):
Suppose (A8.1) $S_0 \leq S_1$, then the survivor average causal effect, $\mathbb{E}[Y_1 - Y_0|S_0 = 1, S_1 = 1]$, is equal to

$$\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1] - \alpha$$

where $\alpha = \mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1]$

Proof:
Under assumption (A8.1) $S_0 \leq S_1$, $S_0 = 1$ and $S_1 = 0$ cannot hold simultaneously, and thus $P(S_1 = 0, S_0 = 1) = 0$, which implies there are no defiers. When there are no defiers, individuals with the observed values of $A = 0$ and $S = 1$ are limited to the always-survivors. We thus have $\mathbb{E}[Y_a|A = 0, S = 1] = \mathbb{E}[Y_a|S_1 = S_0 = 1]$. Therefore, the survivor average causal effect is given by

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|S_0 = 1, S_1 = 1] &= \mathbb{E}[Y_1|A = 0, S = 1] - \mathbb{E}[Y_0|A = 0, S = 1] \\ &= (\mathbb{E}[Y_1|A = 1, S = 1] - \alpha) - \mathbb{E}[Y_0|A = 0, S = 1] \\ &= \mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1] - \alpha\end{aligned}$$

where the second equality follows because $\alpha = \mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1]$, and the third equality follows because $\mathbb{E}[Y_a|A = a, S = 1] = \mathbb{E}[Y|A = a, S = 1]$. This completes the proof. ■

The result states that, to obtain the survivor average causal effect, one can use the crude difference in outcomes Y between the treated and control subjects amongst those who survived, $\mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$, and then subtract the sensitivity analysis parameter α. The result depends critically on the monotonicity assumption (A8.1). Assessing the survivor average causal effect is more complicated when this monotonicity assumption does not hold, but some methods are available (Hayden et al., 2005; Chiba and VanderWeele, 2011).

The parameter α itself is the average difference in the outcome that would have been observed under treatment comparing two different populations: The first population is the population that would have survived under treatment $(A = 1, S = 1)$; the second population is the population that would have survived without treatment $(A = 0, S = 1)$. Because the second population consists of individuals who survived even without treatment, it will likely overall be a healthier population than the population who would have survived with treatment and we thus might expect $\alpha \leq 0$. If so, we have the following immediate corollary.

Corollary (Chiba and VanderWeele, 2011)
Suppose (A8.1) $S_0 \leq S_1$ and (A8.2) $\alpha = \mathbb{E}[Y_1|A = 1, S = 1] - \mathbb{E}[Y_1|A = 0, S = 1] \leq 0$, then

$$\mathbb{E}[Y_1 - Y_0|S_0 = 1, S_1 = 1] \geq \mathbb{E}[Y|A = 1, S = 1] - \mathbb{E}[Y|A = 0, S = 1]$$

We will now consider a somewhat different context. Let A be a randomized treatment, S compliance status (assume that compliance is all or nothing), and Y the outcome. The overall affect of treatment assignment on the outcome, $\mathbb{E}[Y_1 - Y_0]$, is simply given by a comparison of the average outcomes in the treatment versus the control groups, $\mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$. The effect is sometimes referred to as the intent-to-treat estimate. We might, however, instead be interested in the effect of the treatment taken (not simply assigning treatment). Suppose now that we were willing to assume that there was no effect of treatment assignment A on the outcome Y except through the actual treatment taken (compliance status) S. Suppose further that we were willing to assume that there were no defiers—that is, no one who would take the treatment if assigned to the control group, but who would not take the treatment if assigned to the treatment group (no one with $S_0 = 1, S_1 = 0$). We then have the following result. The result is stated so as to allow randomization within strata of covariates C.

Proposition 8.2 (Angrist et al., 1996):
If $S_a \perp\!\!\!\perp A|C$, $Y_a \perp\!\!\!\perp A|C$, $Y_{as} = Y_s$ and if $a^* \leq a$, implies $S_{a^*} \leq S_a$ then $\mathbb{E}[Y_{s=1} - Y_{s=0}|S_{a^*} < S_a, c] = \frac{\mathbb{E}[Y|A=a,c]-\mathbb{E}[Y|A=a^*,c]}{\mathbb{E}[S|A=a,c]-\mathbb{E}[S|A=a^*,c]}$.

Proof:
We have that

$$\begin{aligned}
&\mathbb{E}[Y_a - Y_{a^*}|c] \\
&\quad = \mathbb{E}[Y_a - Y_{a^*}|S_{a^*} < S_a, c]P(S_{a^*} < S_a|c) + \mathbb{E}[Y_a - Y_{a^*}|S_a = S_{a^*}, c]P(S_a = S_{a^*}|c) \\
&\quad = \mathbb{E}[Y_a - Y_{a^*}|S_{a^*} < S_a, c]P(S_{a^*} < S_a|c) + \mathbb{E}[Y_{aS_a} - Y_{a^*S_{a^*}}|S_a = S_{a^*}, c]P(S_a = S_{a^*}|c) \\
&\quad = \mathbb{E}[Y_a - Y_{a^*}|S_{a^*} < S_a, c]P(S_{a^*} < S_a|c) + \mathbb{E}[Y_{S_a} - Y_{S_a}|S_a = S_{a^*}, c]P(S_a = S_{a^*}|c) \\
&\quad = \mathbb{E}[Y_{S=1} - Y_{S=0}|S_{a^*} < S_a, c](\mathbb{E}[S_a|c] - \mathbb{E}[S_{a^*}|c])
\end{aligned}$$

Therefore, $\mathbb{E}[Y_{s=1} - Y_{s=0}|S_{a^*} < S_a, c] = \frac{\mathbb{E}[Y_a - Y_{a^*}|c]}{\mathbb{E}[S_a|c]-\mathbb{E}[S_{a^*}|c]} = \frac{\mathbb{E}[Y_a|c]-\mathbb{E}[Y_{a^*}|c]}{\mathbb{E}[S_a|c]-\mathbb{E}[S_{a^*}|c]} = \frac{\mathbb{E}[Y|A=a,c]-\mathbb{E}[Y|A=a^*,c]}{\mathbb{E}[S|A=a,c]-\mathbb{E}[S|A=a^*,c]}$. ■

With binary A and no covariates C the result reduces to $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1] = \frac{\mathbb{E}[Y|A=1]-\mathbb{E}[Y|A=0]}{\mathbb{E}[S|A=1]-\mathbb{E}[S|A=0]}$. The expression $\frac{\mathbb{E}[Y|A=1]-\mathbb{E}[Y|A=0]}{\mathbb{E}[S|A=1]-\mathbb{E}[S|A=0]}$ is sometimes referred to as the "instrumental variable estimator." Under the assumptions of Proposition 8.2 it is equal to the effect of treatment taken S among the compliers—that is, among those who would taken the treatment if assigned to the treatment group and would not if assigned to the control group. The principal stratum causal effect that is estimated, $\mathbb{E}[Y_1 - Y_0|S_0 = 0, S_1 = 1]$, is sometimes now referred to as the local average treatment effect (LATE) or the complier average causal effect (CACE).

A.8.2. Surrogate Outcomes

The following definitions and results were given in work on signed causal directed acyclic graphs (VanderWeele et al., 2008; VanderWeele and Robins, 2009b, 2010) and have immediate implications for the surrogacy results discussed in the text. Suppose that variable A is a parent of some variable Y and let $\widetilde{pa}_Y$ denote the parents of Y other than A. We say that A has a distributional positive monotonic effect on Y if the survivor function $S = (y|, \widetilde{pa}_Y) = P(Y > y|A = a, \widetilde{pa}_Y)$ is such that whenever $a_1 \geq a_0$ we have $S(y|a_1, \widetilde{pa}_Y) \geq S(y|a_0, \widetilde{pa}_Y)$ for all y and all $\widetilde{pa}_Y$; the variable A is said to have a distributional negative monotonic effect on Y if whenever $a_1 \geq a_0$ we have $S(y|a_1, \widetilde{pa}_Y) \leq S(y|a_0, \widetilde{pa}_Y)$ for all y and all $\widetilde{pa}_Y$. An edge on a causal directed acyclic graph (Pearl, 2009) from A to Y is said to be of positive or negative sign if, respectively, A has a distributional positive or negative monotonic effect on Y; if an edge is neither positive nor negative, it is said to be without a sign. The sign of a path on a causal directed acyclic graph is the product of the signs of the edges that constitute that path. If one of the edges on a path is without a sign then the sign, of the path is said to be undefined.

Proposition 8.3 (VanderWeele, 2013d):
In the causal diagram in Figure 8.4, if (a) $P(Y > y|a,s,u)$ is nondecreasing in a and s for all y,u and (b) $P(S > s|a,u)$ is nondecreasing in a for all s,u then $P(Y_a > y)$ is nondecreasing in a.

Proof:
VanderWeele and Robins (2009b) showed the following lemma: If X denotes some set of nondescendents of A that blocks all backdoor paths from A to Y and if all directed paths between A and Y are of positive sign, then $P(Y > y|a,x)$ is nondecreasing in a for all y; if all directed paths between A and Y are of negative sign, then $P(Y > y|a,x)$ is nonincreasing in a for all y. Proposition 8.3 follows immediately from this lemma. ■

Proposition 8.4 (VanderWeele, 2013d):
In the causal diagram in Figure 8.4, if (a) $\mathbb{E}(Y|a,s,u)$ is nondecreasing in a and s for all u and (b) $P(S > s|a,u)$, is nondecreasing in a for all s,u, then $\mathbb{E}(Y_a)$ is nondecreasing in a.

Proof: To prove this proposition, we use variants of definitions in VanderWeele and Robins (2009b). Suppose that variable A is a parent of some variable Y and let $\widetilde{pa}_Y$ denote the parents of Y other than A. We say that A has an average positive monotonic effect on Y if whenever $a_1 \geq a_0$ we have $\mathbb{E}(Y|a_1,\widetilde{pa}_Y) \geq \mathbb{E}(Y|a_0,\widetilde{pa}_Y)$ for all y and all $\widetilde{pa}_Y$; the variable A is said to have an average negative monotonic effect on Y if whenever $a_1 \geq a_0$ we have $\mathbb{E}(Y|a_1,\widetilde{pa}_Y) \leq \mathbb{E}(Y|a_0,\widetilde{pa}_Y)$ for all y and all $\widetilde{pa}_Y$. A directed path that is of positive sign with the exception that the parent of the final edge may only have an average monotonic effect on the child, rather than a distributional monotonic effect, will be said to be a directed path with mean positive sign. A directed path that is of negative sign with the exception that the parent of the final edge may only have an average monotonic effect on the child, rather than a distributional monotonic effect, will be said to be a directed path with mean negative sign.

It suffices to show that if X denotes some set of nondescendents of A that blocks all backdoor paths from A to Y and if all directed paths between A and Y are of mean positive sign, then $\mathbb{E}(Y|a,x)$ is nondecreasing in a for all x; if all directed paths between A and Y are of mean negative sign, then $\mathbb{E}(Y|a,x)$ is nonincreasing in a for all x. The proof of this follows from the proof of Proposition 4 in VanderWeele and Robins (2009b) by simply replacing "$\mathbb{E}[1(Y > y)|\ldots]$" by "$\mathbb{E}[Y|\ldots]$" wherever the former expression appears in the proof. ■

A.8.3. Instrumental Variables and Mendelian Randomization

Proposition 8.2 gave the general instrumental variable result for local average treatment effects for a binary exposure. For sensitivity analysis, consider Figure 8.14 in which we have both a violation of the exclusion restriction (an arrow for G to Y not through X) and linkage disequilibrium (another genetic marker G_U that was in linkage disequilibrium with G and affected the outcome Y). Suppose first that the

outcome Y were dichotomous and that X were continuous, and we are comparing two levels of the genetic marker G. For simplicity we denote these by $G = 0$ and $G = 1$, respectively. Suppose that the outcome were rare and that we estimated a risk ratio between G and Y of ϕ, possibly conditional on measured baseline covariates, and that we estimated the effect of G on X on the difference scale to be η, conditional on those same covariates. The standard instrumental variable estimate for the effect of X on Y on the risk ratio scale would be $\phi^{1/\eta}$ (Didelez et al., 2010). Under violations of the exclusion restriction and/or in the presence of linkage disequilibrium, as in Figure 8.14, this would be biased. Suppose we assume that G_U is binary and does not interact with G on the risk ratio scale in its effects on Y and that G_U increases the risk of Y by a factor of γ on the risk ratio scale. Suppose also that although the exclusion restriction is violated (i.e., there is a path from G to Y but not through X), the effects of G and X on Y do not themselves interact on the risk ratio scale, and suppose that the effect of the direct path from X to Y is to increase the risk of Y by a factor of λ on the risk ratio scale. Under these assumptions, we can obtain a "corrected" estimate (i.e., what we would have obtained had we been able to control for G_U and had we been able to isolate the $G \rightarrow X \rightarrow Y$ path) by using $[(\phi/\lambda)\{1+(\gamma-1)\pi_1\}/\{1+(\gamma-1)\pi_0\}]^{1/\eta}$, where π_1 and π_0 are the prevalences of G_U amongst those with $G = 1$ and those with $G = 0$, respectively (Conley et al., 2012; VanderWeele and Arah, 2011; Didelez et al., 2010). Corrected confidence intervals for this expression could be obtained by bootstrapping. We could specify values of λ, γ, π_1, and π_0 based on prior studies or we could consider a range of these values and vary them in a sensitivity analysis.

Likewise for a continuous outcome Y, suppose we estimated the effect of G on Y on the difference scale to be ϕ, possibly conditional on measured baseline covariates, and that we estimated the effect of G on X on the difference scale to be η, conditional on those same covariates. The standard instrumental variable estimate for the effect of X on Y on the difference scale would be ϕ/η. Under violations of the exclusion restriction and/or in the presence of linkage disequilibrium, as in Figure 8.14, this would be biased. Suppose we assume G_U is binary and does not interact with G on the additive scale in its effects on Y and that G_U increases Y on average by a difference of γ. Suppose also that although the exclusion restriction is violated, the effects of G and X on Y do not themselves interact on the difference scale, and suppose that the effect of the direct path from X to Y is to increase the average value of Y by a difference of λ. Under these assumptions, we can obtain a "corrected" estimate (i.e., what we would have obtained had we been able to control for G_U and had we been able to isolate the $G \rightarrow X \rightarrow Y$ path so that the exclusion restriction was not violated) by using $(\phi - \gamma\,\delta - \lambda)/\eta$, where δ is the difference in the prevalences of G_U amongst those with $G = 1$ and those with $G = 0$, respectively (Conley et al., 2012; VanderWeele and Arah, 2011). Corrected confidence intervals for this expression could be obtained by bootstrapping. If G_U is specified as continuous rather than binary, then γ can be modified to the effect of a one unit increase in G_U on Y, and δ can be modified to the difference in means of G_U amongst those with $G = 1$ versus $G = 0$. We could again specify values of γ and δ based on prior studies or we could consider a range of these values and vary them in a sensitivity analysis.

A.9. AN INTRODUCTION TO INTERACTION ANALYSIS

A.9.1. Standard Error for RERI

We consider two exposures G and E.

Proposition 9.1 (VanderWeele and Knol, 2014; cf. Hosmer and Lemeshow, 1992): Suppose we fit the model

$$\text{logit}\{P(Y=1|G=g,E=e,C=c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c$$

then the variance of $\widehat{RERI}_{OR} = e^{(g_1-g_0)\widehat{\gamma}_1+(e_1-e_0)\widehat{\gamma}_2+(g_1e_1-g_0e_0)\widehat{\gamma}_3} - e^{(g_1-g_0)\widehat{\gamma}_1+(g_1-g_0)e_0\widehat{\gamma}_3} - e^{(e_1-e_0)\widehat{\gamma}_2+(e_1-e_0)g_0\widehat{\gamma}_3} + 1$ is given by

$$v_{11}K_1^2 + v_{22}K_2^2 + v_{33}K_3^2 + v_{12}K_1K_2 + v_{13}K_1K_3 + v_{23}K_2K_3$$

where v_{ij} is the covariance between $\widehat{\gamma}_i$ and $\widehat{\gamma}_j$ and

$$K_1 = (g_1-g_0)e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} - (g_1-g_0)e^{(g_1-g_0)\gamma_1+(g_1-g_0)e_0\gamma_3}$$

$$K_2 = (e_1-e_0)e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} - (e_1-e_0)e^{(e_1-e_0)\gamma_2+(e_1-e_0)g_0\gamma_3}$$

$$K_3 = (g_1e_1-g_0e_0)e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} - (g_1-g_0)e_0e^{(g_1-g_0)\gamma_1+(g_1-g_0)e_0\gamma_3} - (e_1-e_0)g_0e^{(e_1-e_0)\gamma_2+(e_1-e_0)g_0\gamma_3}.$$

Proof:
Let

$$V = \begin{pmatrix} v_{00} & v_{01} & v_{02} & v_{03} \\ v_{10} & v_{11} & v_{12} & v_{13} \\ v_{20} & v_{21} & v_{22} & v_{23} \\ v_{30} & v_{31} & v_{32} & v_{33} \end{pmatrix}$$

be the covariance matrix for the estimators $(\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3)'$ of $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$. By the delta method the variance of $\widehat{RERI}_{OR} = e^{(g_1-g_0)\widehat{\gamma}_1+(e_1-e_0)\widehat{\gamma}_2+(g_1e_1-g_0e_0)\widehat{\gamma}_3} - e^{(g_1-g_0)\widehat{\gamma}_1+(g_1-g_0)e_0\widehat{\gamma}_3} - e^{(e_1-e_0)\widehat{\gamma}_2+(e_1-e_0)g_0\widehat{\gamma}_3} + 1$ is

$$\text{Var}(\widehat{RERI}_{OR}) = \frac{\partial(RERI_{OR})'}{\partial(\gamma_0,\gamma_1,\gamma_2,\gamma_3)'} V \frac{\partial(RERI_{OR})}{\partial(\gamma_0,\gamma_1,\gamma_2,\gamma_3)'}$$

We have that

$$\frac{\partial(RERI_{OR})}{\partial(\gamma_0,\gamma_1,\gamma_2,\gamma_3)'} = \begin{pmatrix} 0 \\ (g_1-g_0)e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} \\ -(g_1-g_0)e^{(g_1-g_0)\gamma_1+(g_1-g_0)e_0\gamma_3} \\ (e_1-e_0)e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} \\ -(e_1-e_0)e^{(e_1-e_0)\gamma_2+(e_1-e_0)g_0\gamma_3} \\ (g_1e_1-g_0e_0)e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} \\ -(g_1-g_0)e_0e^{(g_1-g_0)\gamma_1+(g_1-g_0)e_0\gamma_3} \\ -(e_1-e_0)g_0e^{(e_1-e_0)\gamma_2+(e_1-e_0)g_0\gamma_3} \end{pmatrix}$$

Let K_1, K_2, and K_3 denote the first, second, and third nonzero expressions in this vector. We then have

$$\begin{aligned}
\text{Var}(\widehat{RERI}_{OR}) &= \frac{\partial(RERI_{OR})'}{\partial(\gamma_0,\gamma_1,\gamma_2,\gamma_3)'} V \frac{\partial(RERI_{OR})}{\partial(\gamma_0,\gamma_1,\gamma_2,\gamma_3)'} \\
&= \begin{pmatrix} 0 \\ K_1 \\ K_2 \\ K_3 \end{pmatrix}' \begin{pmatrix} v_{00} & v_{01} & v_{02} & v_{03} \\ v_{10} & v_{11} & v_{12} & v_{13} \\ v_{20} & v_{21} & v_{22} & v_{23} \\ v_{30} & v_{31} & v_{32} & v_{33} \end{pmatrix} \begin{pmatrix} 0 \\ K_1 \\ K_2 \\ K_3 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ K_1 \\ K_2 \\ K_3 \end{pmatrix}' \begin{pmatrix} v_{01}K_1 + v_{02}K_2 + v_{03}K_3 \\ v_{11}K_1 + v_{12}K_2 + v_{13}K_3 \\ v_{21}K_1 + v_{22}K_2 + v_{23}K_3 \\ v_{31}K_1 + v_{32}K_2 + v_{33}K_3 \end{pmatrix} \\
&= v_{11}K_1^2 + v_{22}K_2^2 + v_{33}K_3^2 + v_{12}K_1K_2 + v_{13}K_1K_3 + v_{23}K_2K_3. \quad \blacksquare
\end{aligned}$$

A.9.2. Interaction Versus Effect Modification

We will let G and E denote two exposures and let Y be an outcome. Let Y_e denote the counterfactual outcome for Y for an individual if E were set to e, and let Y_{ge} denote the counterfactual outcome for an individual if G were set to g and E were set to e. We might say that there is effect modification or effect heterogeneity on the difference scale for the effect of E on Y across strata of G if, for two values of E, e_1, and e_0 say, and two values of G, g_1, and g_0, we have that

$$\mathbb{E}[Y_{e_1} - Y_{e_0}|g_1] - \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0] \neq 0$$

so that the effect of E on Y varies across strata of G.

We might say that there is causal interaction between the effects of G and E on Y if for two values of E, e_1, and e_0 say, and two values of G, g_1 and g_0, we have that

$$\{\mathbb{E}[Y_{g_1e_1}] - \mathbb{E}[Y_{g_1e_0}]\} - \{\mathbb{E}[Y_{g_0e_1}] - \mathbb{E}[Y_{g_0e_0}]\} \neq 0$$

so that the effect of E on Y varies depending on whether we intervene to fix G to g_1 or intervene to fix G to g_0. Effect modification involves intervention on one exposure; causal interaction involves interventions on both exposures. We could likewise define these measures conditional on covariates $C = c$ or on different scales—for example, odds ratio or risk ratio scales (cf. VanderWeele, 2009c). As discussed in the text, for a particular scale, there can be effect modification without there being causal interaction, and there can also be causal interaction without there being effect modification. See VanderWeele (2009c) for numerical examples.

Since effect modification involves intervention on one exposure and causal interaction involves intervention on both exposures, the confounding control requirements likewise differ for effect modification and causal interaction. Essentially,

since effect modification involves intervention on only one exposure, the confounding variables for only one exposure need to be controlled for. Since causal interaction involves intervention on two exposures, confounding control needs to be made for both of the exposures. More formally, the effect modification measure, $\mathbb{E}[Y_{e_1} - Y_{e_0}|g_1] - \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0]$, will be identified with covariates C, if the effect of E on Y is unconfounded conditional on (C, G)—that is, if $Y_e \perp\!\!\!\perp E|(C, G)$. The causal interaction measure, $\{\mathbb{E}[Y_{g_1e_1}] - \mathbb{E}[Y_{g_1e_0}]\} - \{\mathbb{E}[Y_{g_0e_1}] - \mathbb{E}[Y_{g_0e_0}]\}$, will be identified with covariates C, if the effects of G and E jointly are unconfounded conditional on C—that is, if $Y_{ge} \perp\!\!\!\perp (G, E)|C$. There can be cases in which the effect modification measure is identified from observed data but the causal interaction measure is not, and there can be other cases in which the causal interaction measure is identified from observed data but the effect modification measure is not. See VanderWeele (2009c) for examples.

A.9.3. Attributing Effects to Interactions

We will let G and E denote two exposures of interest that may be binary, continuous, or categorical and let Y be an outcome of interest that may be binary or continuous. Let Y_g denote the counterfactual outcome for an individual if G were set to g, let Y_e denote the counterfactual outcome for an individual if E were set to e, and let Y_{ge} denote the counterfactual outcome for an individual if G were set to g and E were set to e. We will say that the effect of G on Y is unconfounded conditional on C if $Y_g \perp\!\!\!\perp G|C$. We will say that the effect of E on Y is unconfounded conditional on C if $Y_e \perp\!\!\!\perp E|C$. We will say the joint effects of G and E on Y are unconfounded conditional on C if $Y_{ge} \perp\!\!\!\perp (G, E)|C$.

Proposition 9.2 (VanderWeele and Tchetgen Tchetgen, 2014): For any two levels e_1 and e_0 of E and any level g_0 of G, we have the decomposition

$$\mathbb{E}[Y_{e_1} - Y_{e_0}|c] = \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c] + \int \{\mathbb{E}[Y_{e_1} - Y_{e_0}|g, c] - \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c]\}\, dP(g)$$

Proof:
We have

$$\begin{aligned}
&\mathbb{E}[Y_{e_1} - Y_{e_0}|c] \\
&\quad = \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c] + \mathbb{E}[Y_{e_1} - Y_{e_0}|c] - \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c] \\
&\quad = \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c] + \int \{\mathbb{E}[Y_{e_1} - Y_{e_0}|g, c] - \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c]\}\, dP(g). \quad \blacksquare
\end{aligned}$$

In Proposition 9.2, we can decompose a total effect, $\mathbb{E}[Y_{e_1} - Y_{e_0}|c]$, into an effect conditional on $G = g_0$, namely, $\mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c]$, and a component that is a summary measure of effect modification, $\int \{\mathbb{E}[Y_{e_1} - Y_{e_0}|g, c] - \mathbb{E}[Y_{e_1} - Y_{e_0}|g_0, c]\} dP(g)$. Note that this decomposition will vary for different values of $G = g_0$. The decomposition here is given at the counterfactual level and is a decomposition of a total effect into an effect conditional on G and a measure of effect

modification. Under assumptions about independence and confounding we can identify each component of the decomposition.

Proposition 9.3 (VanderWeele and Tchetgen Tchetgen, 2014): Suppose that the effect of E on Y is unconfounded conditional on C and on (C, G), then

$$\mathbb{E}[Y_{e_1} - Y_{e_0}|c] = \mathbb{E}[Y|e_1, c] - \mathbb{E}[Y|e_0, c]$$
$$\mathbb{E}[Y_{e_1} - Y_{e_0}|g, c] = \mathbb{E}[Y|g, e_1, c] - \mathbb{E}[Y|g, e_0, c]$$

and we can thus identify the components in Proposition 9.2 and the decomposition can be written in terms of observed data as

$$\mathbb{E}[Y|e_1, c] - \mathbb{E}[Y|e_0, c] = \mathbb{E}[Y|g_0, e_1, c] - \mathbb{E}[Y|g_0, e_0, c] + \int \{\mathbb{E}[Y|g, e_1, c] - \mathbb{E}[Y|g, e_0, c] - \mathbb{E}[Y|g_0, e_1, c] + \mathbb{E}[Y|g_0, e_0, c]\} dP(g)$$

If, moreover, the joint effects of G and E are unconfounded conditional on C, then we can write the decomposition as

$$\mathbb{E}[Y_{e_1} - Y_{e_0}|c] = \mathbb{E}[Y_{g_0 e_1}|c] - \mathbb{E}[Y_{g_0 e_0}|c] + \int \{\mathbb{E}[Y_{g e_1}|c] - \mathbb{E}[Y_{g e_0}|c] - \mathbb{E}[Y_{g_0 e_1}|c] - \mathbb{E}[Y_{g_0 e_0}|c]\} dP(g).$$

Proof:
If the effect of E on Y is unconfounded conditional on C, then $[Y_{e_1} - Y_{e_0}|c] = \mathbb{E}[Y|e_1, c] - \mathbb{E}[Y|e_0, c]$. If the effect of E on Y is unconfounded conditional on (C, G), then we have $\mathbb{E}[Y_{e_1} - Y_{e_0}|g, c] = \mathbb{E}[Y|g, e_1, c] - \mathbb{E}[Y|g, e_0, c]$. If the joint effects of G and E are unconfounded conditional on C, then we have $\mathbb{E}[Y|g, e, c] = \mathbb{E}[Y_{ge}|c]$ and thus

$$\mathbb{E}[Y_{e_1} - Y_{e_0}|c] = \mathbb{E}[Y_{g_0 e_1}|c] - \mathbb{E}[Y_{g_0 e_0}|c] + \int \{\mathbb{E}[Y_{g e_1}|c] - \mathbb{E}[Y_{g e_0}|c] - \mathbb{E}[Y_{g_0 e_1}|c] - \mathbb{E}[Y_{g_0 e_0}|c]\} dP(g). \quad \blacksquare$$

If no covariates are necessary for confounding control and we let $p_{ge} = P(Y = 1|G = g, E = e)$, $p_g = P(Y = 1|G = g)$, and $p_e = P(Y = 1|E = e)$, then the first decomposition in Proposition 9.3 written in terms of the observed data simplifies to

$$(p_{e=1} - p_{e=0}) = (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(G = 1)$$

and the second decomposition written in terms of counterfactuals simplifies to

$$\mathbb{E}[Y_{e=1} - Y_{e=0}] = \mathbb{E}[Y_{01} - Y_{00}|c] + \mathbb{E}[Y_{11} - Y_{10} - Y_{01} + Y_{00}]P(G = 1)$$

On the risk ratio scale, we let $RR_{g=1} = \frac{p_{g=1}}{p_{g=0}} = \frac{P(Y=1|G=1)}{P(Y=1|G=0)}$, $RR_{e=1} = \frac{p_{e=1}}{p_{e=0}} = \frac{P(Y=1|E=1)}{P(Y=1|E=0)}$, and $RR_{ge} = \frac{p_{ge}}{p_{00}} = \frac{P(Y=1|G=g,E=e)}{P(Y=1|G=0,E=0)}$. The decomposition $(p_{e=1} - p_{e=0}) = (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(G = 1)$, when divided by $p_{e=0}$, is

$$(RR_{e=1} - 1) = \kappa RR_{01} + \kappa(RR_{11} - RR_{10} - RR_{01} + 1)P(G = 1)$$

where κ is a scaling factor given by $\kappa = \frac{p_{00}}{p_{e=0}}$. The proportion of the effect of E attributable to interaction is given by

$$PAI_{G=0}(E) \approx \frac{(e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1)P(G=1)}{e^{\gamma_2} + (e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1)P(G=1)}$$

As noted in the text, if we use the logistic regression model

$$\text{logit}\{P(Y=1|G=g,E=e,C=c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c$$

the proportion attributed to interaction can be approximated by $PAI_{G=0}(E) \approx \frac{(e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1)P(G=1)}{e^{\gamma_2} + (e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1)P(G=1)}$.

Note that, by symmetry, from Proposition 9.2, we have the decomposition

$$\mathbb{E}[Y_{g_1} - Y_{g_0}|c] = \mathbb{E}[Y_{g_1} - Y_{g_0}|e_0,c] + \int \{\mathbb{E}[Y_{g_1} - Y_{g_0}|e,c] - \mathbb{E}[Y_{g_1} - Y_{g_0}|e_0,c]\}\, dP(e)$$

This decomposition applies even if G affects E. If G and E were independent so that G did not affect E, then we would have an analogue of Proposition 2 which would be that if the effect of G on Y is unconfounded conditional on C and on (C,E), then we have $\mathbb{E}[Y_{g_1} - Y_{g_0}|c] = \mathbb{E}[Y|g_1,c] - \mathbb{E}[Y|g_0,c]$ and $\mathbb{E}[Y_{g_1} - Y_{g_0}|e,c] = \mathbb{E}[Y|g_1,e,c] - \mathbb{E}[Y|g_0,e,c]$ and we can thus write the decomposition of the total effect of G in terms of observed data as

$$\begin{aligned}\mathbb{E}[Y|g_1,c] - \mathbb{E}[Y|g_0,c] &= \mathbb{E}[Y|g_1,e_0,c] - \mathbb{E}[Y|g_0,e_0,c] \\ &+ \int \{\mathbb{E}[Y|g_1,e,c] - \mathbb{E}[Y|g_0,e,c] - \mathbb{E}[Y|g_1,e_0,c] + \mathbb{E}[Y|g_0,e_0,c]\} dP(e)\end{aligned}$$

If, moreover, the joint effects of G and E are unconfounded conditional on C, then we can write the decomposition as

$$\mathbb{E}[Y_{g_1} - Y_{g_0}|c] = \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c] + \int \{\mathbb{E}[Y_{g_1 e} - Y_{g_0 e}|c] - \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c]\} dP(e)$$

If G affects E, then we can still thus empirically decompose the total effect of E on Y into a conditional effect and the portion due to interaction. However, if G affects E, then the analogue of Proposition 9.3 for G will not apply. We still have the decomposition analogous to that in Proposition 9.2:

$$\mathbb{E}[Y_{g_1} - Y_{g_0}|c] = \mathbb{E}[Y_{g_1} - Y_{g_0}|e_0,c] + \int \{\mathbb{E}[Y_{g_1} - Y_{g_0}|e,c] - \mathbb{E}[Y_{g_1} - Y_{g_0}|e_0,c]\} dP(e)$$

However, the counterfactuals of the form $\mathbb{E}[Y_{g_1} - Y_{g_0}|e_0,c]$ will not be identified and so we cannot empirically estimate the various parts of the decomposition. This is because when G affects E, the analogue of Proposition 9.3 for G would require that the effect of G on Y be unconfounded on (C,E), and this fails because G itself affects E.

Even if G affects E, we would still have the following decomposition:

$$\int \mathbb{E}[Y_{g_1 e} - Y_{g_0 e}|c]\, dP(e) = \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c] + \int \{\mathbb{E}[Y_{g_1 e} - Y_{g_0 e}|c] - \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c]\}\, dP(e)$$

and if the joint effect of G and E are unconfounded conditional on C, all parts of the decomposition are identified from the observed data and we have

$$\int \{\mathbb{E}[Y|g_1, e, c] - \mathbb{E}[Y|g_0, e, c] dP(e) = \mathbb{E}[Y|g_1, e_0, c] - \mathbb{E}[Y|g_0, e_0, c] + \int \{\mathbb{E}[Y|g_1, e, c] - \mathbb{E}[Y|g_0, e, c] - \mathbb{E}[Y|g_1, e_0, c] + \mathbb{E}[Y|g_0, e_0, c]\} dP(e)$$

Note that the two components, $\mathbb{E}[Y|g_1, e_0, c] - \mathbb{E}[Y|g_0, e_0, c]$ and $\int \{\mathbb{E}[Y|g_1, e, c] - \mathbb{E}[Y|g_0, e, c] - \mathbb{E}[Y|g_1, e_0, c] + \mathbb{E}[Y|g_0, e_0, c]\}\, dP(e)$, can be estimated from the observed data, and these two components are exactly the same as components on the right-hand side of the decomposition considered above. However, now their sum is not equal to the total effect $\mathbb{E}[Y_{g_1} - Y_{g_0}|c]$.

Alternatively, if the effect of G on Y is unconfounded conditional on C and the joint effects of G and E are unconfounded conditional on C, then we could instead consider how much of the total effect of G on Y, $\mathbb{E}[Y_{g_1} - Y_{g_0}|c]$, is eliminated by setting E to e_0; that is, we could consider the difference $\mathbb{E}[Y_{g_1} - Y_{g_0}|c] - \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c]$. This proportion-eliminated measure (Robins and Greenland, 1992; VanderWeele, 2013a) may still be of substantial interest even if G affects E because it is a measure of how much of the effect of G could be removed by setting E to e_0. It is in fact what is estimated by the portion due to interaction in the decompositions above when G does not affect E. However, in contrast to the setting when G does not affect E, here, when G does affect E, this difference, $\mathbb{E}[Y_{g_1} - Y_{g_0}|c] - \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c]$, is no longer equal to a measure of interaction. When G affects E, the difference $\mathbb{E}[Y_{g_1} - Y_{g_0}|c] - \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c]$ may be nonzero even if there is no interaction on the additive scale between G and E. In fact if there is no interaction between G and E at the individual level so that the distribution of $Y_{g_1 e_0} - Y_{g_0 e_0}$ does not vary with e_0, then the difference $\mathbb{E}[Y_{g_1} - Y_{g_0}|c] - \mathbb{E}[Y_{g_1 e_0} - Y_{g_0 e_0}|c]$ will be equal to a mediated effect (Robins and Greenland, 1992; VanderWeele, 2013a). Because G affects E it can be nonzero even when there is no interaction, but again, even in this setting, the contrast may still be of interest from a policy perspective.

In summary then, when G affects E, we can still use our decomposition for the second exposure E of a total effect into a conditional effect when $G = 0$ and a component due to interaction. However, the analogous decomposition for G no longer works when G affects E. We have the option of either (i) retaining the same two empirical components as the condition effect and the component due to interaction, but their sum is no longer equal to a total effect, or (ii) proceeding with estimating the difference between the total effect of G and the effect of G when

setting E to some level e_0, but this difference then is no longer simply equal to a measure of interaction. See Chapter 14 for further discussion on the relations between mediation and interaction.

A.10. MECHANISTIC INTERACTION

A.10.1. Sufficient Cause and Epistatic Interactions

Consider a setting with two binary causes, G and E, and a binary outcome Y and covariates C. Let $p_{gec} = P(Y = 1|G = g, E = e, C = c)$. Consider now the potential responses ("counterfactual outcomes") for individuals under different combinations of the exposures so that Y_{ge} denotes the potential outcome that would have occurred had G been set to g and E to e. With binary G and E, because there are four possible exposure combinations, each individual has four potential outcomes: $(Y_{11}, Y_{10}, Y_{01}, Y_{00})$. The four potential outcomes defined what was called an individual's "response type"; because there were four different potential outcomes, there were $2^4 = 16$ different possible response types as in Table 10.4. We will say that G has a positive monotonic effect on Y if Y_{ge} is nondecreasing in g (and that G has a negative monotonic on Y if Y_{ge} is nonincreasing in g). We will say that E has a positive monotonic effect on Y if Y_{ge} is nondecreasing in e (and that E has a negative monotonic on Y if Y_{ge} is nonincreasing in e).

We will say that there is a sufficient cause interaction if there are individuals for whom $Y_{11} = 1$ but $Y_{10} = Y_{01} = 0$. It is possible to empirically test for sufficient cause interactions. This was shown under monotonicity assumptions by Rothman and Greenland (1998), but it was thought not possible without monotonicity assumptions. It was shown that testing was possible without monotonicity assumptions in VanderWeele and Robins (2007b, 2008).

Proposition 10.1 (VanderWeele and Robins, 2007b, 2008): If the effects of G and E on Y are unconfounded conditional on C—that is, if $Y_{ge} \perp\!\!\!\perp (G, E)|C$—then if for some c

$$p_{11c} - p_{10c} - p_{01c} > 0$$

there is a sufficient cause interaction. If G and E have positive monotonic effects on Y and the effects of G and E on Y are unconfounded conditional on C, then if for some c

$$p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$$

there is a sufficient cause interaction.

Proof:
If there is no sufficient cause interaction then $0 \geq Y_{11} - Y_{10} - Y_{01}$ for all individuals. We must thus have $0 \geq \mathbb{E}[Y_{11} - Y_{10} - Y_{01}|c]$ and since $Y_{ge} \perp\!\!\!\perp (G, E)|C$ we have

$$0 \geq \mathbb{E}[Y_{11}|c] - \mathbb{E}[Y_{10}|c] - \mathbb{E}[Y_{01}|c]$$
$$0 \geq \mathbb{E}[Y_{11}|G = 1, E = 1, c] - \mathbb{E}[Y_{10}|G = 1, E = 0, c]$$

$$
\begin{aligned}
&\quad - \mathbb{E}[Y_{01}|G=0,E=1,c] \\
0 &\geq \mathbb{E}[Y|G=1,E=1,c] - \mathbb{E}[Y|G=1,E=0,c] \\
&\quad - \mathbb{E}[Y|G=0,E=1,c] \\
0 &\geq p_{11c} - p_{10c} - p_{01c}
\end{aligned}
$$

Thus if $p_{11c} - p_{10c} - p_{01c} > 0$, then we must have $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$ for some individual.

Now consider the setting in which G and E have positive monotonic effects on Y. Under these monotonicity assumptions if there is an individual for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$, then it is also the case for that individual that $Y_{00} = 0$. Suppose there were no individual for whom $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$; then whenever $Y_{11} = 1$ we must have that either $Y_{10} = 1$ or $Y_{01} = 1$. We also have that $Y_{10} \geq Y_{00}$ and $Y_{01} \geq Y_{00}$. Thus if for some individual it is not the case that $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$, then we must have that $Y_{11} - Y_{10} - Y_{01} + Y_{00} \leq 0$. So if there is no sufficient cause interaction, we must have $Y_{11} - Y_{10} - Y_{01} + Y_{00} \leq 0$ for all individuals and thus

$$
\begin{aligned}
0 &\geq \mathbb{E}[Y_{11} - Y_{10} - Y_{01} + Y_{00}|c] \\
0 &\geq \mathbb{E}[Y_{11}|c] - \mathbb{E}[Y_{10}|c] - \mathbb{E}[Y_{01}|c] + \mathbb{E}[Y_{00}|c] \\
0 &\geq \mathbb{E}[Y_{11}|G=1,E=1,c] - \mathbb{E}[Y_{10}|G=1,E=0,c] \\
&\quad - \mathbb{E}[Y_{01}|G=0,E=1,c] + \mathbb{E}[Y_{00}|G=0,E=0,c] \\
0 &\geq \mathbb{E}[Y|G=1,E=1,c] - \mathbb{E}[Y|G=1,E=0,c] \\
&\quad - \mathbb{E}[Y|G=0,E=1,c] + \mathbb{E}[Y_{00}|G=0,E=0,c] \\
0 &\geq p_{11c} - p_{10c} - p_{01c} + p_{00c}
\end{aligned}
$$

Thus if $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$, then we must have $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$ for some individual. ■

Within this setting of two binary causes of interest, Greenland and Poole (1988) also noted that one could conceive of nine different possible sufficient causes A_0, A_1G, $A_2\overline{G}$, A_3E, $A_4\overline{E}$, A_5GE, $A_6\overline{G}E$, $A_7G\overline{E}$, and $A_8\overline{G}\overline{E}$, each involving the presence of G, or the absence of G (denoted by $\overline{G}$) or being unrelated to G, and likewise each involving the presence of E, or the absence of E (denoted by $\overline{E}$) or being unrelated to E, and finally each also possibly involving an additional background cause A_j. Greenland and Poole noted that the presence or absence of different sufficient causes (indicated by $U_j = 1$ or $U_j = 0$) would give rise to different response types (cf. Greenland and Brumback, 2002).

Let $\bigvee$ denote the disjunctive operator defined for binary X and W as $X \bigvee W = 1$ if and only if at least one of X or W is 1. We could potentially re-express the potential outcomes as a disjunction of the various sufficient causes. We will say that any set of variables $(A_0, \ldots, A_8)$ satisfying

$$Y_{ge} = A_0 \bigvee A_1 g \bigvee A_2(1-g) \bigvee A_3 e \bigvee A_4(1-e) \bigvee A_5 ge$$
$$\bigvee A_6(1-g)e \bigvee A_7 g(1-e) \bigvee A_8(1-g)(1-e)$$

constitutes a sufficient cause representation of the potential outcomes. In Section A.10.2 we will show that there are individuals for whom $Y_{11} = 1$ but $Y_{10} = Y_{01} = 0$ if and only if in every sufficient cause representation for the potential outcomes we have $A_5 \neq 0$ (VanderWeele and Robins, 2008).

We will say that there is an "epistatic interaction" or a "singular interaction" if there are individuals for whom $Y_{11} = 1$ but $Y_{10} = Y_{01} = Y_{00} = 0$.

Proposition 10.2 (VanderWeele, 2010c):
Suppose the effects of G and E on Y are unconfounded conditional on C, that is, $Y_{ge} \perp\!\!\!\perp (G,E)|C$. If for some c we have

$$p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$$

then there is an epistatic interaction. If at least one of G or E has a positive monotonic effect on Y, then

$$p_{11c} - p_{10c} - p_{01c} > 0$$

implies an epistatic interaction. If at both G and E have positive monotonic effect on Y, then

$$p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$$

implies an epistatic interaction.

Proof:
If there is no epistatic interaction, then $0 \geq Y_{11} - Y_{10} - Y_{01} - Y_{01}$ for all individuals. We must thus have $0 \geq \mathbb{E}[Y_{11} - Y_{10} - Y_{01} - Y_{00}|c]$ and since $Y_{ge} \perp\!\!\!\perp (G,E)|C$ we have

$$\begin{aligned}
0 &\geq \mathbb{E}[Y_{11} - Y_{10} - Y_{01} - Y_{00}|c] \\
0 &\geq \mathbb{E}[Y_{11}|c] - \mathbb{E}[Y_{10}|c] - \mathbb{E}[Y_{01}|c] - \mathbb{E}[Y_{00}|c] \\
0 &\geq \mathbb{E}[Y_{11}|G=1,E=1,c] - \mathbb{E}[Y_{10}|G=1,E=0,c] - \mathbb{E}[Y_{01}|G=0,E=1,c] \\
&\quad - \mathbb{E}[Y_{00}|G=0,E=0,c]
\end{aligned}$$

$$0 \geq \mathbb{E}[Y|G=1,E=1,c] - \mathbb{E}[Y|G=1,E=0,c] - \mathbb{E}[Y|G=0,E=1,c] - \mathbb{E}[Y_{00}|G=0,E=0,c]$$

$$0 \geq p_{11c} - p_{10c} - p_{01c} - p_{00c}$$

Thus if $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$, then we must have $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$ for some individual.

Now consider the setting in which at least one of G or E has a positive monotonic effect on Y: If $p_{11c} - p_{10c} - p_{01c} > 0$, then by Proposition 10.1 we have a sufficient cause interaction—that is, an individual with $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$, and since at least one of G or E has a positive monotonic effect on Y, we must also have for this individual that $Y_{00} = 0$.

Now consider the setting in which both G and E have positive monotonic effects on Y: If $p_{11c} - p_{10c} - p_{01c} + p_{00c}$, then by Proposition 10.1 we have a sufficient cause interaction—that is, an individual with $Y_{11} = 1$ and $Y_{10} = Y_{01} = 0$—and since G and E have positive monotonic effects on Y, we must also have for this individual that $Y_{00} = 0$. ■

The proofs of the conditions for epistatic interaction involving variables with three levels in Section 10.6 follow similar logic to that in Proposition 10.2 and are thus omitted. See VanderWeele (2010c) for details.

We can relate the conditions $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$, $p_{11c} - p_{10c} - p_{01c} > 0$, and $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$ to interactions in statistical models. Under a linear probability model

$$P(Y = 1|G = g, E = e) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg$$

we can rewrite the condition $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$ as $\alpha_3 > 0$, the condition $p_{11c} - p_{10c} - p_{01c} > 0$ as $\alpha_3 > \alpha_0$, and the condition $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$ as $\alpha_3 > 2\alpha_0$.

If we let $RR_{gec} = \frac{p_{gec}}{p_{00c}}$ denote the risk ratios and *let* $RERI_c = RR_{11c} - RR_{10c} - RR_{01c} + 1 = \frac{p_{11c}}{p_{00c}} - \frac{p_{10c}}{p_{00c}} - \frac{p_{01c}}{p_{00c}} + 1$ denote the relative excess risk due to interaction, then we can rewrite the condition $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$ as $RERI_c > 0$, the condition $p_{11c} - p_{10c} - p_{01c} > 0$ as $RERI_c > 1$, and the condition $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$ as $RERI_c > 2$.

The conditions expressed in terms of the coefficients of the linear probability model or in terms of $RERI_c$ are equivalent to the corresponding condition $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$, $p_{11c} - p_{10c} - p_{01c} > 0$, and $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$, respectively.

We can also give conditions on a multiplicative scale (i.e., using a log-linear model) that imply the respective conditions $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$, $p_{11c} - p_{10c} - p_{01c} > 0$, and $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$, as stated in the next proposition, but these will no longer be equivalent to those conditions.

Proposition 10.3 (VanderWeele, 2009d, 2010c):
Consider the log-linear model

$$\log\{P(Y = 1|G = g, E = e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg$$

Suppose $\beta_1 \geq 0$ and $\beta_2 \geq 0$, then $\beta_3 > 0$ implies $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$, $\beta_3 > \log(2)$ implies $p_{11c} - p_{10c} - p_{01c} > 0$, and $\beta_3 > \log(3)$ implies $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$.

Proof:
The condition $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$ can be written as

$$\begin{aligned} RR_{11c} - RR_{10c} - RR_{01c} + 1 &= e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1 \\ &= \tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} + \tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} \\ &\quad -e^{\beta_2} + \tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} - 1 \\ &= e^{\beta_1}(e^{\beta_2+\beta_3} - 1) - (e^{\beta_2} - 1). \end{aligned}$$

If $\beta_3 > 0$ and also the main effects β_1 and β_2 are non-negative, then $e^{\beta_2+\beta_3} > e^{\beta_2}$ and so $(e^{\beta_2+\beta_3} - 1) > (e^{\beta_2} - 1)$ and so $e^{\beta_1}(e^{\beta_2+\beta_3} - 1) > (e^{\beta_2} - 1)$ and thus $e^{\beta_1}(e^{\beta_2+\beta_3} - 1) - (e^{\beta_2} - 1) > 0$. Thus if $\beta_3 > 0$ and also the main effects β_1 and β_2 are non-negative, then we must have $RR_{11c} - RR_{10c} - RR_{01c} + 1 > 0$ and $p_{11c} - p_{10c} - p_{01c} + p_{00c} > 0$.

The condition $p_{11c} - p_{10c} - p_{01c} > 0$ can be written as

$$\begin{aligned} RR_{11c} - RR_{10c} - RR_{01c} &= e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} \\ &= \tfrac{1}{2}e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} + \tfrac{1}{2}e^{\beta_1+\beta_2+\beta_3} - e^{\beta_2} \\ &= e^{\beta_1}(\tfrac{1}{2}e^{\beta_2+\beta_3} - 1) + e^{\beta_2}(\tfrac{1}{2}e^{\beta_1+\beta_3} - 1) \end{aligned}$$

If $\beta_3 > \log(2)$ and also the main effects β_1 and β_2 are non-negative, then each of the terms, $(\frac{1}{2}e^{\beta_2+\beta_3} - 1)$ and $(\frac{1}{2}e^{\beta_1+\beta_3} - 1)$, will be positive and thus we will have that $RR_{11c} - RR_{10c} - RR_{01c} > 0$ and consequently also $p_{11c} - p_{10c} - p_{01c} > 0$.

The condition $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$ can be written as

$$\begin{aligned} RR_{11c} - RR_{10c} - RR_{01c} - 1 &= e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} - 1 \\ &= \tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} + \tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} - e^{\beta_2} + \tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} - 1 \\ &= e^{\beta_1}(\tfrac{1}{3}e^{\beta_2+\beta_3} - 1) + e^{\beta_2}(\tfrac{1}{3}e^{\beta_1+\beta_3} - 1) + (\tfrac{1}{3}e^{\beta_1+\beta_2+\beta_3} - 1) \end{aligned}$$

If $\beta_3 > \log(3)$ and also the main effects β_1 and β_2 are non-negative, then each of the terms $(\frac{1}{3}e^{\beta_2+\beta_3} - 1)$, $(\frac{1}{3}e^{\beta_1+\beta_3} - 1)$, and $(\frac{1}{3}e^{\beta_1+\beta_2+\beta_3} - 1)$ will be positive and thus we will have that $RR_{11c} - RR_{10c} - RR_{01c} - 1 > 0$ and consequently also $p_{11c} - p_{10c} - p_{01c} - p_{00c} > 0$. ■

A.10.2. Extension to n-Way Sufficient Cause Interaction

Consider s binary factors, $G_1, \ldots, G_s$, and let Y denote some binary outcome. We will let Ω denote the sample space of individuals in the population and we will use ω for a particular sample point. Let $Y_{g_1 \ldots g_s}(\omega)$ denote the counterfactual value of Y for individual ω if the cause G_j were set to the value g_j for $j = 1, \ldots, s$. Since $G_1, \ldots, G_s$ are

binary, all of the potential outcomes for an individual ω could be listed in a vector with 2^s components and this vector. A set of binary causes $G_1,\ldots,G_m$ for Y is said to constitute *a sufficient cause* for Y if for all values $g_1,\ldots,g_s$ such that $g_1\ldots g_m = 1$ we have that $Y_{g_1\ldots g_s}(\omega) = 1$ for all $\omega \in \mathbf{\Omega}$. A set of binary causes $G_1,\ldots,G_m$ is said to constitute *a minimal sufficient cause* for Y if $G_1,\ldots,G_m$ constitute a sufficient cause for Y and no proper subset of $G_1,\ldots,G_m$ also constitutes a sufficient cause for Y. A set of sufficient causes for Y, $M_1,\ldots,M_n$, each of which may be some product of binary causes of Y, is said to be *determinative* for Y if for all $\omega \in \mathbf{\Omega}$, $Y_{g_1\ldots g_s}(\omega) = 1$ if and only if $g_1,\ldots,g_s$ are such that $M_1 \bigvee M_2 \bigvee \ldots \bigvee M_n = 1$. If $M_1,\ldots,M_n$ is a determinative set of (minimal) sufficient causes for Y such that there is no proper subset of $M_1,\ldots,M_n$ that is also a determinative set of (minimal) sufficient causes for Y, then $M_1,\ldots,M_n$ is said to constitute a *nonredundant determinative set of (minimal) sufficient causes* for Y. Note that minimality makes references to the components in a particular conjunction—that each component is necessary for the conjunction to be sufficient for the outcome Y. Nonredundancy makes reference to a disjunction of conjunctions, that each conjunction is necessary for the disjunction to be determinative. The following results are all given in VanderWeele and Richardson (2012), though the proofs there make use of somewhat different (more compact and less transparent) notation and sometimes follow a slightly different structure or are given in slightly greater generality.

Proposition 10.4 (VanderWeele and Richardson, 2012):
For each possible conjunction $W_i = F_1^i \ldots F_{n_i}^i$, where each F_k^i member of the set $\{G_1,\ldots,G_s\}$ or is the complement of such a member, there exists a binary variable $A_i(\omega)$ that is a function of the potential outcome vector $Y(\omega)$ such that $Y(\omega) = \bigvee_i A_i(\omega) F_1^i(\omega) \ldots F_{n_i}^i(\omega)$ and such that $Y_{g_1\ldots g_s}(\omega) = \bigvee_i A_i(\omega) w_i(g_1,\ldots,g_s)$ where $w_i(g_1,\ldots,g_s) = 1$ if $F_1^i \ldots F_{n_i}^i = 1$ when $(G_1,\ldots,G_s) = (g_1,\ldots,g_s)$ and 0 otherwise.

Proof:
For $W_i = F_1^i \ldots F_{n_i}^i$ we construct the corresponding binary variable A_i as follows. Recall that each condition of the form $F_k^i = 1$ places a restriction on one of $G_1,\ldots,G_s$, that it be either 1 or 0. Let $A_i(\omega) = 1$ if $Y_{g_1\ldots g_s}(\omega) = 1$ whenever $g_1,\ldots,g_s$ are such that

$$F_1^i \ldots F_{n_i}^i = 1$$

and if there does not exist a j such that $Y_{g_1\ldots g_s}(\omega) = 1$ whenever $g_1,\ldots,g_s$ are such that

$$F_1^i \ldots F_{j-1}^i F_{j+1}^i \ldots F_{n_i}^i = 1$$

Otherwise, let $A_i(\omega) = 0$. We will show that

$$Y_{g_1\ldots g_s}(\omega) = \bigvee_i A_i(\omega) w_i(g_1,\ldots,g_s)$$

Consider ω and $g_1,\ldots,g_s$ such that

$$\bigvee_i A_i(\omega)w_i(g_1,\ldots,g_s) = 1$$

Then there exists an i such that $A_i(\omega)w_i(g_1,\ldots,g_s) = 1$. Since $A_i(\omega) = 1$ we have that $Y_{g_1\ldots g_s}(\omega) = 1$ whenever $g_1,\ldots,g_s$ are such that $F_1^i \ldots F_{n_i}^i = 1$; and since $w_i(g_1,\ldots,g_s) = 1$ we have $(G_1,\ldots,G_s) = (g_1,\ldots,g_s)$ implies that $F_1^i \ldots F_{n_i}^i = 1$ and thus we have that $Y_{g_1\ldots g_s}(\omega) = 1$. Now we must show that if $Y_{g_1\ldots g_s}(\omega) = 1$, then there exists an i such that $A_i(\omega)w_i(g_1,\ldots,g_s) = 1$. The potential outcome $Y_{g_1\ldots g_s}(\omega)$ is a function of $(\omega,g_1,\ldots,g_s)$. Let $(\omega^*,g_1^*,\ldots,g_s^*)$ be such that $Y_{g_1^*\ldots g_s^*}(\omega^*) = 1$. Consider the ordered set $(g_1^*,\ldots,g_s^*)$. If for any j,

$$g_1 = g_1^*,\ldots,g_{j-1} = g_{j-1}^*, g_{j+1} = g_{j+1}^*,\ldots,g_m = g_m^* \Rightarrow Y_{g_1\ldots g_s}(\omega^*) = 1$$

remove g_j^* from $(g_1^*,\ldots,g_s^*)$. Continue to remove those g_j^* from this set which are not needed to maintain the implication $Y_{g_1\ldots g_s}(\omega^*) = 1$. Suppose the set that remains is $(g_{h_1}^*,\ldots,g_{h_u}^*)$. We then have that

$$g_{h_1} = g_{h_1}^*,\ldots,g_{h_u} = g_{h_u}^* \Rightarrow Y_{g_1\ldots g_s}(\omega^*) = 1$$

and that for no j,

$$g_{h_1} = g_{h_1}^*,\ldots,g_{h_{j-1}} = g_{h_{j-1}}^*, g_{h_{j+1}} = g_{h_{j+1}}^*,\ldots,g_{h_u} = g_{h_u}^* \Rightarrow Y_{g_1\ldots g_s}(\omega^*) = 1$$

Define F_j as the indicator $F_j = 1_{(G_{h_j} = g_{h_j}^*)}$, then for some i, $W_i = F_1\ldots F_u$. Since the two conditions just given above hold, we must have that $Y_{g_1\ldots g_s}(\omega^*) = 1$ whenever $g_1,\ldots,g_s$ are such that

$$F_1\ldots F_u = 1$$

and that for no j is it the case that $Y_{g_1\ldots g_s}(\omega^*) = 1$ whenever $g_1,\ldots,g_s$ are such that

$$F_1\ldots F_{j-1}F_{j+1}\ldots F_u = 1$$

Thus $A_i(\omega^*) = 1$. Since $(G_1,\ldots,G_s) = (g_1^*,\ldots,g_s^*) \Rightarrow G_{h_1} = g_{h_1}^*,\ldots,G_{h_u} = g_{h_u}^* \Rightarrow F_1\ldots F_u = 1$ we have that $w_i(g_1^*,\ldots,g_s^*) = 1$ and so $A_i(\omega^*)w_i(g_1^*,\ldots,g_s^*) = 1$ and so $\bigvee_i A_i(\omega^*)w_i(g_1^*,\ldots,g_s^*) = 1$. We have thus shown $Y_{g_1\ldots g_s}(\omega) = \bigvee_i A_i(\omega)w_i(g_1,\ldots,g_s)$. From this it also immediately follows that $Y(\omega) = \bigvee_i A_i(\omega)F_1^i(\omega)\ldots F_{n_i}^i(\omega)$ since

$$\begin{aligned} Y(\omega) &= Y_{G_1(\omega)\ldots G_s(\omega)}(\omega) = \bigvee_i A_i(\omega)w_i\{G_1(\omega),\ldots,G_s(\omega)\} \\ &= \bigvee_i A_i(\omega)F_1^i(\omega)\ldots F_{n_i}^i(\omega). \quad \blacksquare \end{aligned}$$

Proposition 10.4 allows for the construction of variables A_i such that the A_i variables along with $G_1,\ldots,G_s$ and their complements can be used to form a determinative set of sufficient causes for Y which replicate a given set of potential outcomes. The conjunctions $A_iF_1^i\ldots F_{n_i}^i$ are sufficient for Y, and the disjunction of these conjunctions is determinative for Y. Each F_k^i in these conjunctions is a cause of Y since each F_k^i is either a member of the set $\{G_1,\ldots,G_s\}$ or is the complement of such a member. The variables A_i are logical constructs and may or may not allow for interpretation; it may not be possible to intervene on these logical constructs. The A_i variables essentially represent unmeasured or unknown factors that complete the particular sufficient cause. Although it may not be possible to intervene on A_i, we will still refer to conjunctions of the form $A_iF_1^i\ldots F_{n_i}^i$ as sufficient causes for Y. Note that the logical constructs A_i, being functions of the potential outcomes themselves, are not affected by any of the causes or interventions $G_1,\ldots,G_s$. If the counterfactual response pattern for every individual in the population is identical—that is, if the causes $G_1,\ldots,G_s$ completely determine the outcome Y—then no additional variables A_i are needed to form a determinative set of sufficient causes for Y. A determinative set of sufficient causes for Y can then be constructed simply from the set of binary causes $G_1,\ldots,G_s$ and their complements. If for some i, $A_i(\omega)=0$ for all ω, then the conjunction in which A_i appears will be suppressed from the disjunction. If for some i, $A_i(\omega)=1$ for all ω, then A_i will be suppressed from the conjunction in which it appears, but the terms involving $\{G_1,\ldots,G_s\}$ and their complements will appear in the conjunction for the sufficient cause.

A determinative set of sufficient causes will not in general be unique. More generally, any set of binary variables $A_i(\omega)$ constructed from the potential outcomes calY(ω) such that the disjunction $\bigvee_i A_iF_1^i\ldots F_{n_i}^i$ replicates the potential outcome response patterns for the entire population we will call a sufficient cause representation for Y. Suppose that $G_1,\ldots,G_s$ are binary causes of some binary outcome Y. For each possible conjunction $W_i=F_1^i\ldots F_{n_i}^i$, where each F_k^i is either a member of the set $\{G_1,\ldots,G_s\}$ or is the complement of such a member, let $A_i(\omega)$ be any binary variable that is a function of the potential outcome vector calY(ω). If $Y_{g_1\ldots g_s}(\omega)=\bigvee_i A_i(\omega)w_i(g_1,\ldots,g_s)$ where $w_i(g_1,\ldots,g_s)=1$ if $F_1^i\ldots F_{n_i}^i=1$ when $(G_1,\ldots,G_s)=(g_1,\ldots,g_s)$ and 0 otherwise, then the disjunction $\bigvee_i A_iF_1^i\ldots F_{n_i}^i$ is said to be *a sufficient cause representation for* Y. For any sufficient cause representation, $\bigvee_i A_iF_1^i\ldots F_{n_i}^i$, each conjunction $A_iF_1^i\ldots F_{n_i}^i$ is a sufficient cause for the outcome Y and the collection of conjunctions of the form $A_iF_1^i\ldots F_{n_i}^i$ constitutes a determinative set of sufficient causes for Y. If the conjunctions in a particular sufficient cause representation are minimal sufficient causes, then we will refer to the representation as a minimal sufficient cause representation. If the conjunctions in a particular sufficient cause representation are nonredundant, then we will refer to the representation as a nonredundant sufficient cause representation.

Suppose that $F_1,\ldots,F_m$ are such that each F_k is a member of the set of binary causes $\{G_1,\ldots,G_s\}$ or is the complement of such a member, then $F_1,\ldots,F_m$ is

said to exhibit *a minimal sufficient cause interaction* if in every nonredundant minimal sufficient cause representation for Y there exists within the representation a sufficient cause which contains $F_1,\ldots,F_m$ within its conjunction. Corresponding to the definition of a minimal sufficient cause interaction is that of a sufficient cause interaction which makes reference to all sufficient cause representations and not just nonredundant minimal sufficient cause representations. A conjunction of $F_1,\ldots,F_m$, where each F_k is a member of the set of binary causes $\{G_1,\ldots,G_s\}$ or is the complement of such a member is said to exhibit a *sufficient cause interaction (or to be irreducible)* if within every sufficient cause representation for Y there exists some sufficient cause that contains $F_1,\ldots,F_m$ within its conjunction.

Proposition 10.5 (VanderWeele and Richardson, 2012):
The conjunction of $F_1,\ldots,F_m$ is irreducible if and only if $F_1,\ldots,F_m$ exhibits a minimal sufficient cause interaction.

Proof:
If $F_1 \ldots F_m$ is irreducible, then within any sufficient cause representation $\bigvee_i A_i F_1^i \ldots F_{n_i}^i$ there exists some sufficient cause that contains within its conjunction $F_1,\ldots,F_m$ and so it immediately follows that in every nonredundant minimal sufficient cause representation for Y there will exist within the representation a sufficient cause that contains $F_1,\ldots,F_m$ in its conjunction. If $F_1 \ldots F_m$ is not irreducible, then there exists some representation $\bigvee_i A_i F_1^i \ldots F_{n_i}^i$ such that no sufficient cause $A_i F_1^i \ldots F_{n_i}^i$ contains within its conjunction $F_1,\ldots,F_m$. This representation $\bigvee_i A_i F_1^i \ldots F_{n_i}^i$ can be made into a nonredundant minimal sufficient cause representation by iteratively discarding the components of each conjunction $A_i F_1^i \ldots F_{n_i}^i$ which are not necessary to preserve the implication $A_i F_1^i \ldots F_{n_i}^i \Rightarrow Y = 1$ and then iteratively discarding any redundant minimal sufficient causes. Clearly no sufficient cause of this resulting nonredundant minimal sufficient causation representation will contain $F_1,\ldots,F_m$ within its conjunction. ■

Before proceeding with empirical tests for sufficient cause interactions, we will also need one additional concept, that of an intermediate variable. Some cause I of Y will be said to be *an intermediate variable* of other causes of Y, $G_1,\ldots,G_m$, if $I_{g_1 \ldots g_m}$, the counterfactual value of I intervening to set $G_1,\ldots,G_m$ to $g_1,\ldots,g_m$, is not independent of the values of $g_1,\ldots,g_m$. Thus if I is a cause of Y which is affected by other causes of Y, $G_1,\ldots,G_m$, then I may be conceived to be an intermediate variable between $G_1,\ldots,G_m$ and Y. We will also need an additional "consistency" assumption: If I is some set of variables such that no variable in I is an intermediate variable between $G_1,\ldots,G_m$ and Y, then $Y_{G_1=g_1,\ldots,G_m=g_m,I=I(\omega)}(\omega) = Y_{G_1=g_1,\ldots,G_m=g_m}(\omega)$. This consistency assumption concerning intermediate variables states that if I is not an intermediate variable between $G_1,\ldots,G_m$ and Y, then the counterfactual value of Y intervening to set $G_1,\ldots,G_m$ to $g_1,\ldots,g_m$ is the same as the counterfactual value of Y intervening to set $G_1,\ldots,G_m$ to $g_1,\ldots,g_m$ and I to the value it in fact was. This is as it should be since the assumption that I is not an intermediate variable between

$G_1,\ldots,G_m$ and Y implies that the interventions on $G_1,\ldots,G_m$ do not affect I. We can now relate sufficient cause interactions to counterfactual outcomes and give empirical tests for sufficient cause interactions.

Proposition 10.6 (VanderWeele and Richardson, 2012):
Let $G_1,\ldots,G_m$ be some subset (with the subscripts relabeled if necessary) of $G_1,\ldots,G_s$ and suppose that none of $G_{m+1},\ldots,G_s$ are intermediate variables between $G_1,\ldots,G_m$ and Y. The following two implications hold: (i) If there exists $\omega \in \Omega$ such that $Y_{g_1\ldots g_m}(\omega) = 1$ when $g_1 = \ldots = g_m = 1$ but $Y_{g_1\ldots g_m}(\omega) = 0$ for all $g_1,\ldots,g_m$ such that $\sum_{i=1}^{m} g_i = m-1$, then $G_1,\ldots,G_m$ have a sufficient cause interaction; (ii) if $G_1,\ldots,G_m$ have a sufficient cause interaction, then there exists $\omega^* \in \Omega$ and values $g^*_{m+1},\ldots,g^*_s$ of $G_{m+1},\ldots,G_s$ such that $Y_{g_1\ldots g_m g^*_{m+1}\ldots g^*_s}(\omega^*) = 1$ when $g_1 = \ldots = g_m = 1$ but such that $Y_{g_1\ldots g_m g^*_{m+1}\ldots g^*_s}(\omega^*) = 0$ for any $g_1,\ldots,g_m$ such that $\sum_{i=1}^{m} g_i = m-1$.

Proof:
Suppose that $G_1,\ldots,G_m$ do not have a minimal sufficient cause interaction, then there exists a nonredundant minimal sufficient cause representation $\bigvee_i A_i F^i_1 \ldots F^i_{n_i}$ such that there is no sufficient cause within the representation that contains $G_1,\ldots,G_m$ in its conjunction. Note that $G_{m+1},\ldots,G_s$ are the members of $\{G_1,\ldots,G_s\}$ other than $G_1,\ldots,G_m$. Consider any $\omega \in \Omega$ such that $Y_{g_1\ldots g_m}(\omega) = 0$ whenever $\sum_{i=1}^{m} g_i = m-1$. Suppose $G_{m+1}(\omega) = g_{m+1},\ldots,G_s(\omega) = g_s$. Define $J_{m+1},\ldots,J_s$ by $J_{m+1} = 1(G_{m+1} = g_{m+1}),\ldots,J_s = 1(G_s = g_s)$. For every $W_i = F^i_1 \ldots F^i_{n_i}$ for which the F^i_k's consist only of some subset of the elements of $G_1,\ldots,G_{j-1},G_{j+1},\ldots G_m,J_{m+1},\ldots,J_s$, we must have $A_i(\omega) = 0$ since for each j, $Y_{g_1\ldots g_m}(\omega) = 0$ when $g_j = 0$ and $g_i = 1$ for $i \neq j$. By assumption there was no sufficient cause W_i within the representation that included $G_1,\ldots,G_m$ in its conjunction. Thus for every $W_i = F^i_1 \ldots F^i_{n_i}$ for which the F^i_k's consist only of some subset of the elements of $G_1,\ldots,G_m,J_{m+1},\ldots,J_s$ we must have $A_i(\omega) = 0$ and so we have that $Y_{G_1=1,\ldots,G_m=1,G_{m+1}=g_{m+1},\ldots,G_s=g_s}(\omega) = 0$. Furthermore, since none of $G_{m+1},\ldots,G_s$ are intermediate variables between $G_1,\ldots,G_m$ and Y we have that

$$Y_{G_1=1,\ldots,G_m=1}(\omega) = Y_{G_1=1,\ldots,G_m=1,G_{m+1}=g_{m+1},\ldots,G_s=g_s}(\omega) = 0$$

There thus exists no $\omega \in \Omega$ such that $Y_{G_1=1,\ldots,G_m=1}(\omega) = 1$ but $Y_{g_1\ldots g_m}(\omega) = 0$ whenever $\sum_{i=1}^{m} g_i = m-1$. We now prove the second part of the proposition. Suppose that $G_1,\ldots,G_m$ do have a minimal sufficient cause interaction. Consider some sufficient cause representation $\bigvee_i A_i F^i_1 \ldots F^i_{n_i}$. Since $G_1,\ldots,G_m$ exhibit a minimal sufficient cause interaction and thus a sufficient cause interaction, it follows that there exists a sufficient cause within the representation with $G_1,\ldots,G_m$ in its conjunction. Let $A_l G_1 \ldots G_m H_1 \ldots H_u$ denote any such sufficient cause where each of $H_1,\ldots,H_u$ are members of the set $\{G_{m+1},\ldots,G_{m+u}\}$, with indices reordered if necessary, or complements of such members. Consider any ω^* for which $A_l(\omega^*) \neq 0$. For each $j \in \{1,\ldots,m\}$ let

$$K_j(\omega^*) = \{(g_{m+u+1},\ldots,g_s) : Y_{g_1\ldots g_s}(\omega^*) = 0 \quad \text{whenever } g_1,\ldots,g_s \text{ are such that } G_1 \ldots G_{j-1} G_{j+1} \ldots G_m H_1 \ldots H_u = 1\}$$

Suppose now that $\bigcap_{j=1}^{m} K_j(\omega^*)$ is empty, then there exist $K_1,\ldots,K_{s-m-u}$, each of which are members of the set $\{G_{m+u+1},\ldots,G_s\}$ or complements of such members and there exists a j such that $Y_{g_1\ldots g_s}(\omega^*) = 1$ whenever $g_1,\ldots,g_s$ are such that $G_1\ldots G_{j-1}G_{j+1}\ldots G_mH_1\ldots H_uK_1\ldots K_{s-m-u} = 1$. Let A_v be the A_i variable in the representation $\bigvee_i A_iF_1^i\ldots F_{n_i}^i$ corresponding to the conjunction $G_1\ldots G_{j-1}G_{j+1}\ldots G_mH_1\ldots H_uK_1\ldots K_{s-m-u}$. We may redefine the A_i variables as follows. Let $\widetilde{A}_v(\omega^*) = 1$ and $\widetilde{A}_v(\omega) = A_v(\omega)$ for $\omega \neq \omega^*$. Let $\widetilde{A}_l(\omega^*) = 0$ and $\widetilde{A}_l(\omega) = A_l(\omega)$ for $\omega \neq \omega^*$. Let $\widetilde{A}_i(\omega) = A_i(\omega)$ for all ω for $i \notin \{v,l\}$. Then $\bigvee_i \widetilde{A}_iF_1^i\ldots F_{n_i}^i$ also constitutes a sufficient cause representation for Y. We may follow the above reasoning for each ω for which $A_l(\omega) \neq 0$; and thus if $\bigcap_{j=1}^{m} K_j(\omega)$ is empty for all ω such that $A_l(\omega) \neq 0$, the conjunction $A_lG_1\ldots G_mH_1\ldots H_u$ can be eliminated from the representation by redefining the A_i variables as indicated above. Since $A_lG_1\ldots G_mH_1\ldots H_u$ was an arbitrary sufficient cause with $G_1,\ldots,G_m$ in its conjunction, it follows that if for every sufficient cause with $G_1,\ldots,G_m$ in its conjunction the corresponding set $\bigcap_{j=1}^{m} K_j(\omega)$ is empty for all ω, then every sufficient cause with $G_1,\ldots,G_m$ in its conjunction can be eliminated from the representation by redefining the A_i variables but then $G_1,\ldots,G_m$ cannot exhibit a sufficient cause interaction. Thus there must exist some sufficient cause $A_lG_1\ldots G_mH_1\ldots H_u$ and some ω^* such that $\bigcap_{j=1}^{m} K_j(\omega^*)$ is non-empty. Let $(g^*_{m+u+1},\ldots,g^*_s) \in \bigcap_{j=1}^{m} K_j(\omega^*)$ and let $g^*_{m+1},\ldots,g^*_u$ be the values of $g_{m+1},\ldots,g_u$ that correspond to $H_1,\ldots,H_u$, then $Y_{1\ldots 1g^*_{m+1}\ldots g^*_s}(\omega^*) = 1$ but $Y_{g_1\ldots g_mg^*_{m+1}\ldots g^*_s}(\omega^*) = 0$ whenever $\sum_{i=1}^{m} g_i = m-1$. ■

The conditions provided in Proposition 10.6 have obvious analogues if one or more of $G_1,\ldots,G_m$ are replaced with their complements. Proposition 10.6 and the one that follows below both require that none of $G_{m+1},\ldots,G_s$ be intermediate variables between $G_1,\ldots,G_m$ and Y. This condition is satisfied trivially if $G_1,\ldots,G_m$ are all of the causes of Y under consideration, that is, $G_1,\ldots,G_s$. The assumption is necessary because it might otherwise be possible that one of the causes $G_{m+1},\ldots,G_s$, say G_s, is in fact effectively a conjunction of $G_1,\ldots,G_m$ or some subset of these variables. In such a case, G_s might serve as a proxy for a minimal sufficient cause interaction term and thereby allow for a representation in which the conjunction of $G_1,\ldots,G_m$ is not present in any sufficient cause. Note that the presence of a sufficient cause interaction may depend upon the context of which other causes $G_{m+1},\ldots,G_s$ are being considered in the sufficient cause representations. However, the condition that there exists $\omega \in \Omega$ such that $Y_{g_1\ldots g_m}(\omega) = 1$ when $g_1 = \ldots = g_m = 1$ but $Y_{g_1\ldots g_m}(\omega) = 0$ for all $g_1,\ldots,g_m$ such that $\sum_{i=1}^{m} g_i = m-1$ does not make reference to $G_{m+1},\ldots,G_s$ and thus provides a condition for a sufficient cause interaction that is not dependent on the context. If this condition holds, then the conjunction $G_1\ldots G_m$ will be present in any sufficient cause representation regardless of which other causes $G_{m+1},\ldots,G_s$ are being considered in

the sufficient cause representations so long as these other causes of Y, $G_{m+1},\ldots,G_s$, are not themselves effects of $G_1,\ldots,G_m$.

Proposition 10.7 (VanderWeele and Richardson, 2012):
Let $G_1,\ldots,G_m$, be some subset (with the subscripts relabeled if necessary) of $G_1,\ldots,G_s$ and suppose that none of $G_{m+1},\ldots,G_s$ are intermediate variables between $G_1,\ldots,G_m$ and Y. Let C be any set of variables which suffices to control for the confounding of the causal effects of $G_1,\ldots,G_m$ on Y—that is, such that $Y_{g_1\ldots g_m} \perp\!\!\!\perp \{G_1,\ldots,G_m\}|C$ then if for any value c of C we have that

$$\begin{aligned}
&\mathbb{E}(Y|G_1=1,\ldots,G_m=1,C=c)\\
&\quad-\mathbb{E}(Y|G_1=0,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c)\\
&\quad-\mathbb{E}(Y|G_1=1,G_2=0,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c)\\
&\quad-\cdots-\\
&\quad-\mathbb{E}(Y|G_1=1,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=0,C=c)>0
\end{aligned}$$

then $G_1,\ldots,G_m$ have a sufficient cause interaction.

Proof:
We prove the contrapositive. Suppose there were no sufficient cause interaction between $G_1,\ldots,G_m$, then by Proposition 10.6 it would follow that there is no $\omega\in\Omega$ such that $Y_{1\ldots1}(\omega)=1$ but such that $Y_{g_1\ldots g_m}(\omega)=0$ whenever $\sum_i g_i=m-1$. From this it follows that for all $\omega\in\Omega$ we have $Y_{1\ldots1}(\omega)-Y_{01\ldots1}(\omega)-\ldots-Y_{1\ldots10}(\omega)\leq 0$ and so $E\{Y_{1\ldots1}(\omega)-Y_{01\ldots1}(\omega)-\ldots-Y_{1\ldots10}(\omega)|C\}\leq 0$. Since $Y_{g_1\ldots g_m} \perp\!\!\!\perp \{G_1,\ldots,G_m\}|C$ we have that

$$\begin{aligned}
&\mathbb{E}(Y|G_1=1,\ldots,G_m=1,C=c)\\
&\quad-\mathbb{E}(Y|G_1=0,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c)\\
&\quad-\mathbb{E}(Y|G_1=1,G_2=0,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c)\\
&\quad-\cdots-\\
&\quad-\mathbb{E}(Y|G_1=1,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=0,C=c)\\
&=E\{Y_{1\ldots1}(\omega)|G_1=1,\ldots,G_m=1,C=c\}\\
&\quad-E\{Y_{011\ldots1}(\omega)|G_1=0,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c\}\\
&\quad-E\{Y_{101\ldots1}(\omega)|G_1=1,G_2=0,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c\}\\
&\quad-\cdots-\\
&\quad-E\{Y_{1\ldots10}(\omega)|G_1=1,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=0,C=c\}\\
&=E\{Y_{1\ldots1}(\omega)-Y_{01\ldots1}(\omega)-\ldots-Y_{1\ldots10}(\omega)|C\}\leq 0
\end{aligned}$$

This completes the proof. ■

Even if it is the case that the conditions of Proposition 10.7 fail, there might still be a sufficient cause with $G_1,\ldots,G_m$ in its conjunction. The condition provided in

Proposition 10.7 is sufficient but not necessary for the presence of a sufficient cause interaction. In some instances it is also possible to empirically detect the presence of combinations of different sufficient cause interactions or counterfactual response patterns (cf. Ramsahai, 2013).

We will say that $G_1,\ldots,G_m$ have *positive monotonic effects* on Y if for all individuals ω we have $Y_{g_1\ldots g_m}(\omega) \geq Y_{g_1'\ldots g_m'}(\omega)$ whenever $g_i \geq g_i'$ for $i = 1,\ldots,m$. Similarly, we will say that $G_1,\ldots,G_m$ have *negative monotonic effects* on Y if for all individuals ω we have $Y_{g_1\ldots g_m}(\omega) \leq Y_{g_1'\ldots g_m'}(\omega)$ whenever $g_i \geq g_i'$ for $i = 1,\ldots,m$. Under monotonicity we can form more powerful tests because, as will be seen below, it is possible to add to the conditions of Proposition 10.7 several terms corresponding to counterfactual outcomes for combinations of causes fixed by what we will now define as a subordinate set. Let $\mathcal{U}(m) = \{(g_1,\ldots,g_m) \in \{0,1\}^m : \sum_{i=1}^m g_i = m-1\}$ and let $\mathcal{Q}(m) = \{(g_1,\ldots,g_m) \in \{0,1\}^m : \sum_{i=1}^m g_i = m-2\}$ then we say that $\mathcal{S}$ is a *subordinate set of order m* if $\mathcal{S}$ consists of $m-1$ members of $\mathcal{Q}$ such that for any $m-1$ distinct members of $\mathcal{U}$, $u_1,\ldots,u_{m-1}$, the members of $\mathcal{S}$ can be ordered $s_1,\ldots,s_{m-1}$ so that $s_i \leq u_i$ for $i = 1,\ldots,m-1$.

The set $\mathcal{U}(m)$ has m members and the set $\mathcal{Q}(m)$ has $\binom{m}{2}$ members. The requirement that each member $(g_1,\ldots,g_m)$ of the set $\mathcal{Q}(m)$ be such that $\sum_{i=1}^m g_i = m-2$ is simply that each member of $\mathcal{Q}(m)$ have $m-2$ g_i's with the value 1 and two g_i's with the value of 0 and the requirement that each member $(g_1,\ldots,g_m)$ of the set $\mathcal{U}(m)$ be such that $\sum_{i=1}^m g_i = m-1$ is simply that for some j, $g_j = 0$ and for $i \neq j$ we have that $g_i = 1$. There will in general be many possible subordinate sets $\mathcal{S}$ of a particular order. Although there is only one subordinate set of order 2:

$$\{(0,0)\}$$

it can be shown that there are three subordinate sets of order 3:

$$\{(1,0,0),(0,1,0)\}$$
$$\{(1,0,0),(0,0,1)\}$$
$$\{(0,1,0),(0,0,1)\}$$

Proposition 10.8 (VanderWeele and Richardson, 2012):
Let $G_1,\ldots,G_m$ be some subset (with the subscripts relabeled if necessary) of $G_1,\ldots,G_s$ and suppose that $G_1,\ldots,G_m$ have monotonic effects on Y and that none of $G_{m+1},\ldots,G_s$ are intermediate variables between $G_1,\ldots,G_m$ and Y. Let $U = \{(g_1,\ldots,g_m) \in \{0,1\}^m : \sum_{i=1}^m g_i = m-1\}$. If there exists an ω such that $Y_{G_1=1\ldots G_m=1}(\omega) - \sum_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) + \sum_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) > 0$ for some subordinate set S of order m then $G_1,\ldots,G_m$ have a sufficient cause interaction.

Proof:
Suppose that for some $\omega \in \Omega$ we have that

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) \leq 0$$

If $Y_{1\ldots1}(\omega) = 0$, then $Y_{g_1\ldots g_m}(\omega) = 0$ for all $g_1,\ldots,g_m$ since $G_1,\ldots,G_m$ have monotonic effects on Y and so

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) = 0$$

If $Y_{1\ldots1}(\omega) \neq 0$, then we must have that $Y_{1\ldots1}(\omega) = 1$ and since $Y_{1\ldots1}(\omega) - \sum_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) \leq 0$ there must exist some $(g_1',\ldots,g_m') \in \mathcal{U}$ such that $Y_{g_1'\ldots g_m'}(\omega) = 1$. Let $\mathcal{U}' = \mathcal{U}\backslash(g_1',\ldots,g_m')$. For any choice of the subordinate set $\mathcal{S}$ we have that

$$\begin{aligned} &Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) \\ &\quad = Y_{1\ldots1}(\omega) - Y_{g_1'\ldots g_m'}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}'} Y_{g_1\ldots g_m}(\omega) \\ &\qquad + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) \end{aligned}$$

Now $Y_{1\ldots1}(\omega) - Y_{g_1'\ldots g_m'}(\omega) = 1 - 1 = 0$ and furthermore

$$-\sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}'} Y_{g_1\ldots g_m}(\omega) + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) \leq 0$$

since each of the two sums has $m-1$ terms and since, because $\mathcal{S}$ is a subordinate set, each term in the sum over $\mathcal{U}'$ can be matched with a term in the sum over $\mathcal{S}$ so that, because of the assumption that $G_1,\ldots,G_m$ have monotonic effects on Y, the term in the sum over $\mathcal{U}'$ will be at least as large as the term in the sum over $\mathcal{S}$. Thus we have that

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) \leq 0$$

We have shown that if

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) \leq 0$$

then

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) \leq 0$$

for any choice of a subordinate set $\mathcal{S}$. From this it follows that if for some choice of a subordinate set $\mathcal{S}$ we have that

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) + \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{S}} Y_{g_1\ldots g_m}(\omega) > 0$$

then we must also have that

$$Y_{1\ldots1}(\omega) - \sum\nolimits_{(g_1,\ldots,g_m)\in\mathcal{U}} Y_{g_1\ldots g_m}(\omega) > 0$$

and so by Proposition 10.6, $G_1,\ldots,G_m$ have a sufficient cause interaction. ■

The proof of the following result essentially follows that of Proposition 10.7 and is thus omitted.

Proposition 10.9 (VanderWeele and Richardson, 2012):
Let $G_1,\ldots,G_m$, be some subset (with the subscripts relabeled if necessary) of $G_1,\ldots,G_s$ and suppose that $G_1,\ldots,G_m$ have monotonic effects on Y and that none of $G_{m+1},\ldots,G_s$ are intermediate variables between $G_1,\ldots,G_m$ and Y. Let C be any set of variables which suffice to control for the confounding of the causal effects of $G_1,\ldots,G_m$ on Y—that is, such that $Y_{g_1\ldots g_m} \perp\!\!\!\perp \{G_1,\ldots,G_m\}|C$ and let S be any subordinate set of order m then if for any value c of C we have that

$$\begin{aligned}
&\mathbb{E}(Y|G_1=1,\ldots,G_m=1,C=c)\\
&\quad-\mathbb{E}(Y|G_1=0,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c)\\
&\quad-\mathbb{E}(Y|G_1=1,G_2=0,G_3=1,\ldots,G_{m-1}=1,G_m=1,C=c)\\
&\quad-\ldots-\\
&\quad-\mathbb{E}(Y|G_1=1,G_2=1,G_3=1,\ldots,G_{m-1}=1,G_m=0,C=c)\\
&\quad+\sum\nolimits_{(g_1,\ldots,g_m)\in S}\mathbb{E}(Y|G_1=g_1,\ldots,G_m=g_m,C=c)>0
\end{aligned}$$

then $G_1,\ldots,G_m$ have a sufficient cause interaction.

A.10.3. Other Extensions: Sufficient Cause Interactions with Dichotomized Continuous Exposures and Under Independence of Background Causes

Suppose now that G is a binary exposure and that E is a continuous non-negative exposure. That E is non-negative is not strictly necessary; provided that E is bounded below by some value K, a non-negative variable E could be constructed simply by adding K to all values. For ease of exposition we will thus assume throughout that this has been done and that E is non-negative. Let Y_{ge} denote the counterfactual outcome Y for an individual if, possibly contrary to fact, G had been set to g and E had been set to e. We say that that there is causal interaction between G and E comparing exposure levels (g,e) and (g',e') if there is some individual such that $Y_{ge}=1$ but $Y_{g'e}=Y_{ge'}=0$; for example, with $g=1,g'=0$, we would have a causal interaction if there is some individual ω for whom the outcome Y would occur if exposure G were present and if the continuous exposure were at level e, but the outcome would not occur if either the binary exposure were removed or if the continuous exposure were moved to level e'. We will say that the effects of G and E on Y are positive monotonic if, for all ω, $Y_{ge}(\omega)$ is nondecreasing in g and e, respectively. For some arbitrary cutoff $h>0$, define $X=1(E>h)$. Note that if the effects of G and E on Y are unconfounded conditional on C, then it will also be the case that Y_{1e} is independent of (G,X) conditional on C.

If X is a dichotomization of E, then counterfactuals of the form $Y_{gx}(\omega)$ are not well-defined since X only indicates whether or not $E>h$ and an individual's outcome may in fact depend on the precise level of E, not simply whether or not it is

greater than some arbitrary cutoff, h. The next proposition gives a result on the conclusions that one can draw about causal interactions on the underlying continuous scale for E if one applies the test for sufficient cause interaction using a dichotomized exposure.

Proposition 10.10 (VanderWeele et al., 2011):
Suppose the effects of binary G and continuous exposure E on Y are unconfounded conditional on C and that the effect of E on Y is positive monotonic. Let $X = 1(E > h)$ and let $p_{gxc} = \mathbb{E}(Y|G = g, X = x, C = c)$. If

$$p_{11c} - p_{10c} - p_{01c} > 0$$

then there exists some individual ω such that $G(\omega) = 1$, $\mathbb{E}(\omega) = e > h$, and $Y_{1e} = 1$ but $Y_{0h} = Y_{10} = 0$.

Proof:
Suppose that for binary exposure G and dichotomized exposure $X = 1(E > h)$, if $p_{11c} - p_{10c} - p_{01c} > 0$, then we have

$$\begin{aligned}
0 &< p_{11c} - p_{10c} - p_{01c} \\
&= P(Y|G = 1, E > h, c) - P(Y|G = 1, E \leq h, c) - P(Y|G = 0, E > h, c) \\
&= P(Y|G = 1, E > h, c) - \int_{e \leq h} P(Y|G = 1, E = e, c) p(e|G = 1, E \leq h, c)\, de \\
&\quad - \int_{e > h} P(Y|G = 0, E = e, c) p(e|G = 0, E > h, c)\, de \\
&= P(Y|G = 1, E > h, c) - \int_{e \leq h} P(Y_{1e}|G = 1, E = e, c) p(e|G = 1, E \leq h, c)\, de \\
&\quad - \int_{e > h} P(Y_{0e}|G = 0, E = e, c) p(e|G = 0, E > h, c)\, de \\
&\leq P(Y|G = 1, E > h, c) - \int_{e \leq h} P(Y_{10}|G = 1, E = e, c) p(e|G = 1, E \leq h, c)\, de \\
&\quad - \int_{e > h} P(Y_{0h}|G = 0, E = e, c) p(e|G = 0, E > h, c)\, de \\
&= P(Y|G = 1, E > h, c) - \int_{e \leq h} P(Y_{10}|c) p(e|G = 1, E \leq h, c)\, de \\
&\quad - \int_{e > h} P(Y_{0h}|c) p(e|G = 0, E > h, c)\, de \\
&= P(Y|G = 1, E > h, c) - P(Y_{10}|c) - P(Y_{0h}|c) \\
&= P(Y|G = 1, E > h, c) - P(Y_{10}|G = 1, E > h, c) - P(Y_{0h}|G = 1, E > h, c) \\
&= P(Y - Y_{10} - Y_{0h}|G = 1, E > h, c)
\end{aligned}$$

where the third equality follows by consistency, the less-than-or-equal-to inequality follows by monotonicity, and the second to last and fourth to last equalities follow by unconfoundedness. If there were no individual ω with $G(\omega) = 1, \mathbb{E}(\omega) > h, C(\omega) = c$ and $Y(\omega) = 1$ but $Y_{10}(\omega) = Y_{0h}(\omega) = 0$, then we would have that $P(Y - Y_{10} - Y_{0h}|G = 1, E > h, c) \leq 0$. Thus if $p_{11c} - p_{10c} - p_{01c} > 0$, then there must be some individual with $G(\omega) = 1, \mathbb{E}(\omega) > h, C(\omega) = c$ and $Y(\omega) = 1$ but $Y_{10}(\omega) = Y_{0h}(\omega) = 0$. ■

See also VanderWeele et al. (2011) and Berzuini and Dawid (2012) for further results that hold when exposures have been dichotomized which apply when the distribution of the two exposures are statistically independent.

In the setting of gene–environment interaction it is often not unreasonable to assume that the distribution of two exposures are statistically independent. However, some of the early literature on sufficient cause interaction made a much stronger assumption that the background causes $(A_0, \ldots, A_8)$ in a sufficient cause representation as in Section A.10.1 were themselves independent of one another. Since these background causes are generally unknown, such an assumption cannot be assessed empirically and it is difficult to even know what it entails or to argue for it on substantive grounds. However, under this quite strong assumption, further results about sufficient cause interaction can be obtained.

Proposition 10.11:
Suppose for some variables $(A_0, \ldots, A_8)$, that are mutually independent conditional on C, we have

$$Y_{ge} = A_0 \bigvee A_1 g \bigvee A_2(1-g) \bigvee A_3 e \bigvee A_4(1-e) \bigvee A_5 ge \bigvee A_6(1-g)e \bigvee A_7 g(1-e) \bigvee A_8(1-g)(1-e)$$

and suppose the effect of G and E on Y are unconfounded conditional on C, then if

$$\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \neq 1$$

at least one of A_5, A_6, A_7, A_8 must be nonzero. If, in addition, the effects of G and E on Y are monotonic, then $\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} < 1$—or, equivalently, $(p_{11} - p_{10} - p_{01} + p_{00}) - p_{10}p_{01} + p_{11}p_{00} > 0$—implies that A_5 is nonzero.

Proof:
Define $P(A_0|C = c) = a_0^c, P(A_1|C = c) = a_1^c, P(A_2|C = c) = a_2^c, P(A_3|C = c) = a_3^c$ and $P(A_4|C = c) = a_4^c$. Suppose A_5, A_6, A_7, A_8 are all zero, then we have

$$\begin{aligned}(1 - p_{00}) &= 1 - \mathbb{E}(Y|C = c, G = 0, E = 0) \\ &= \{1 - \mathbb{E}(A_0 \bigvee A_2 \bigvee A_4|C = c)\} \\ &= \mathbb{E}(\overline{A}_0\overline{A}_2\overline{A}_4|C = c)\end{aligned}$$

$$= \mathbb{E}(\overline{A}_0|C=c)\mathbb{E}(\overline{A}_2|C=c)\mathbb{E}(\overline{A}_4|C=c)$$
$$= (1-a_0^c)(1-a_2^c)(1-a_4^c)$$

and

$$(1-p_{01}) = 1-\mathbb{E}(Y|C=c,G=0,E=1)$$
$$= \{1-\mathbb{E}(A_0 \bigvee A_2 \bigvee A_3|C=c)\}$$
$$= (1-a_0^c)(1-a_2^c)(1-a_3^c)$$

and

$$(1-p_{10}) = 1-\mathbb{E}(Y|C=c,G=1,E=0)$$
$$= \{1-\mathbb{E}(A_0 \bigvee A_1 \bigvee A_4|C=c)\}$$
$$= (1-a_0^c)(1-a_1^c)(1-a_4^c)$$

and

$$(1-p_{11}) = 1-\mathbb{E}(Y|C=c,G=1,E=1)$$
$$= \{1-\mathbb{E}(A_0 \bigvee A_1 \bigvee A_3|C=c)\}$$
$$= (1-a_0^c)(1-a_1^c)(1-a_3^c)$$

Then

$$(1-p_{11})(1-p_{00}) = (1-a_0^c)^2(1-a_1^c)(1-a_2^c)(1-a_3^c)(1-a_4^c)$$
$$= (1-p_{01})(1-p_{10})$$

Thus if $\frac{(1-p_{11})(1-p_{00})}{(1-p_{10})(1-p_{01})} \neq 1$, then at least one of A_5, A_6, A_7, A_8 must be nonzero. If the effect of G and E on Y are monotonic, we must have $A_5 \neq 0$. ■

A.10.4. Antagonism

In Section 10.9 we gave empirical conditions for antagonistic response types 2, 12, and 14. By appropriate recoding of the exposure and/or the outcome, the results in Proposition 10.2 can also be used to empirically test for response types 3,5,8,9, and 15. Specifically, Proposition 10.2 implies that one could test for individuals of type 8—that is, for whom $Y_{11} = 1$ but $Y_{10} = Y_{01} = Y_{00} = 0$—by testing $p_{11} - p_{10} - p_{01} - p_{00} > 0$. This test did not require any monotonicity assumptions or assumptions about the absence of other response types. Other conditions in Proposition 10.2 are given for this specific response type if monotonicity assumptions can be made. By appropriately recoding the exposure and/or the outcome of interest, the results of Proposition 10.2 give empirical tests for any of response types $2, 3, 5, 8, 9, 12, 14, 15$, both with and without monotonicity assumptions. Each

of these response types can thus, in some instances, be empirically detected without making any assumptions about monotonicity or about the absence of other response types.

We now demonstrate that the procedure in Section 10.9.5 in the text to test for any form of causal co-action for the presence of an outcome is valid. If $p_{AB} - p_{(1-A)B} - p_{A(1-B)} + p_{(1-A)(1-B)} > 0$ and if A and B have positive monotonic effects on Y, then Proposition 10.1 implies that there is causal co-action between A and B; if $p_{AB} - p_{(1-A)B} - p_{A(1-B)} + p_{(1-A)(1-B)} \leq 0$, then we cannot draw conclusions about causal co-action between for Y between A and B; because the ordering of the levels were chosen so that p_{AB} is the largest of $p_{AB}, p_{(1-A)B}, p_{A(1-B)}, p_{(1-A)(1-B)}$, only positive monotonic effects of A and B on Y are possible so no other form of causal co-action for Y can be detected using monotonicity assumptions. Without monotonicity, if $p_{AB} - p_{(1-A)B} - p_{A(1-B)} > 0$, then Proposition 10.1 implies that there is causal co-action for Y between A and B. If $p_{AB} - p_{(1-A)B} - p_{A(1-B)} > 0$ holds, then $p_{(1-A)(1-B)} - p_{(1-A)B} - p_{A(1-B)} > 0$ might also hold, in which case one could also empirically detect causal co-action for Y between $(1 - A)$ and $(1 - B)$. If $p_{AB} - p_{(1-A)B} - p_{A(1-B)} \leq 0$, then we cannot draw conclusions about causal co-action for Y between A and B; because the ordering of the levels were chosen so that p_{AB} is the largest of $p_{AB}, p_{(1-A)B}, p_{A(1-B)}, p_{(1-A)(1-B)}$, we cannot have $p_{(1-A)(1-B)} - p_{(1-A)B} - p_{A(1-B)} > 0$ or $p_{(1-A)B} - p_{AB} - p_{(1-A)(1-B)} > 0$ or $p_{A(1-B)} - p_{AB} - p_{(1-A)(1-B)} > 0$ and so we cannot detect any specific form of causal co-action for Y without monotonicity. The results for case–control studies follow by dividing the inequalities by $p_{(1-A)(1-B)}$ in the argument above.

We now demonstrate that the procedure in Section 10.9.6 to test for any form of causal co-action for the absence of an outcome is valid. The arguments for the results concerning causal co-action for $\overline{Y}$ using risks are analogous to those for risks in Section 10.9.5, under recoding of the outcome. For case–control studies, we would have causal co-action for $\overline{Y}$ between A and B under both factors having negative monotonic effects on Y (positive monotonic effects on $\overline{Y}$) if $(1 - p_{AB}) - (1 - p_{(1-A)B}) - (1 - p_{A(1-B)}) + (1 - p_{(1-A)(1-B)}) > 0$—that is, if $RR_{(1-A)B} + RR_{A(1-B)} > RR_{AB} + RR_{(1-A)(1-B)}$. If $RR_{(1-A)B} + RR_{A(1-B)} \leq RR_{AB} + RR_{(1-A)(1-B)}$, then we cannot draw conclusions about causal co-action between for $\overline{Y}$ between A and B; because the ordering of the levels were chosen so that p_{AB} is the smallest of $p_{AB}, p_{(1-A)B}, p_{A(1-B)}, p_{(1-A)(1-B)}$, only negative monotonic effects of A and B on Y are possible so no other form of causal co-action can be detected using monotonicity assumptions. Without monotonicity, we would have causal co-action for $\overline{Y}$ between A and B if $(1 - p_{AB}) - (1 - p_{(1-A)B}) - (1 - p_{A(1-B)}) > 0$—that is, if $p_{(1-A)B} + p_{A(1-B)} > p_{AB} + 1$, which, dividing by $p_{(1-A)(1-B)}$, is $RR_{(1-A)B} + RR_{A(1-B)} > RR_{AB} + \frac{1}{p_{(1-A)(1-B)}}$. If $RR_{(1-A)B} + RR_{A(1-B)} \leq RR_{AB} + \frac{1}{p_{(1-A)(1-B)}}$, then $(1 - p_{AB}) - (1 - p_{(1-A)B}) - (1 - p_{A(1-B)}) \leq 0$ and we cannot draw conclusion about causal co-action between for $\overline{Y}$ between A and B; because the ordering of the levels were chosen so that p_{AB} is the smallest of $p_{AB}, p_{(1-A)B}, p_{A(1-B)}, p_{(1-A)(1-B)}$, we cannot have $(1 - p_{(1-A)(1-B)}) - (1 - p_{(1-A)B}) - (1 - p_{A(1-B)}) > 0$ or $(1 - p_{(1-A)B}) - (1 - p_{(1-A)(1-B)}) -$

$(1 - p_{AB}) > 0$ or $(1 - p_{A(1-B)}) - (1 - p_{(1-A)(1-B)}) - (1 - p_{AB}) > 0$ and thus cannot detect any specific form of causal co-action for $\overline{Y}$ without monotonicity.

A.11. BIAS ANALYSIS FOR INTERACTIONS

A.11.1. Sensitivity Analysis and Robustness for Additive Interaction

We will let G and E denote our two factors or exposures of interest. These might be genetic and environmental factors, respectively, or any two exposures. We will let Y denote the outcome of interest. The two exposures and the outcome may be binary or continuous. We let Y_{ge} denote the counterfactual outcome or potential outcome for Y for each individual if possible, contrary to fact the first exposure had been set to g and the second to e. Thus, if the two exposures were both binary, then for each individual there would be four counterfactual or potential outcomes, Y_{11}, Y_{10}, Y_{01}, and Y_{00}. If the two factors were both randomized, we could consistently estimate $\mathbb{E}[Y_{11}]$ by $\mathbb{E}[Y|G = 1, E = 1]$, $\mathbb{E}[Y_{10}]$ by $\mathbb{E}[Y|G = 1, E = 0]$, and so on. In an observational study, the exposures are not randomized and estimates are potentially subject to confounding. Thus an investigator instead typically tries to collect data on a set of covariates C that suffices to control for this confounding. Essentially, within strata of C the groups with different exposure status should be comparable. More formally we use $A \perp\!\!\!\perp B|C$ to denote that A is independent of B conditional on C. We say that the effects of G and E on Y are unconfounded given C if for all g and e, $Y_{ge} \perp\!\!\!\perp \{G, E\}|C$. If the effects of G and E on Y are unconfounded given C, then we can consistently estimate $\mathbb{E}[Y_{ge}|c]$ by $\mathbb{E}[Y|G = g, E = e, C = c]$. Often the set of measured covariates C will not suffice to control for confounding. Instead we might hypothesize a set of unmeasured confounders U such that the effects of G and E on Y are unconfounded given $\{C, U\}$, that is, $Y_{ge} \perp\!\!\!\perp \{G, E\}|\{C, U\}$. If we do not have data on U, we cannot stratify on or otherwise adjust for U.

If we only have data on C and we are interested in interaction on the additive scale, then we would typically use the following measure for additive interaction

$$\mathbb{E}[Y|g_1, e_1, c] - \mathbb{E}[Y|g_0, e_1, c] - \mathbb{E}[Y|g_1, e_0, c] + \mathbb{E}[Y|g_0, e_0, c]$$

An additive interaction measure of 0 corresponds to exact additivity (i.e., no additive interaction). If the effects of G and E on Y are unconfounded given C, then this will consistently estimate the true causal interaction on the additive scale:

$$\mathbb{E}[Y_{g_1e_1} - Y_{g_0e_1} - Y_{g_1e_0} + Y_{g_0e_0}|c]$$

If, however, there are one or more unmeasured confounding variables U such that the effects of G and E on Y are unconfounded given $\{C, U\}$ but the effects are not unconfounded given only C, then the estimate will not be consistent for the causal interaction. If the effects of G and E on Y are unconfounded given $\{C, U\}$, that is, $Y_{ge} \perp\!\!\!\perp \{G, E\}|\{C, U\}$, then the true causal interaction conditional on $C = c$ above

in is equal to

$$\sum_u \{\mathbb{E}[Y|g_1,e_1,c,u] - \mathbb{E}[Y|g_0,e_1,c,u] - \mathbb{E}[Y|g_1,e_0,c,u] + \mathbb{E}[Y|g_0,e_0,c,u]\}P(u|c)$$

which we cannot estimate without data on U.

We define the bias on the additive scale, B_{add}, as the difference between the interaction estimate with the data and the true causal interaction contrast in conditional on $C = c$, that is,

$$\begin{aligned} B_{add} = {} & \mathbb{E}[Y|g_1,e_1,c] - \mathbb{E}[Y|g_0,e_1,c] - \mathbb{E}[Y|g_1,e_0,c] + \mathbb{E}[Y|g_0,e_0,c] \\ & - \mathbb{E}[Y_{g_1e_1} - Y_{g_0e_1} - Y_{g_1e_0} + Y_{g_0e_0}|c] \end{aligned}$$

The following result gives a general formula for the bias for the interaction on the additive scale, B_{add}, in terms of various sensitivity analysis parameters.

Proposition 11.1 (VanderWeele et al., 2012b):
Suppose that for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$ and for any particular reference level u' of U define $\gamma_{ij}(u) = \mathbb{E}[Y|g_i,e_j,c,u] - \mathbb{E}[Y|g_i,e_j,c,u']$. We then have that

$$\begin{aligned} B_{add} = {} & \sum_u \gamma_{11}(u)\{P(u|g_1,e_1,c) - P(u|c)\} - \sum_u \gamma_{10}(u)\{P(u|g_1,e_0,c) - P(u|c)\} \\ & - \sum_u \gamma_{01}(u)\{P(u|g_0,e_1,c) - P(u|c)\} + \sum_u \gamma_{00}(u)\{P(u|g_0,e_0,c) - P(u|c)\} \end{aligned}$$

Proof:
If for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$, then

$$\begin{aligned} \mathbb{E}[Y_{ge}|c] &= \sum_u \mathbb{E}[Y_{ge}|c,u]P(u|c) \\ &= \sum_u \mathbb{E}[Y_{ge}|g,e,c,u]P(u|c) \\ &= \sum_u \mathbb{E}[Y|g,e,c,u]P(u|c) \end{aligned}$$

where the first equality follows by the law of iterated expectations, the second by $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$, and the third by consistency. Thus, for any fixed reference value of u' of U

$$\begin{aligned} \mathbb{E}[Y|g,e,c] - \mathbb{E}[Y_{ge}|c] &= \sum_u \mathbb{E}[Y|g,e,c,u]P(u|g,e,c) - \sum_u \mathbb{E}[Y|g,e,c,u]P(u|c) \\ &= \sum_u \{\mathbb{E}[Y|g,e,c,u] - \mathbb{E}[Y|g,e,c,u']\}\{P(u|g,e,c) - P(u|c)\} \end{aligned}$$

By applying this equality for (g_1,e_1), (g_1,e_0), (g_0,e_1), and (g_0,e_0), the result for B_{add} follows. ■

The use of Proposition 11.1 requires specifying what might be interpreted as the effect of U in each strata of G and E, $\gamma_{ij}(u) = \mathbb{E}[Y|g_i,e_j,c,u] - \mathbb{E}[Y|g_i,e_j,c,u']$, and also the distribution of U in each strata of G and E, $P(u|g_i,e_j,c)$, along with the prevalence of U overall, $P(u|c)$. Each of these could be taken as sensitivity analysis

parameters. Under the simplifying assumption that U does not interact with one of the two factors on the additive scale, the expression for the bias on the additive scale, B_{add}, simplifies considerably as stated in the next corollary.

Corollary
Suppose that the effect of G and E on Y are unconfounded conditional on $\{C, U\}$. Suppose further that U is binary and that for fixed c, $\mathbb{E}[Y|g, e_1, c, U = 1] - \mathbb{E}[Y|g, e_1, c, U = 0] = \gamma_1$ and $\mathbb{E}[Y|g, e_0, c, U = 1] - \mathbb{E}[Y|g, e_0, c, U = 0] = \gamma_0$ are constant across strata of g so that G does not interact with U on the additive scale and let $\delta_1 = P(U = 1|g_1, e_1, c) - P(U = 1|g_0, e_1, c)$ and $\delta_0 = P(U = 1|g_1, e_0, c) - P(U = 1|g_0, e_0, c)$, then

$$B_{add} = \delta_1\gamma_1 - \delta_0\gamma_0$$

Proof:
If $\mathbb{E}[Y|g, e_1, c, U = 1] - \mathbb{E}[Y|g, e_1, c, U = 0] = \gamma_1$ and $\mathbb{E}[Y|g, e_0, c, U = 1] - \mathbb{E}[Y|g, e_0, c, U = 0] = \gamma_0$ are constant across strata of g, then

$$\begin{aligned}
B_{add} &= \{\mathbb{E}[Y|g_1, e_1, c, U = 1] - \mathbb{E}[Y|g_1, e_1, c, U = 0]\}\{P(U = 1|g_1, e_1, c) - P(U = 1|c)\} \\
&\quad -\{\mathbb{E}[Y|g_0, e_1, c, U = 1] - \mathbb{E}[Y|g_0, e_1, c, U = 0]\}\{P(U = 1|g_0, e_1, c) - P(U = 1|c)\} \\
&\quad -\{\mathbb{E}[Y|g_1, e_0, c, U = 1] - \mathbb{E}[Y|g_1, e_0, c, U = 0]\}\{P(U = 1|g_1, e_0, c) - P(U = 1|c)\} \\
&\quad +\{\mathbb{E}[Y|g_0, e_0, c, U = 1] - \mathbb{E}[Y|g_0, e_0, c, U = 0]\}\{P(U = 1|g_0, e_0, c) - P(U = 1|c)\} \\
&= \gamma_1\{P(U = 1|g_1, e_1, c) - P(u|c)\} - \gamma_1\{P(U = 1|g_0, e_1, c) - P(u|c)\} \\
&\quad -\gamma_0\{P(U = 1|g_1, e_0, c) - P(U = 1|c)\} + \gamma_0\{P(U = 1|g_0, e_0, c) - P(U = 1|c)\} \\
&= \gamma_1\{P(U = 1|g_1, e_1, c) - P(U = 1|g_0, e_1, c)\} - \gamma_0\{P(U = 1|g_1, e_0, c) - P(U = 1|g_0, e_0, c)\} \\
&= \gamma_1\delta_1 - \gamma_0\delta_0 \quad \blacksquare
\end{aligned}$$

Proposition 11.2 (VanderWeele et al., 2012b):
Suppose that the effects of G and E on Y are unconfounded conditional on $\{C, U\}$ and we have $G \times E$ independence in the sense that $\{E, U\} \perp\!\!\!\perp G|C$; then if U does not interact with G on the additive scale in the sense that $\mathbb{E}[Y|g, e, c, u] - \mathbb{E}[Y|g, e, c, u']$ is constant across g, we have $B_{add} = 0$.

Proof:
If there no interaction between G and U on the additive scale, then we have $\gamma_{1j}(u) = \gamma_{0j}(u)$. By Proposition 11.1 and $\{E, U\} \perp\!\!\!\perp G|C$ we have

$$\begin{aligned}
B_{add} &= \sum\nolimits_u \gamma_{11}(u)\{P(u|g_1, e_1, c) - P(u|c)\} - \sum\nolimits_u \gamma_{10}(u)\{P(u|g_1, e_0, c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \gamma_{01}(u)\{P(u|g_0, e_1, c) - P(u|c)\} + \sum\nolimits_u \gamma_{00}(u)\{P(u|g_0, e_0, c) - P(u|c)\} \\
&= \sum\nolimits_u \gamma_{11}(u)\{P(u|e_1, c) - P(u|c)\} - \sum\nolimits_u \gamma_{10}(u)\{P(u|e_0, c) - P(u|c)\} \\
&\quad - \sum\nolimits_u \gamma_{11}(u)\{P(u|e_1, c) - P(u|c)\} + \sum\nolimits_u \gamma_{10}(u)\{P(u|e_0, c) - P(u|c)\} \\
&= 0
\end{aligned}$$

This completes the proof. $\blacksquare$

Proposition 11.3 (VanderWeele et al., 2012b):
Suppose for all g and e and for binary U_1, U_2 we have $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U_1,U_2\}$ and we have $G \times E$ independence in the sense that $\{G,U_1\} \perp\!\!\!\perp E|C$ and $\{E,U_2\} \perp\!\!\!\perp G|C$; then if G does not interact with U_2 on the additive scale in the sense that $\mathbb{E}[Y|g,e,c,u_1,u_2] - \mathbb{E}[Y|g,e,c,u_1,u_2']$ does not vary with g, and if E does not interact with U_1 on the additive scale in the sense that $\mathbb{E}[Y|g,e,c,u_1,u_2] - \mathbb{E}[Y|g,e,c,u_1',u_2]$ does not vary with e, and if U_1 does not interact with U_2 on the additive scale in the sense that $\mathbb{E}[Y|g,e,c,u_1,u_2] - \mathbb{E}[Y|g,e,c,u_1,u_2']$ does not vary with u_1, we have $B_{add} = 0$.

Proof:
If we let $U = (U_1, U_2)$ and $u' = (0,0)$, then by Proposition 11.1 we have that

$$\begin{aligned} B_{add} = &\sum\nolimits_u \gamma_{11}(u)\{P(u|g_1,e_1,c) - P(u|c)\} - \sum\nolimits_u \gamma_{01}(u)\{P(u|g_0,e_1,c) - P(u|c)\} \\ &- \sum\nolimits_u \gamma_{10}(u)\{P(u|g_1,e_0,c) - P(u|c)\} + \sum\nolimits_u \gamma_{00}(u)\{P(u|g_0,e_0,c) - P(u|c)\} \end{aligned}$$

Now

$$\begin{aligned} &\sum\nolimits_u \gamma_{ij}(u)\{P(u|g_i,e_j,c) - P(u|c)\} \\ &= \gamma_{ij}(1,1)\{P(U=(1,1)|g_i,e_j,c) - P(U=(1,1)|c)\} \\ &\quad + \gamma_{ij}(1,0)\{P(U=(1,0)|g_i,e_j,c) - P(U=(1,0)|c)\} \\ &\quad + \gamma_{ij}(0,1)\{P(U=(0,1)|g_i,e_j,c) - P(U=(0,1)|c)\} \\ &= \gamma_{ij}(1,1)\{P(U_1=1|g_i,c)P(U_2=1|e_j,c) - P(U_1=1|c)P(U_2=1|c)\} \\ &\quad + \gamma_{ij}(1,0)\{P(U_1=1|g_i,c)P(U_2=0|e_j,c) - P(U_1=1|c)P(U_2=0|c)\} \\ &\quad + \gamma_{ij}(0,1)\{P(U_1=0|g_i,c)P(U_2=1|e_j,c) - P(U_1=0|c)P(U_2=1|c)\} \end{aligned}$$

Let $\tau_{ij} = \mathbb{E}[Y|g_i,e_j,c,U_1=1,U_2=1] - \mathbb{E}[Y|g_i,e_j,c,U_1=1,U_2=0]$ so that

$$\begin{aligned} \gamma_{ij}(1,1) &= \mathbb{E}[Y|g_i,e_j,c,U_1=1,U_2=1] - \mathbb{E}[Y|g_i,e_j,c,U_1=0,U_2=0] \\ &= \tau_{ij} + \gamma_{ij}(1,0) \end{aligned}$$

Summing the expression above over $i = 0,1$ and $j = 0,1$ and noting that because G and U_2 do not interact on the additive scale, $\tau_{1j} = \tau_{0j}$ and $\gamma_{1j}(0,1) = \gamma_{0j}(0,1)$ and because G and U_2 do not interact on the additive scale, $\gamma_{i1}(1,0) = \gamma_{i0}(1,0)$, we then have that

$$\begin{aligned} B_{add} = &\tau_{11}\{P(U_1=1|g_1,c)P(U_2=1|e_1,c) - P(U_1=1|g_0,c)P(U_2=1|e_1,c)\} \\ &- \tau_{10}\{P(U_1=1|g_1,c)P(U_2=1|e_1,c) - P(U_1=1|g_0,c)P(U_2=1|e_1,c)\} \\ &+ \gamma_{11}(1,0)\{P(U_1=1|g_1,c)P(U_2=1|e_1,c) - P(U_1=1|g_1,c)P(U_2=1|e_0,c)\} \\ &- \gamma_{01}(1,0)\{P(U_1=1|g_0,c)P(U_2=1|e_1,c) - P(U_1=1|g_0,c)P(U_2=1|e_0,c)\} \\ &+ \gamma_{11}(1,0)\{P(U_1=1|g_1,c)P(U_2=0|e_1,c) - P(U_1=1|g_1,c)P(U_2=0|e_0,c)\} \\ &- \gamma_{01}(1,0)\{P(U_1=1|g_0,c)P(U_2=0|e_1,c) - P(U_1=1|g_0,c)P(U_2=0|e_0,c)\} \end{aligned}$$

$$+\gamma_{11}(0,1)\{P(U_1=0|g_1,c)P(U_2=1|e_1,c)-P(U_1=0|g_0,c)P(U_2=1|e_1,c)\}$$
$$-\gamma_{10}(0,1)\{P(U_1=0|g_1,c)P(U_2=1|e_0,c)-P(U_1=0|g_0,c)P(U_2=1|e_0,c)\}$$

Furthermore, because U_1 and U_2 do not interact on the additive scale, $\tau_{1j} = \gamma_{1j}(0,1)$ we can group the first and seventh term, the second and eighth, the third and fifth, and the fourth and sixth to get

$$\begin{aligned} B_{add} &= \tau_{11}\{P(U_2=1|e_1,c)-P(U_2=1|e_1,c)\}-\tau_{10}\{P(U_2=1|e_1,c) \\ &\quad -P(U_2=1|e_1,c)\}+\gamma_{11}(1,0)\{P(U_1=1|g_1,c)-P(U_1=1|g_1,c)\} \\ &\quad -\gamma_{01}(1,0)\{P(U_1=1|g_0,c)-P(U_1=1|g_0,c)\} \\ &= 0 \end{aligned}$$

This completes the proof. ■

A.11.2. Sensitivity Analysis and Robustness for Multiplicative Interaction

On the multiplicative scale if we had data on covariates C, we would typically use the following measure for multiplicative interaction on the risk ratio scale:

$$\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_1,c]} \Big/ \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}$$

A multiplicative interaction measure of 1 corresponds to exact multiplicativity (i.e., no multiplicative interaction). If the effects of G and E on Y are unconfounded given C, then this will consistently estimate the true causal interaction on the multiplicative scale:

$$\frac{\mathbb{E}[Y_{g_1e_1}|c]}{\mathbb{E}[Y_{g_0e_1}|c]} \Big/ \frac{\mathbb{E}[Y_{g_1e_0}|c]}{\mathbb{E}[Y_{g_0e_0}|c]}$$

If, however, there are one or more unmeasured confounding variables U such that the effects of G and E on Y are unconfounded given $\{C,U\}$ (i.e., $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$) but the effects are not unconfounded given only C, then the estimate in will not be consistent for the causal interaction. We can then define the bias on the multiplicative scale as

$$B_{mult} = \left\{\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_1,c]} \Big/ \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}\right\} \Big/ \left\{\frac{\mathbb{E}[Y_{g_1e_1}|c]}{\mathbb{E}[Y_{g_0e_1}|c]} \Big/ \frac{\mathbb{E}[Y_{g_1e_0}|c]}{\mathbb{E}[Y_{g_0e_0}|c]}\right\}$$

The following result gives a general formula for the bias for the interaction on the multiplicative scale, B_{mult}, in terms of various sensitivity analysis parameters.

Proposition 11.4 (VanderWeele et al., 2012b):
Suppose that for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$ and for any particular reference

level u' of U define $\gamma_{ij}(u) = \frac{\mathbb{E}(Y|g_i,e_j,c,u)}{\mathbb{E}(Y|g_i,e_j,c,u')}$, then we have that

$$B_{mult} = \frac{\frac{\sum_u \gamma_{11}(u)P(u|g_1,e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{01}(u)P(u|g_0,e_1,c)}{\sum_u \gamma_{01}(u)P(u|c)}}{\frac{\sum_u \gamma_{10}(u)P(u|g_1,e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_0,c)}{\sum_u \gamma_{00}(u)P(u|c)}}$$

Proof:
If for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$, then as in the proof of Proposition 11.1, $\mathbb{E}[Y_{ge}|c] = \sum_u \mathbb{E}[Y|g,e,c,u]P(u|c)$ and thus for any fixed value of u' of U we have that

$$\begin{aligned}\mathbb{E}[Y|g,e,c]/\mathbb{E}[Y_{ge}|c] &= \sum\nolimits_u \mathbb{E}[Y|g,e,c,u]P(u|g,e,c) \Big/ \sum\nolimits_u \mathbb{E}[Y|g,e,c,u]P(u|c) \\ &= \sum\nolimits_u \{\mathbb{E}[Y|g,e,c,u]/\mathbb{E}[Y|g,e,c,u']\}P(u|g,e,c) \\ &\quad \Big/ \sum\nolimits_u \{\mathbb{E}[Y|g,e,c,u]/\mathbb{E}[Y|g,e,c,u']\}P(u|c)\end{aligned}$$

By applying this equality for (g_1,e_1), (g_1,e_0), (g_0,e_1) and (g_0,e_0) and taking ratios the result for B_{mult} follows. ■

Under the simplifying assumption that U does not interact on the multiplicative scale with one of the two factors, the expression for the bias on the multiplicative scale, B_{mult}, simplifies considerably as stated in the next corollary.

Corollary
Suppose that the effect of G and E on Y are unconfounded conditional on $\{C,U\}$. Suppose further that U is binary and that $\mathbb{E}[Y|g,e_1,c,U=1]/\mathbb{E}[Y|g,e_1,c,U=0] = \gamma_1$ and $\mathbb{E}[Y|g,e_0,c,U=1]/\mathbb{E}[Y|g,e_0,c,U=0] = \gamma_0$ are constant across strata of g so that G does not interact with U on the multiplicative scale, then

$$B_{mult} = \frac{1+(\gamma_1-1)P(U=1|g_1,e_1,c)}{1+(\gamma_1-1)P(U=1|g_0,e_1,c)} \Big/ \frac{1+(\gamma_0-1)P(U=1|g_1,e_0,c)}{1+(\gamma_0-1)P(U=1|g_0,e_0,c)}$$

Proof:
If U is binary and $\mathbb{E}[Y|g,e_1,c,U=1]/\mathbb{E}[Y|g,e_1,c,U=0] = \gamma_1$ and $\mathbb{E}[Y|g,e_0,c,U=1]/\mathbb{E}[Y|g,e_0,c,U=0] = \gamma_0$ are constant across strata of g, then

$$\begin{aligned}B_{mult} &= \frac{\frac{\sum_u \gamma_1(u)P(u|g_1,e_1,c)}{\sum_u \gamma_1(u)P(u|c)} \Big/ \frac{\sum_u \gamma_0(u)P(u|g_0,e_1,c)}{\sum_u \gamma_0(u)P(u|c)}}{\frac{\sum_u \gamma_1(u)P(u|g_1,e_0,c)}{\sum_u \gamma_1(u)P(u|c)} \Big/ \frac{\sum_u \gamma_0(u)P(u|g_0,e_0,c)}{\sum_u \gamma_0(u)P(u|c)}} \\ &= \frac{\sum_u \gamma_1(u)P(u|g_1,e_1,c) \Big/ \sum_u \gamma_0(u)P(u|g_0,e_1,c)}{\sum_u \gamma_1(u)P(u|g_1,e_0,c) \Big/ \sum_u \gamma_0(u)P(u|g_0,e_0,c)} \\ &= \frac{1+(\gamma_1-1)P(U=1|g_1,e_1,c)}{1+(\gamma_1-1)P(U=1|g_0,e_1,c)} \Big/ \frac{1+(\gamma_0-1)P(U=1|g_1,e_0,c)}{1+(\gamma_0-1)P(U=1|g_0,e_0,c)}.\end{aligned}$$ ■

Proposition 11.5 (VanderWeele et al., 2012b):
Suppose that the effect of G and E on Y are unconfounded conditional on $\{C,U\}$

and we have $G \times E$ independence in the sense that $\{E,U\} \perp\!\!\!\perp G|C$; then if U does not interact with G on the multiplicative scale in the sense that $\frac{\mathbb{E}(Y|g,e,c,u)}{\mathbb{E}(Y|g,e,c,u')}$ is constant across g, then $B_{mult} = 1$.

Proof:
If there is no interaction between G and U on the multiplicative scale, then we have $\gamma_{1j}(u) = \gamma_{0j}(u)$. By Proposition 11.4 and $\{E,U\} \perp\!\!\!\perp G|C$ we have

$$\begin{aligned} B_{mult} &= \frac{\frac{\sum_u \gamma_{11}(u)P(u|g_1,e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{01}(u)P(u|g_0,e_1,c)}{\sum_u \gamma_{01}(u)P(u|c)}}{\frac{\sum_u \gamma_{10}(u)P(u|g_1,e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_0,c)}{\sum_u \gamma_{00}(u)P(u|c)}} \\ &= \frac{\frac{\sum_u \gamma_{11}(u)P(u|e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{11}(u)P(u|e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)}}{\frac{\sum_u \gamma_{10}(u)P(u|e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{10}(u)P(u|e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)}} = 1 \end{aligned}$$

This completes the proof. ■

Proposition 11.6 (VanderWeele, 2012b):
Suppose for all g and e and for binary U_1, U_2 we have $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U_1,U_2\}$ and we have $G \times E$ independence in the sense that $\{G,U_1\} \perp\!\!\!\perp E|C$ and $\{E,U_2\} \perp\!\!\!\perp G|C$; then if G does not interact with U_2 on the multiplicative scale in the sense that $\mathbb{E}[Y|g,e,c,u_1,u_2]/\mathbb{E}[Y|g,e,c,u_1,u_2']$ does not vary with g, and if E does not interact with U_1 on the multiplicative scale in the sense that $\mathbb{E}[Y|g,e,c,u_1,u_2]/\mathbb{E}[Y|g,e,c,u_1',u_2]$ does not vary with e, and if U_1 does not interact with U_2 on the multiplicative scale in the sense that $\mathbb{E}[Y|g,e,c,u_1,u_2]/\mathbb{E}[Y|g,e,c,u_1,u_2']$ does not vary with u_1, then $B_{mult} = 1$.

Proof:
If we let $U = (U_1,U_2)$ and $u' = (0,0)$, then by Proposition 11.4, we have

$$B_{mult} = \frac{\frac{\sum_u \gamma_{11}(u)P(u|g_1,e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{01}(u)P(u|g_0,e_1,c)}{\sum_u \gamma_{01}(u)P(u|c)}}{\frac{\sum_u \gamma_{10}(u)P(u|g_1,e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)} \Big/ \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_0,c)}{\sum_u \gamma_{00}(u)P(u|c)}}$$

Let $\tau_{ij} = \mathbb{E}[Y|g_i,e_j,c,U_1 = 1,U_2 = 1]/\mathbb{E}[Y|g_i,e_j,c,U_1 = 1,U_2 = 0]$ so that $\gamma_{ij}(1,1) = \tau_{ij} \times \gamma_{ij}(1,0)$. We then have that

$$\begin{aligned} &\sum_u \gamma_{ij}(u)P(u|g_i,e_j,c) \Big/ \Big\{\sum_u \gamma_{ij}(u)P(u|c)\Big\} \\ &= [\gamma_{ij}(1,1)P(U = (1,1)|g_i,e_j,c) + \gamma_{ij}(1,0)P(U = (1,0)|g_i,e_j,c) \\ &\quad + \gamma_{ij}(0,1)P(U = (0,1)|g_i,e_j,c) + P(U = (0,0)|g_i,e_j,c)] \Big/ \\ &\quad [\gamma_{ij}(1,1)P(U = (1,1)|c) + \gamma_{ij}(1,0)P(U = (1,0)|c) \\ &\quad + \gamma_{ij}(0,1)P(U = (0,1)|c) + P(U = (0,0)|c)] \\ &= [\gamma_{ij}(1,1)P(U_1 = 1|g_i,c)P(U_2 = 1|e_j,c) \end{aligned}$$

$$
\begin{aligned}
&+\gamma_{ij}(1,0)P(U_1=1|g_i,c)P(U_2=0|e_j,c)\\
&+\gamma_{ij}(0,1)P(U_1=0|g_i,c)P(U_2=1|e_j,c)+P(U_1=0|g_i,c)P(U_2=0|e_j,c)]/\\
&[\gamma_{ij}(1,1)P(U_1=1|c)P(U_2=1|c)+\gamma_{ij}(1,0)P(U_1=1|c)P(U_2=0|c)\\
&+\gamma_{ij}(0,1)P(U_1=0|c)P(U_2=1|c)+P(U_1=0|c)P(U_2=0|c)]
\end{aligned}
$$

Using this in Proposition 11.4 for B_{mult} for $i=0,1$ and $j=0,1$ and noting that because G and U_2 do not interact on the multiplicative scale, $\tau_{1j}=\tau_{0j}$ and $\gamma_{1j}(0,1)=\gamma_{0j}(0,1)$ and because G and U_2 do not interact on the multiplicative scale, $\gamma_{i1}(1,0)=\gamma_{i0}(1,0)$, and because U_1 and U_2 do not interact on the multiplicative scale we have $\tau_{1j}=\gamma_{1j}(0,1)$ we then have that

$$
\begin{aligned}
B_{mult} &= \frac{\left\{\begin{array}{r}\tau_{11}\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=1|e_1,c)+\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=0|e_1,c)\\ +\gamma_{11}(0,1)P(U_1=0|g_1,c)P(U_2=1|e_1,c)+P(U_1=0|g_1,c)P(U_2=0|e_1,c)\end{array}\right\}}{\left\{\begin{array}{r}\tau_{10}\gamma_{10}(1,0)P(U_1=1|g_1,c)P(U_2=1|e_0,c)+\gamma_{10}(1,0)P(U_1=1|g_1,c)P(U_2=0|e_0,c)\\ +\gamma_{10}(0,1)P(U_1=0|g_1,c)P(U_2=1|e_0,c)+P(U_1=0|g_1,c)P(U_2=0|e_0,c)\end{array}\right\}}\\
&\quad\frac{\left\{\begin{array}{r}\tau_{01}\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=1|e_1,c)+\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=0|e_1,c)\\ +\gamma_{01}(0,1)P(U_1=0|g_0,c)P(U_2=1|e_1,c)+P(U_1=0|g_0,c)P(U_2=0|e_1,c)\end{array}\right\}}{\left\{\begin{array}{r}\tau_{00}\gamma_{00}(1,0)P(U_1=1|g_0,c)P(U_2=1|e_0,c)+\gamma_{00}(1,0)P(U_1=1|g_0,c)P(U_2=0|e_0,c)\\ +\gamma_{00}(0,1)P(U_1=0|g_0,c)P(U_2=1|e_0,c)+P(U_1=0|g_0,c)P(U_2=0|e_0,c)\end{array}\right\}}\\
&= \frac{\left\{\begin{array}{r}\gamma_{11}(0,1)\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=1|e_1,c)+\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=0|e_1,c)\\ +\gamma_{11}(0,1)P(U_1=0|g_1,c)P(U_2=1|e_1,c)+P(U_1=0|g_1,c)P(U_2=0|e_1,c)\end{array}\right\}}{\left\{\begin{array}{r}\gamma_{10}(0,1)\gamma_{10}(1,0)P(U_1=1|g_1,c)P(U_2=1|e_0,c)+\gamma_{10}(1,0)P(U_1=1|g_1,c)P(U_2=0|e_0,c)\\ +\gamma_{10}(0,1)P(U_1=0|g_1,c)P(U_2=1|e_0,c)+P(U_1=0|g_1,c)P(U_2=0|e_0,c)\end{array}\right\}}\\
&\quad\frac{\left\{\begin{array}{r}\gamma_{01}(0,1)\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=1|e_1,c)+\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=0|e_1,c)\\ +\gamma_{01}(0,1)P(U_1=0|g_0,c)P(U_2=1|e_1,c)+P(U_1=0|g_0,c)P(U_2=0|e_1,c)\end{array}\right\}}{\left\{\begin{array}{r}\gamma_{00}(0,1)\gamma_{00}(1,0)P(U_1=1|g_0,c)P(U_2=1|e_0,c)+\gamma_{00}(1,0)P(U_1=1|g_0,c)P(U_2=0|e_0,c)\\ +\gamma_{00}(0,1)P(U_1=0|g_0,c)P(U_2=1|e_0,c)+P(U_1=0|g_0,c)P(U_2=0|e_0,c)\end{array}\right\}}\\
&= \frac{\left\{\begin{array}{r}\gamma_{11}(0,1)\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=1|e_1,c)+\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=0|e_1,c)\\ +\gamma_{11}(0,1)P(U_1=0|g_1,c)P(U_2=1|e_1,c)+P(U_1=0|g_1,c)P(U_2=0|e_1,c)\end{array}\right\}}{\left\{\begin{array}{r}\gamma_{10}(0,1)\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=1|e_0,c)+\gamma_{11}(1,0)P(U_1=1|g_1,c)P(U_2=0|e_0,c)\\ +\gamma_{10}(0,1)P(U_1=0|g_1,c)P(U_2=1|e_0,c)+P(U_1=0|g_1,c)P(U_2=0|e_0,c)\end{array}\right\}}\\
&\quad\frac{\left\{\begin{array}{r}\gamma_{11}(0,1)\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=1|e_1,c)+\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=0|e_1,c)\\ +\gamma_{11}(0,1)P(U_1=0|g_0,c)P(U_2=1|e_1,c)+P(U_1=0|g_0,c)P(U_2=0|e_1,c)\end{array}\right\}}{\left\{\begin{array}{r}\gamma_{10}(0,1)\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=1|e_0,c)+\gamma_{01}(1,0)P(U_1=1|g_0,c)P(U_2=0|e_0,c)\\ +\gamma_{10}(0,1)P(U_1=0|g_0,c)P(U_2=1|e_0,c)+P(U_1=0|g_0,c)P(U_2=0|e_0,c)\end{array}\right\}}\\
&= \frac{\left\{\begin{array}{r}\times[\gamma_{11}(1,0)P(U_1=1|g_1,c)+P(U_1=0|g_1,c)]\end{array}\right\}}{\left\{\begin{array}{r}\times[\gamma_{11}(1,0)P(U_1=1|g_1,c)+P(U_1=0|g_1,c)]\end{array}\right\}}\\
&\quad\frac{}{\left\{\begin{array}{r}\times[\gamma_{11}(1,0)P(U_1=1|g_0,c)+P(U_1=0|g_0,c)]\end{array}\right\}}\\
&= 1
\end{aligned}
$$

This completes the proof. ■

A.11.3. Sensitivity Analysis for the Relative Excess Risk Due to Interaction

The relative excess risk due to interaction ($RERI$) conditional on $C = c$ is defined as

$$\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} + 1$$

If the outcome is rare so that odds ratios approximate risk ratios, then each term $\mathbb{E}[Y|g,e,c]/\mathbb{E}[Y|g_0,e_0,c]$ can be approximated by the estimated odds ratio from the logistic regression.

Define the causal $RERI$ conditional on $C = c$ by

$$RERI_c = \frac{\mathbb{E}[Y_{g_1e_1}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} - \frac{\mathbb{E}[Y_{g_1e_0}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} - \frac{\mathbb{E}[Y_{g_0e_1}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} + 1$$

If the effects of G and E on Y were unconfounded conditional on (C,U) but data were only available on C, the estimated $RERI$ with the data would in general be biased for the true causal $RERI$ conditional on $C = c$ because of the unmeasured confounding due to U.

Proposition 11.7 (VanderWeele et al., 2012b):
Suppose that for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$ and for any particular reference level u' of U define $\gamma_{ij}(u) = \frac{\mathbb{E}(Y|g_i,e_j,c,u)}{\mathbb{E}(Y|g_i,e_j,c,u')}$, then we have that

$$RERI_c = \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_0,c)}{\sum_u \gamma_{00}(u)P(u|c)} \left[\frac{\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{11}(u)P(u|g_1,e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)}} - \frac{\frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{10}(u)P(u|g_1,e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)}} - \frac{\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{01}(u)P(u|g_0,e_1,c)}{\sum_u \gamma_{01}(u)P(u|c)}} \right] + 1$$

Proof:
Using the result in the proof of Proposition 11.1, we have that

$$\frac{\mathbb{E}[Y_{g_ie_j}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} = \frac{\mathbb{E}[Y|g_i,e_j,c]}{\mathbb{E}[Y|g_0,e_0,c]} \Bigg/ \left[\frac{\sum_u \gamma_{ij}(u)P(u|g_i,e_j,c)}{\sum_u \gamma_{ij}(u)P(u|c)} \Bigg/ \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_0,c)}{\sum_u \gamma_{00}(u)P(u|c)} \right]$$

$$= \frac{\frac{\mathbb{E}[Y|g_i,e_j,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{ij}(u)P(u|g_1,e_1,c)}{\sum_u \gamma_{ij}(u)P(u|c)}} \times \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_1,c)}{\sum_u \gamma_{00}(u)P(u|c)}$$

Thus,

$$RERI_c = \frac{\mathbb{E}[Y_{g_1e_1}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} - \frac{\mathbb{E}[Y_{g_1e_0}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} - \frac{\mathbb{E}[Y_{g_0e_1}|c]}{\mathbb{E}[Y_{g_0e_0}|c]} + 1$$

$$= \frac{\sum_u \gamma_{00}(u)P(u|g_0,e_0,c)}{\sum_u \gamma_{00}(u)P(u|c)} \left[\frac{\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{11}(u)P(u|g_1,e_1,c)}{\sum_u \gamma_{11}(u)P(u|c)}} - \frac{\frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{10}(u)P(u|g_1,e_0,c)}{\sum_u \gamma_{10}(u)P(u|c)}} - \frac{\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_{01}(u)P(u|g_0,e_1,c)}{\sum_u \gamma_{01}(u)P(u|c)}} \right] + 1. \quad \blacksquare$$

To apply the result, one would again specify the effect of U, $\gamma_{ij}(u) = \frac{\mathbb{E}(Y|g_i,e_j,c,u)}{\mathbb{E}(Y|g_i,e_j,c,u')}$, in each of the $G \times E$ strata along with the distribution of U, $P(u|g_1,e_0,c)$, for each of the $G \times E$ strata. The corrected causal *RERI* could then be computed using the expression in Proposition 11.7. Under simplifying assumptions that U is binary with a constant effect across $G \times E$, a more straightforward adjustment approach is possible as stated in the following Corollary, which follows immediately from Proposition 11.7.

Corollary
Suppose that for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$ and suppose that U is binary and $\gamma = \frac{\mathbb{E}(Y|g,e,c,U=1)}{\mathbb{E}(Y|g,e,c,U=0)}$ is constant over g and e, then we have that

$$RERI_c = \frac{\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{1+(\gamma-1)P(U=1|g_1,e_1,c)}{1+(\gamma-1)P(U=1|g_0,e_0,c)}} - \frac{\frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{1+(\gamma-1)P(U=1|g_1,e_0,c)}{1+(\gamma-1)P(U=1|g_0,e_0,c)}} - \frac{\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{1+(\gamma-1)P(U=1|g_0,e_1,c)}{1+(\gamma-1)P(U=1|g_0,e_0,c)}} + 1.$$

A simpler result is also possible under $G \times E$ independence.

Proposition 11.8 (VanderWeele et al., 2012b):
Suppose that for all g and e, $Y_{ge} \perp\!\!\!\perp \{G,E\}|\{C,U\}$ and we have $G \times E$ independence in the sense that $\{E,U\} \perp\!\!\!\perp G|C$. Suppose further that U is binary and that $\mathbb{E}[Y|g,e_1,c,U=1]/\mathbb{E}[Y|g,e_1,c,U=0] = \gamma_1$and $\mathbb{E}[Y|g,e_0,c,U=1]/\mathbb{E}[Y|g,e_0,c,U=0] = \gamma_0$ are constant across strata of g so that G does not interact with U on the multiplicative scale then

$$RERI_c = \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} + 1$$

where

$$\kappa = \frac{1+(\gamma_1-1)P(U=1|e_1,c)}{1+(\gamma_1-1)P(U=1|c)} \Big/ \frac{1+(\gamma_0-1)P(U=1|e_0,c)}{1+(\gamma_0-1)P(U=1|c)}$$

Proof:
If U does not interaction with G on the multiplicative scale, then $\gamma_{10}(u) = \gamma_{00}(u) = \gamma_0(u)$ and $\gamma_{11}(u) = \gamma_{01}(u) = \gamma_1(u)$ and if we have $G \times E$ independence, then by Proposition 11.7,

$$RERI_c = \frac{\sum_u \gamma_0(u)P(u|e_0,c)}{\sum_u \gamma_0(u)P(u|c)} \left[\frac{\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_1(u)P(u|e_1,c)}{\sum_u \gamma_1(u)P(u|c)}} - \frac{\frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_0(u)P(u|e_0,c)}{\sum_u \gamma_0(u)P(u|c)}} \right.$$

$$-\frac{\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]}}{\frac{\sum_u \gamma_1(u)P(u|e_1,c)}{\sum_u \gamma_1(u)P(u|c)}}\Bigg] + 1$$

$$= \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_1,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{\mathbb{E}[Y|g_1,e_0,c]}{\mathbb{E}[Y|g_0,e_0,c]} - \frac{1}{\kappa}\frac{\mathbb{E}[Y|g_0,e_1,c]}{\mathbb{E}[Y|g_0,e_0,c]} + 1$$

where

$$\kappa = \frac{\sum_u \gamma_1(u)P(u|e_1,c)}{\sum_u \gamma_1(u)P(u|c)} \Big/ \frac{\sum_u \gamma_0(u)P(u|e_0,c)}{\sum_u \gamma_0(u)P(u|c)}$$

$$= \frac{1+(\gamma_1-1)P(U=1|e_1,c)}{1+(\gamma_1-1)P(U=1|c)} \Big/ \frac{1+(\gamma_0-1)P(U=1|e_0,c)}{1+(\gamma_0-1)P(U=1|c)}. \quad \blacksquare$$

A.11.4. Measurement Error and Additive Interaction

Let G and E denote two binary exposures of interest and let Y denote a binary outcome and let $p_{ge} = \mathbb{E}(Y|G=g,E=e)$. We will consider testing and correction for the standard measure for additive interaction:

$$p_{11} - p_{10} - p_{01} + p_{00} > 0$$

$$p_{11} - p_{10} - p_{01} > 0$$

and

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

Let G^* and E^* denote the potentially mismeasured exposures. Let $d_1 = P(G = 0|G^* = 1)$ and $d_2 = P(E = 0|E^* = 1)$ and let $u_1 = P(G = 1|G^* = 0)$ and $u_2 = P(E = 1|E^* = 0)$. These misclassification probabilities are equal to 1 minus the positive and negative predictive values, respectively, of the measured exposures for the true exposures. We say that the misclassification is nondifferential if $P(Y|G=i,E=j,G^*=l,E^*=m) = P(Y|G=i,E=j)$. Non-differential misclassification implies that the measured exposures gives no information about the outcome beyond the information it gives about the true exposures. We say that the misclassification is independent if the events of G and E being misclassified are independent—that is, that $P(G=i,E=j|G^*=l,E^*=m) = P(G=i|G^*=l)P(E=j|E^*=m)$. A sufficient condition for this is that $P(g^*,e^*|g,e) = P(g^*|g)P(e^*|e)$ with the distributions of G and E statistically independent. Let $p^*_{ge} = \mathbb{E}(Y|G^*=g,E^*=e)$. We then have the following results.

Proposition 11.9 (VanderWeele, 2012b):
If misclassification of G and E is nondifferential and independent and if $d_i + u_i < 1$, then

$$p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00} > 0$$

implies $p_{11} - p_{10} - p_{01} + p_{00} > 0$ and $(p_{11} - p_{10} - p_{01} + p_{00}) = (p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00})/\{(1-d_1-u_1)(1-d_2-u_2)\}$.

Proof:
We can rewrite $p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00}$ as follows:

$$\begin{aligned}
&(1-d_1)(1-d_2)p_{11} - (1-d_1)u_2p_{11} - u_1(1-d_2)p_{11} + u_1u_2p_{11}\\
&\quad + (1-d_1)d_2p_{10} - (1-d_1)(1-u_2)p_{10} - u_1d_2p_{10} + u_1(1-u_2)p_{10}\\
&\quad + d_1(1-d_2)p_{01} - d_1u_2p_{01} - (1-u_1)(1-d_2)p_{01} + (1-u_1)u_2p_{01}\\
&\quad + d_1d_2p_{00} - d_1(1-u_2)p_{00} - (1-u_1)d_2p_{00} + (1-u_1)(1-u_2)p_{00}\\
&\quad = (1-d_1-u_1)(1-d_2-u_2)(p_{11} - p_{10} - p_{01} + p_{00})
\end{aligned}$$

Thus if $d_i + u_i < 1$ and $p^*_{11} - p^*_{10} - p^*_{01} + p^*_{00} > 0$ then $p_{11} - p_{10} - p_{01} + p_{00} > 0$. ■

Proposition 11.10 (VanderWeele, 2012b):
If misclassification of G and E is nondifferential and independent and if $d_i < 1/2$ and $u_i < 1/4$, then

$$p^*_{11} - p^*_{10} - p^*_{01} > 0$$

implies $p_{11} - p_{10} - p_{01} > 0$.

Proof:
We can rewrite $p^*_{11} - p^*_{10} - p^*_{01}$ as follows

$$\begin{aligned}
&(1-d_1)(1-d_2)p_{11} - (1-d_1)u_2p_{11} - u_1(1-d_2)p_{11}\\
&\quad + (1-d_1)d_2p_{10} - (1-d_1)(1-u_2)p_{10} - u_1d_2p_{10}\\
&\quad + d_1(1-d_2)p_{01} - d_1u_2p_{01} - (1-u_1)(1-d_2)p_{01}\\
&\quad + d_1d_2p_{00} - d_1(1-u_2)p_{00} - (1-u_1)d_2p_{00}
\end{aligned}$$

Let $V = (1-d_1)(1-d_2) - (1-d_1)u_2 - u_1(1-d_2)$. We then rewrite $p^*_{11} - p^*_{10} - p^*_{01}$ as

$$Vp_{11} - (V+u_2)p_{10} - (V+u_1)p_{01} + \{d_1d_2 - d_1(1-u_2) - (1-u_1)d_2\}p_{00}$$

Thus if $p^*_{11} - p^*_{10} - p^*_{01} > 0$, then

$$V(p_{11} - p_{10} - p_{01}) > u_2p_{10} + u_1p_{01} + \{d_1(1-u_2) + (1-u_1)d_2 - d_1d_2\}p_{00}$$

If $d_i < 1/2$ and $u_i < 1/4$, then

$$\begin{aligned}
d_1(1-u_2) + (1-u_1)d_2 - d_1d_2 &= d_1(1-u_2-\tfrac{1}{2}d_2) + d_2(1-u_1-\tfrac{1}{2}d_1)\\
&> d_1(1-\tfrac{1}{4}-\tfrac{1}{2}\tfrac{1}{2}) + d_2(1-\tfrac{1}{4}-\tfrac{1}{2}\tfrac{1}{2}) > 0
\end{aligned}$$

Thus, $u_2p_{10} + u_1p_{01} + \{d_1(1-u_2) + (1-u_1)d_2 - d_1d_2\}p_{00} \geq 0$, and so if $p^*_{11} - p^*_{10} - p^*_{01} > 0$, then $V(p_{11} - p_{10} - p_{01}) > 0$. The result follows if $V > 0$. If $d_i < 1/2$ and $u_i < 1/4$, then

$$\begin{aligned}
V &= (1-d_1)(1-d_2) - (1-d_1)u_2 - u_1(1-d_2)\\
&= (1-d_1)\{\tfrac{1}{2}(1-d_2) - u_2\} + (1-d_2)\{\tfrac{1}{2}(1-d_1) - u_1\}
\end{aligned}$$

$$> (1-d_1)(\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{4})+(1-d_2)(\tfrac{1}{2}\tfrac{1}{2}-\tfrac{1}{4}) \geq 0. \quad \blacksquare$$

Proposition 11.11 (VanderWeele, 2012b):
If misclassification of G and E is nondifferential and independent and if $d_i \leq 1/3$ and $u_i \leq 1/4$ then

$$p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00} > 0$$

implies $p_{11} - p_{10} - p_{01} - p_{00} > 0$.

Proof:
We can rewrite $p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00}$ as follows

$$\begin{aligned}
&(1-d_1)(1-d_2)p_{11} - (1-d_1)u_2p_{11} - u_1(1-d_2)p_{11} - u_1u_2p_{11} \\
&+(1-d_1)d_2p_{10} - (1-d_1)(1-u_2)p_{10} - u_1d_2p_{10} - u_1(1-u_2)p_{10} \\
&+d_1(1-d_2)p_{01} - d_1u_2p_{01} - (1-u_1)(1-d_2)p_{01} - (1-u_1)u_2p_{01} \\
&+d_1d_2p_{00} - d_1(1-u_2)p_{00} - (1-u_1)d_2p_{00} - (1-u_1)(1-u_2)p_{00}
\end{aligned}$$

Let $V = (1-d_1)(1-d_2) - (1-d_1)u_2 - u_1(1-d_2) - u_1u_2$. Rewrite $p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00}$ as

$$\begin{aligned}
&Vp_{11} - (V+u_1+u_2)p_{10} - (V+u_1+u_2)p_{01} \\
&\quad - \{V + 2(d_1+d_2+u_1u_2 - d_1d_2 - d_1u_2 - u_1d_2)\}p_{00}.
\end{aligned}$$

Thus, if $p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00} > 0$, then

$$\begin{aligned}
V(p_{11} - p_{10} - p_{01} - p_{00}) > &(u_1+u_2)(p_{10}+p_{00}) \\
&+ 2(d_1+d_2+u_1u_2 - d_1d_2 - d_1u_2 - u_1d_2)p_{00}
\end{aligned}$$

Denote the right side of the inequality by H. If $d_i \leq 1/3$ and $u_i \leq 1/4$, then

$$\begin{aligned}
&(d_1+d_2+u_1u_2 - d_1d_2 - d_1u_2 - u_1d_2) \\
&\quad = d_1\left(1-u_2-\tfrac{1}{2}d_2\right) + d_2\left(1-u_1-\tfrac{1}{2}d_1\right) + u_1u_2 \\
&\quad \geq d_1\left(1-\tfrac{1}{4}-\tfrac{1}{2}\tfrac{1}{3}\right) + d_2\left(1-\tfrac{1}{4}-\tfrac{1}{2}\tfrac{1}{3}\right) + u_1u_2 \geq 0
\end{aligned}$$

Thus, H is non-negative, and consequently, if $p^*_{11} - p^*_{10} - p^*_{01} - p^*_{00} > 0$, then $V(p_{11} - p_{10} - p_{01} - p_{00}) > 0$. The result follows provided $V > 0$. If $d_i \leq 1/3$ and $u_i \leq 1/4$, then

$$\begin{aligned}
V &= (1-d_1)(1-d_2) - (1-d_1)u_2 - u_1(1-d_2) - u_1u_2 \\
&= (1-d_1)\{\tfrac{3}{8}(1-d_2) - u_2\} + (1-d_2)\{\tfrac{3}{8}(1-d_1) - u_1\} \\
&\quad + \{\tfrac{1}{4}(1-d_1)(1-d_2) - u_1u_2\} \\
&\geq (1-d_1)(\tfrac{3}{8}\tfrac{2}{3}-\tfrac{1}{4}) + (1-d_2)(\tfrac{3}{8}\tfrac{2}{3}-\tfrac{1}{4}) + \{\tfrac{1}{4}\tfrac{2}{3}\tfrac{2}{3} - \tfrac{1}{4}\tfrac{1}{4}\} > 0. \quad \blacksquare
\end{aligned}$$

A.11.5. Measurement Error and Multiplicative Interaction

Proposition 11.12 (Garcia-Closas et al., 1998):
Suppose G and E are independent among the controls (i.e., when $Y = 0$), that E is mismeasured, and that the misclassification of E is non-differential in the sense that $P(E^* = e^*|E = e, G = g, Y = d)$ is independent of g, then in the two logistic regression models

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg$$
$$\text{logit}\{P(Y = 1|G = g, E^* = e^*)\} = \gamma_0^* + \gamma_1^* g + \gamma_2^* e^* + \gamma_3^* e^* g$$

$\gamma_3^* \neq 0$ implies $\gamma_3 \neq 0$. Moreover, if E is binary, then $|\gamma_3^*| \leq |\gamma_3|$.

Proof:
We have that

$$P(y|g, e, c) = \frac{P(g|e, y)P(y|e)}{P(g|e)}$$

and thus

$$\exp(\gamma_3) = \frac{\frac{P(Y=1|g_1,e_1)}{P(Y=0|g_1,e_1)} \Big/ \frac{P(Y=1|g_1,e_0)}{P(Y=0|g_1,e_0)}}{\frac{P(Y=1|g_0,e_1)}{P(Y=0|g_0,e_1)} \Big/ \frac{P(Y=1|g_0,e_0)}{P(Y=0|g_0,e_0)}}$$
$$= \frac{\frac{P(g_1|e_1,Y=1)/P(g_0|e_1,Y=1)}{P(g_1|e_0,Y=1)/P(g_0|e_0,Y=1)}}{\frac{P(g_1|e_1,Y=0)/P(g_0|e_1,Y=0)}{P(g_1|e_0,Y=0)/P(g_0|e_0,Y=0)}}$$

Let $OR_{AB|C=c}$ denote the odds ratio of A and B conditional on $C = c$. We thus have $\exp(\gamma_3) = OR_{GE|Y=1}/OR_{GE|Y=0}$. We likewise have that $\exp(\gamma_3^*) = OR_{GE^*|Y=1}/OR_{GE^*|Y=0}$. If $\gamma_3 = 0$ and G and E are independent among the controls, then we have $OR_{GE|Y=0} = 1$; and thus, since $1 = \exp(\gamma_3) = OR_{GE|Y=1}/OR_{GE|Y=0}$ we must have $OR_{GE|Y=1} = 1$. We thus have that G and E are independent conditional on $Y = y$. Under the assumption that the misclassification of E is nondifferential with respect to G, we have

$$P(g|e^*, y) = \sum\nolimits_e P(g|e^*, e, y)P(e|e^*, y) = \sum\nolimits_e P(g|e, y)P(e|e^*, y)$$
$$= \sum\nolimits_e P(g|y)P(e|e^*, y) = P(g|y)$$

Thus G and E^* are independent conditional on $Y = y$ and so $OR_{GE^*|Y=1} = 1$ and $OR_{GE^*|Y=0} = 1$ and thus $\exp(\gamma_3^*) = OR_{GE^*|Y=1}/OR_{GE^*|Y=0} = 1$ and $\gamma_3^* = 0$. Thus if $\gamma_3^* \neq 0$, then $\gamma_3 \neq 0$.

If E is binary, then nondifferential misclassification of E with respect to G will bias $OR_{GE^*|Y=1}$ toward 1 compared to $OR_{GE|Y=1}$. Under independence of G and E among the controls we have, by the argument above, that $OR_{GE^*|Y=0} = OR_{GE|Y=0} = 1$. Thus $\exp(\gamma_3^*) = OR_{GE^*|Y=1}/OR_{GE^*|Y=0} = OR_{GE^*|Y=1}$ will be biased toward 1 compared with $\exp(\gamma_3) = OR_{GE|Y=1}/OR_{GE|Y=0} = OR_{GE|Y=1}$.

A.12. INTERACTION IN GENETICS: INDEPENDENCE AND BOOSTING POWER

Proposition 12.1 (Piegorsch et al,. 1994):
Suppose that G and E are independent conditional on C, then

$$\frac{P(g_1|e_1,Y=1,c)/P(g_0|e_1,Y=1,c)}{P(g_1|e_0,Y=1,c)/P(g_0|e_0,Y=1,c)}\frac{P(Y=1|g_1,e_1,c)/P(Y=1|g_1,e_0,c)}{P(Y=1|g_0,e_1,c)/P(Y=1|g_0,e_0,c)}.$$

Proof:
The case-only estimator is given by

$$\frac{P(g_1|e_1,Y=1,c)/P(g_0|e_1,Y=1,c)}{P(g_1|e_0,Y=1,c)/P(g_0|e_0,Y=1,c)}$$

By Bayes' theorem, we have that

$$\begin{aligned}P(G=g|E=e,Y=1,c) &= \frac{P(Y=1|G=g,E=e,c)P(G=g|E=e,c)}{P(Y=1|E=e,c)}\\ &= \frac{P(Y=1|G=g,E=e,c)P(G=g,c)}{P(Y=1|E=e,c)}\end{aligned}$$

where the second equality follows by independence of G and E. Thus for the case-only estimator, we have

$$\begin{aligned}\frac{P(g_1|e_1,Y=1,c)/P(g_0|e_1,Y=1,c)}{P(g_1|e_0,Y=1,c)/P(g_0|e_0,Y=1,c)} &= \frac{\frac{P(Y=1|g_1,e_1,c)P(g_1|c)}{P(Y=1|e_1,c)} \Big/ \frac{P(Y=1|g_0,e_1,c)P(g_0|c)}{P(Y=1|e_1,c)}}{\frac{P(Y=1|g_1,e_0,c)P(g_1|c)}{P(Y=1|e_0,c)} \Big/ \frac{P(Y=1|g_0,e_0,c)P(g_0|c)}{P(Y=1|e_0,c)}}\\ &= \frac{P(Y=1|g_1,e_1,c)/P(Y=1|g_1,e_0,c)}{P(Y=1|g_0,e_1,c)/P(Y=1|g_0,e_0,c)}\end{aligned}$$

where the final expression is simply the multiplicative interaction on the risk ratio scale. Similar reasoning would hold conditional on covariates C if G and E were conditionally independent given C. ■

A.13. POWER AND SAMPLE SIZE CALCULATIONS FOR INTERACTION ANALYSIS

A.13.1. Power and Sample-Size Calculations for Interaction for Continuous Outcomes

For the linear regression model

$$\mathbb{E}[Y|G=g,E=e] = \tau_0 + \tau_1 g + \tau_2 e + \tau_3 ge$$

suppose we wish to use a Wald test for the null hypothesis $\tau_3 = 0$. The sample size required to detect an multiplicative interaction of magnitude $\tau_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{cts}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where V_{cts} is the variance of $\widehat{\tau}_3$ under the alternative that $\tau_3 = \eta$. Likewise, we can calculate the power for a given sample size using $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{cts})}\right\}$. The variance V_{cts} can be derived as follows. The likelihood is given by

$$L(\tau_0, \tau_1, \tau_2, \tau_3) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\{y_i - (\tau_0 + \tau_1 g + \tau_2 e + \tau_3 ge)\}^2}{2\sigma^2}\right]$$

where σ^2 is the variance of the error term in the regression model for Y—that is, the variance of Y conditional on G and E. The log-likelihood is thus given by

$$l(\tau_0, \tau_1, \tau_2, \tau_3) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\{y_i - (\tau_0 + \tau_1 g + \tau_2 e + \tau_3 ge)\}^2}{2\sigma^2}$$

The second derivative is given by

$$\frac{\partial^2 l(\tau_0, \tau_1, \tau_2, \tau_3)}{\partial(\tau_0, \tau_1, \tau_2, \tau_3)^2} = \sum_{i=1}^{n} -\frac{1}{\sigma^2}\begin{pmatrix} 1 & g_i & e_i & g_i e_i \\ g_i & g_i & g_i e_i & g_i e_i \\ e_i & g_i e_i & e_i & g_i e_i \\ g_i e_i & g_i e_i & g_i e_i & g_i e_i \end{pmatrix}$$

The expected information matrix is then given by

$$\begin{aligned} I &= E\left[\frac{1}{\sigma^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}\right] \\ &= E\left[E\left[\frac{1}{\sigma^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix} | G, E\right]\right] \end{aligned}$$

which we may write as

$$\frac{1}{\sigma^2} M_1 \pi_{00} + \frac{1}{\sigma^2} M_2 \pi_{10} + \frac{1}{\sigma^2} M_3 \pi_{01} + \frac{1}{\sigma^2} M_4 \pi_{11}$$

where $\pi_{ge} = P(G = g, E = e)$ and

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad M_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

If we let $L = \frac{1}{\sigma^2}\pi_{00}$, $F = \frac{1}{\sigma^2}\pi_{10}$, $J = \frac{1}{\sigma^2}\pi_{01}$, and $R = \frac{1}{\sigma^2}\pi_{11}$ we then have

$$I = \begin{pmatrix} L+F+J+R & F+R & J+R & R \\ F+R & F+R & R & R \\ J+R & R & J+R & R \\ R & R & R & R \end{pmatrix}$$

The inverse of this matrix is

$$I^{-1} = \begin{pmatrix} \frac{1}{L} & -\frac{1}{L} & -\frac{1}{L} & \frac{1}{L} \\ -\frac{1}{L} & \frac{1}{L}+\frac{1}{F} & \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} \\ -\frac{1}{L} & \frac{1}{L} & \frac{1}{L}+\frac{1}{J} & -\frac{1}{L}-\frac{1}{J} \\ \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} & -\frac{1}{L}-\frac{1}{J} & \frac{1}{L}+\frac{1}{F}+\frac{1}{J}+\frac{1}{R} \end{pmatrix}$$

from which it follows $V_{cts} = \sigma^2(\frac{1}{\pi_{00}} + \frac{1}{\pi_{10}} + \frac{1}{\pi_{01}} + \frac{1}{\pi_{11}})$.

A.13.2. Power and Sample-Size Calculations for Binary Outcomes: Multiplicative Interaction

For the logistic regression model

$$\text{logit}\{P(Y = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

suppose we wish to use a Wald test for the null hypothesis $\gamma_3 = 0$. The sample size required to detect an multiplicative interaction of magnitude $\gamma_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{mult(OR)}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1 - \alpha/2)$th and βth quantiles respectively of the standard normal distribution and where $V_{mult(OR)}$ is the variance of $\widehat{\gamma}_3$ under the alternative that $\gamma_3 = \eta$. Likewise, we can calculate the power for a given sample size

using $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{mult(OR)})}\right\}$. The variance $V_{mult(OR)}$ can be derived as follows. The likelihood is given by

$$L(\gamma_0,\gamma_1,\gamma_2,\gamma_3) = \prod_{i=1}^{n}\left\{\frac{e^{\gamma_0+\gamma_1 g_i+\gamma_2 e_i+\gamma_3 g_i e_i}}{1+e^{\gamma_0+\gamma_1 g_i+\gamma_2 e_i+\gamma_3 g_i e_i}}\right\}^{y_i} \times\left\{\frac{1}{1+e^{\gamma_0+\gamma_1 g_i+\gamma_2 e_i+\gamma_3 g_i e_i}}\right\}^{1-y_i}$$

and the log-likelihood by

$$l(\gamma_0,\gamma_1,\gamma_2,\gamma_3) = \sum_{i=1}^{n} y_i(\gamma_0+\gamma_1 g_i+\gamma_2 e_i+\gamma_3 g_i e_i)+\log(1+e^{\gamma_0+\gamma_1 g_i+\gamma_2 e_i+\gamma_3 g_i e_i})$$

As in Demidenko (2008), the second derivative is given by

$$\frac{\partial^2 l(\gamma_0,\gamma_1,\gamma_2,\gamma_3)}{\partial(\gamma_0,\gamma_1,\gamma_2,\gamma_3)^2} = \sum_{i=1}^{n}\frac{-Q_i}{(1+Q_i)^2}\begin{pmatrix} 1 & g_i & e_i & g_i e_i \\ g_i & g_i & g_i e_i & g_i e_i \\ e_i & g_i e_i & e_i & g_i e_i \\ g_i e_i & g_i e_i & g_i e_i & g_i e_i \end{pmatrix}$$

where $Q_i = e^{\gamma_0+\gamma_1 g_i+\gamma_2 e_i+\gamma_3 g_i e_i}$. Let $Q = e^{\gamma_0+\gamma_1 G+\gamma_2 E+\gamma_3 GE}$. The expected information matrix is then given by

$$I = E\left[\frac{Q}{(1+Q)^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}\right]$$
$$= E\left[E\left[\frac{Q}{(1+Q)^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}|G,E\right]\right]$$

which we may write as

$$\frac{e^{\gamma_0}}{(1+e^{\gamma_0})^2}M_1\pi_{00} + \frac{e^{\gamma_0+\gamma_1}}{(1+e^{\gamma_0+\gamma_1})^2}M_2\pi_{10} + \frac{e^{\gamma_0+\gamma_2}}{(1+e^{\gamma_0+\gamma_2})^2}M_3\pi_{01} + \frac{e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}{(1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3})^2}M_4\pi_{11}$$

where $\pi_{ge} = P(G=g, E=e)$ and

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad M_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

If we let $L = \frac{e^{\gamma 0}}{(1+e^{\gamma 0})^2}\pi_{00}$, $F = \frac{e^{\gamma 0+\gamma 1}}{(1+e^{\gamma 0+\gamma 1})^2}\pi_{10}$, $J = \frac{e^{\gamma 0+\gamma 2}}{(1+e^{\gamma 0+\gamma 2})^2}\pi_{01}$, and $R = \frac{e^{\gamma 0+\gamma 1+\gamma 2+\gamma 3}}{(1+e^{\gamma 0+\gamma 1+\gamma 2+\gamma 3})^2}\pi_{11}$ we then have

$$I = \begin{pmatrix} L+F+J+R & F+R & J+R & R \\ F+R & F+R & R & R \\ J+R & R & J+R & R \\ R & R & R & R \end{pmatrix}$$

The inverse of this matrix is

$$I^{-1} = \begin{pmatrix} \frac{1}{L} & -\frac{1}{L} & -\frac{1}{L} & \frac{1}{L} \\ -\frac{1}{L} & \frac{1}{L}+\frac{1}{F} & \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} \\ -\frac{1}{L} & \frac{1}{L} & \frac{1}{L}+\frac{1}{J} & -\frac{1}{L}-\frac{1}{J} \\ \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} & -\frac{1}{L}-\frac{1}{J} & \frac{1}{L}+\frac{1}{F}+\frac{1}{J}+\frac{1}{R} \end{pmatrix}$$

from which it follows $V_{mult(OR)} = \frac{1}{L}+\frac{1}{F}+\frac{1}{J}+\frac{1}{R}$.

A.13.3. Derivations for Case–Control Exposure Probabilities from the Probabilities in the Underlying Population

Here we derive the proportions in each joint exposure group in the case–control sample, $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$, from the proportion in each joint exposure group in the underlying population, $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$, under an assumption that the outcome is rare. We will use $P^*(\cdot)$ to denote probabilities in the case–control sample and $P(\cdot)$ to denote probabilities in the underlying population. We have that

$$\begin{aligned} \pi^*_{ge} &= P^*(G=g, E=e) \\ &= P^*(G=g, E=e|Y=0)P^*(Y=0) + P^*(G=g, E=e|Y=1)P^*(Y=1) \\ &= P(G=g, E=e|Y=0)P^*(Y=0) + P(G=g, E=e|Y=1)P^*(Y=1) \\ &\approx \pi_{ge}P^*(Y=0) + P(G=g, E=e|Y=1)P^*(Y=1) \end{aligned}$$

where the final equality follows because the outcome is rare and thus the exposure distribution among the controls will approximate that in the underlying population. We then also have that

$$\begin{aligned} P(G=g, E=e|Y=1) &= \frac{P(Y=1|G=g, E=e)P(G=g, E=e)}{P(Y=1)} \\ &= \frac{P(Y=1|G=g, E=e)P(G=g, E=e)}{\sum_{i,j} P(Y=1|G=i, E=j)P(G=i, E=j)} \end{aligned}$$

$$= \frac{\frac{P(Y=1|G=g,E=e)}{P(Y=1|G=0,E=0)}\pi_{ge}}{\sum_{i,j}\frac{P(Y=1|G=i,E=j)}{P(Y=1|G=0,E=0)}\pi_{ij}}$$
$$\approx \frac{\frac{P(Y=1|G=g,E=e)/\{1-P(Y=1|G=g,E=e)\}}{P(Y=1|G=0,E=0)/\{1-P(Y=1|G=0,E=0)\}}\pi_{ge}}{\sum_{i,j}\frac{P(Y=1|G=i,E=j)/\{1-P(Y=1|G=i,E=j)\}}{P(Y=1|G=0,E=0)/\{1-P(Y=1|G=0,E=0)\}}\pi_{ij}}$$

where the final equality follows from the rare outcome assumption which implies that risk ratios approximate odds ratio. The odds ratios can then be obtained from the specification of the parameters of the logistic regression model and we thus obtain

$$\pi^*_{00} \approx \pi_{00}P^*(Y=0) + \frac{\pi_{00}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$
$$\pi^*_{10} \approx \pi_{10}P^*(Y=0) + \frac{e^{\gamma_1}\pi_{10}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$
$$\pi^*_{01} \approx \pi_{01}P^*(Y=0) + \frac{e^{\gamma_2}\pi_{01}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$
$$\pi^*_{11} \approx \pi_{11}P^*(Y=0) + \frac{e^{\gamma_1+\gamma_2+\gamma_3}\pi_{11}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)$$

Under this rare outcome assumption we can also obtain $\gamma_0 = \log\{P^*(Y=1|G=0,E=0)/P^*(Y=0|G=0,E=0)\}$, the log odds of baseline probability of the outcome in doubly unexposed group in the case–control sample, from $P^*(Y=0)$ and $P^*(Y=1)$ because

$$\begin{aligned}
&P^*(Y=1|G=0,E=0)\\
&= \frac{P^*(G=0,E=0|Y=1)P^*(Y=1)}{P^*(G=0,E=0)}\\
&= \frac{P(G=0,E=0|Y=1)P^*(Y=1)}{P^*(G=0,E=0)}\\
&\approx \frac{\frac{\pi_{00}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)}{\pi_{00}P^*(Y=0)+\frac{\pi_{00}}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}}P^*(Y=1)}\\
&= \frac{P^*(Y=1)}{\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3}P^*(Y=0)+P^*(Y=1)}\\
&= 1/\{1+(\pi_{00}+\pi_{10}e^{\gamma_1}+\pi_{01}e^{\gamma_2}+\pi_{11}e^{\gamma_1+\gamma_2+\gamma_3})P^*(Y=0)/P^*(Y=1)\}
\end{aligned}$$

If instead the proportions in each joint exposure group in the case–control sample are specified, $\pi^*_{00}, \pi^*_{10}, \pi^*_{01}, \pi^*_{11}$, then we could obtain γ_0 by numerically solving

$$P^*(Y=0) = \frac{\pi^*_{00}}{1+e^{\gamma_0}} + \frac{\pi^*_{10}}{1+e^{\gamma_0+\gamma_1}} + \frac{\pi^*_{01}}{1+e^{\gamma_0+\gamma_2}} + \frac{\pi^*_{11}}{1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}$$

for γ_0. If the joint or marginal exposure probabilities are specified separately for the cases and controls then under an assumption of a rare outcome, the distribution of the exposures amongst the controls could be used as an approximation to $\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}$ or π_g, π_e, Δ.

A.13.4. Multiplicative Interaction with Case-Only Data

Consider the logistic regression model:

$$\text{logit}\{P(G = 1|E = e, Y = 1)\} = \theta_0 + \theta_1 e$$

As noted in Chapters 12 and 13, under independence of G and E, θ_1 will equal the multiplicative interaction on the risk ratio scale. The sample size required to detect an interaction of magnitude $\theta_1 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{CO}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1 - \alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where V_{CO} is the variance of $\widehat{\theta}_1$ under the alternative that $\theta_1 = \eta$. The variance V_{CO} is given by Yang et al. (1997; cf. VanderWeele, 2011g) as

$$V_{CO} = (m_1 + m_2 + m_3 + m_4)\left(\frac{1}{m_1} + \frac{1}{m_2} + \frac{1}{m_3} + \frac{1}{m_4}\right)$$

with

$$\begin{aligned} m_1 &= (1 - \pi_g)(1 - \pi_e) \\ m_2 &= \pi_g(1 - \pi_e)\exp(\gamma_1) \\ m_3 &= (1 - \pi_g)\pi_e \exp(\gamma_2) \\ m_4 &= \pi_g \pi_e \exp(\gamma_1 + \gamma_2 + \gamma_3). \end{aligned}$$

A.13.5. Additive Interaction in Cohort Studies Using Linear Risk Model

For the linear model,

$$P(Y = 1|G = g, E = e) = \theta_0 + \theta_1 g + \theta_2 e + \theta_3 ge$$

suppose we wish to use a Wald test for the null hypothesis $\theta_3 = 0$. The sample size required to detect an multiplicative interaction of magnitude $\theta_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where V is the variance of $\widehat{\theta}_3$ under the alternative that $\theta_3 = \eta$. Likewise, we can calculate the power for a given sample size using $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V)}\right\}$. The variance V can be derived as follows. The likelihood is given by

$$L(\theta_0,\theta_1,\theta_2,\theta_3) = \prod_{i=1}^{n} (\theta_0 + \theta_1 g_i + \theta_2 e_i + \theta_3 g_i e_i)^{y_i} \times \{1 - (\theta_0 + \theta_1 g_i + \theta_2 e_i + \theta_3 g_i e_i)\}^{1-y_i}$$

and the log-likelihood by

$$l(\theta_0,\theta_1,\theta_2,\theta_3) = \sum_{i=1}^{n} y_i \log(\theta_0 + \theta_1 g_i + \theta_2 e_i + \theta_3 g_i e_i) + \log\{1 - (\theta_0 + \theta_1 g_i + \theta_2 e_i + \theta_3 g_i e_i)\}(1 - y_i)$$

The second derivative is given by

$$\frac{\partial^2 l(\theta_0,\theta_1,\theta_2,\theta_3)}{\partial(\theta_0,\theta_1,\theta_2,\theta_3)^2} = \sum_{i=1}^{n} \frac{-y_i + 2y_i Q_i - {Q_i}^2}{{Q_i}^2(1-Q_i)^2} \begin{pmatrix} 1 & g_i & e_i & g_i e_i \\ g_i & g_i & g_i e_i & g_i e_i \\ e_i & g_i e_i & e_i & g_i e_i \\ g_i e_i & g_i e_i & g_i e_i & g_i e_i \end{pmatrix}$$

where $Q_i = \theta_0 + \theta_1 g_i + \theta_2 e_i + \theta_3 g_i e_i$. Let $Q = \theta_0 + \theta_1 G + \theta_2 E + \theta_3 GE$. The expected information matrix is then given by

$$I = E\left[\frac{Y - 2YQ + Q^2}{Q^2(1-Q)^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}\right]$$

$$= E\left[E\left[\frac{Y - 2YQ + Q^2}{Q^2(1-Q)^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}|G,E\right]\right]$$

$$= E\left[E\left[\frac{1}{Q(1-Q)}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}|G,E\right]\right]$$

which we may write as

$$\frac{1}{(\theta_0)(1-\theta_0)}M_1\pi_{00} + \frac{1}{(\theta_0+\theta_1)\{1-(\theta_0+\theta_1)\}}M_2\pi_{10} + \frac{1}{(\theta_0+\theta_1+\theta_2)\{1-(\theta_0+\theta_1+\theta_2)\}}M_3\pi_{01} + \frac{1}{(\theta_0+\theta_1+\theta_2+\theta_3)\{1-(\theta_0+\theta_1+\theta_2+\theta_3)\}}M_4\pi_{11}$$

where

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad M_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

If we let $L' = \frac{1}{(\theta_0)(1-\theta_0)}\pi_{00}, F' = \frac{1}{(\theta_0+\theta_1)\{1-(\theta_0+\theta_1)\}}\pi_{10}, J' = \frac{1}{(\theta_0+\theta_2)\{1-(\theta_0+\theta_2)\}}\pi_{01}$, and $R' = \frac{1}{(\theta_0+\theta_1+\theta_2+\theta_3)\{1-(\theta_0+\theta_1+\theta_2+\theta_3)\}}\pi_{11}$, we then have

$$I = \begin{pmatrix} L'+F'+J'+R' & F'+R' & J'+R' & R' \\ F'+R' & F'+R' & R' & R' \\ J'+R' & R' & J'+R' & R' \\ R' & R' & R' & R' \end{pmatrix}$$

The inverse of this matrix is

$$I^{-1} = \begin{pmatrix} \frac{1}{L'} & -\frac{1}{L'} & -\frac{1}{L'} & \frac{1}{L'} \\ -\frac{1}{L'} & \frac{1}{L'}+\frac{1}{F'} & \frac{1}{L'} & -\frac{1}{L'}-\frac{1}{F'} \\ -\frac{1}{L'} & \frac{1}{L'} & \frac{1}{L'}+\frac{1}{J'} & -\frac{1}{L'}-\frac{1}{J'} \\ \frac{1}{L'} & -\frac{1}{L'}-\frac{1}{F'} & -\frac{1}{L'}-\frac{1}{J'} & \frac{1}{L'}+\frac{1}{F'}+\frac{1}{J'}+\frac{1}{R'} \end{pmatrix}$$

from which it follows $V = \frac{1}{L'} + \frac{1}{F'} + \frac{1}{J'} + \frac{1}{R'}$.

A.13.6. Additive Interaction in Cohort Studies Using Logistic Regression and RERI

As noted in Appendix 13.2, for the logistic regression model

$$\text{logit}\{P(Y=1|G=g,E=e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 ge$$

the variance–covariance matrix for the maximum likelihood estimate of $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ was given by

$$\begin{pmatrix} \frac{1}{L} & -\frac{1}{L} & -\frac{1}{L} & \frac{1}{L} \\ -\frac{1}{L} & \frac{1}{L}+\frac{1}{F} & \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} \\ -\frac{1}{L} & \frac{1}{L} & \frac{1}{L}+\frac{1}{J} & -\frac{1}{L}-\frac{1}{J} \\ \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} & -\frac{1}{L}-\frac{1}{J} & \frac{1}{L}+\frac{1}{F}+\frac{1}{J}+\frac{1}{R} \end{pmatrix},$$

where

$$L = \frac{e^{\gamma_0}}{(1+e^{\gamma_0})^2}\pi_{00}$$

$$F = \frac{e^{\gamma_0+\gamma_1}}{(1+e^{\gamma_0+\gamma_1})^2}\pi_{10}$$
$$J = \frac{e^{\gamma_0+\gamma_2}}{(1+e^{\gamma_0+\gamma_2})^2}\pi_{01}$$
$$R = \frac{e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3}}{(1+e^{\gamma_0+\gamma_1+\gamma_2+\gamma_3})^2}\pi_{11}$$

From the delta method, it follows that the variance of $\widehat{RERI} = e^{\widehat{\gamma}_1+\widehat{\gamma}_2+\widehat{\gamma}_3} - e^{\widehat{\gamma}_1} - e^{\widehat{\gamma}_2} + 1$ is given by

$$\begin{pmatrix} 0 \\ e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} \\ e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_2} \\ e^{\gamma_1+\gamma_2+\gamma_3} \end{pmatrix}' \begin{pmatrix} \frac{1}{L} & -\frac{1}{L} & -\frac{1}{L} & \frac{1}{L} \\ -\frac{1}{L} & \frac{1}{L}+\frac{1}{F} & \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} \\ -\frac{1}{L} & \frac{1}{L} & \frac{1}{L}+\frac{1}{J} & -\frac{1}{L}-\frac{1}{J} \\ \frac{1}{L} & -\frac{1}{L}-\frac{1}{F} & -\frac{1}{L}-\frac{1}{J} & \frac{1}{L}+\frac{1}{F}+\frac{1}{J}+\frac{1}{R} \end{pmatrix}$$
$$\times \begin{pmatrix} 0 \\ e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} \\ e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_2} \\ e^{\gamma_1+\gamma_2+\gamma_3} \end{pmatrix}$$
$$= \begin{pmatrix} 0 \\ e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} \\ e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_2} \\ e^{\gamma_1+\gamma_2+\gamma_3} \end{pmatrix}' \begin{pmatrix} -\frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} + \frac{1}{L}(e^{\gamma_1}+e^{\gamma_2}) \\ \frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} - (\frac{1}{L}+\frac{1}{F})e^{\gamma_1} - \frac{1}{L}e^{\gamma_2} \\ \frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} - \frac{1}{L}e^{\gamma_1} - (\frac{1}{L}+\frac{1}{J})e^{\gamma_2} \\ (-\frac{1}{L}+\frac{1}{R})e^{\gamma_1+\gamma_2+\gamma_3} + (\frac{1}{L}+\frac{1}{F})e^{\gamma_1} + (\frac{1}{L}+\frac{1}{J})e^{\gamma_2} \end{pmatrix}$$

$$= e^{\gamma_1+\gamma_2+\gamma_3}\left\{\frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} - \left(\frac{1}{L}+\frac{1}{F}\right)e^{\gamma_1} - \frac{1}{L}e^{\gamma_2} + \frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} - \frac{1}{L}e^{\gamma_1}\right.$$
$$\left. - \left(\frac{1}{L}+\frac{1}{J}\right)e^{\gamma_2} + \left(-\frac{1}{L}+\frac{1}{R}\right)e^{\gamma_1+\gamma_2+\gamma_3} + \left(\frac{1}{L}+\frac{1}{F}\right)e^{\gamma_1} + \left(\frac{1}{L}+\frac{1}{J}\right)e^{\gamma_2}\right\}$$
$$-e^{\gamma_1}\left\{\frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} - \left(\frac{1}{L}+\frac{1}{F}\right)e^{\gamma_1} - \frac{1}{L}e^{\gamma_2}\right\}$$
$$-e^{\gamma_2}\left\{\frac{1}{L}e^{\gamma_1+\gamma_2+\gamma_3} - \frac{1}{L}e^{\gamma_1} - \left(\frac{1}{L}+\frac{1}{J}\right)e^{\gamma_2}\right\}$$
$$= \left(\frac{1}{L}+\frac{1}{R}\right)e^{2(\gamma_1+\gamma_2+\gamma_3)} - \frac{2}{L}e^{2\gamma_1+\gamma_2+\gamma_3} - \frac{2}{L}e^{\gamma_1+2\gamma_2+\gamma_3}$$
$$+ \left(\frac{1}{L}+\frac{1}{F}\right)e^{2\gamma_1} + \left(\frac{1}{L}+\frac{1}{J}\right)e^{2\gamma_2} + \frac{2}{L}e^{\gamma_1+\gamma_2}.$$

A.13.7. Multiplicative and Additive Interaction for the log-linear model

For the log-linear model,

$$\log\{P(Y=1|G=g,E=e)\} = \kappa_0 + \kappa_1 g + \kappa_2 e + \kappa_3 ge$$

suppose we wish to use a Wald test for the null hypothesis $\kappa_3 = 0$. The sample size required to detect an multiplicative interaction of magnitude $\kappa_3 = \eta$ with significance level α and power β is

$$n = \frac{(Z_{1-\alpha/2} + Z_\beta)^2 V_{mult(RR)}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles respectively of the standard normal distribution and where $V_{mult(RR)}$ is the variance of $\widehat{\kappa}_3$ under the alternative that $\kappa_3 = \eta$. Likewise, we can calculate the power for a given sample size using $Power = \Phi^{-1}\left\{-Z_{1-\alpha/2} + \eta\sqrt{(n/V_{mult(RR)})}\right\}$. The variance $V_{mult(RR)}$ can be derived as follows. The likelihood is given by

$$L(\kappa_0,\kappa_1,\kappa_2,\kappa_3) = \prod_{i=1}^{n} e^{(\kappa_0+\kappa_1 g_i+\kappa_2 e_i+\kappa_3 g_i e_i)y_i}\{1 - e^{(\kappa_0+\kappa_1 g_i+\kappa_2 e_i+\kappa_3 g_i e_i)}\}^{1-y_i}$$

and the log-likelihood is given by

$$\begin{aligned} l(\kappa_0,\kappa_1,\kappa_2,\kappa_3) = &\sum_{i=1}^{n} y_i(\kappa_0 + \kappa_1 g_i + \kappa_2 e_i + \kappa_3 g_i e_i) \\ &+ \log\{1 - e^{(\kappa_0+\kappa_1 g_i+\kappa_2 e_i+\kappa_3 g_i e_i)})\}(1-y_i) \end{aligned}$$

The second derivative is given by

$$\frac{\partial^2 l(\kappa_0,\kappa_1,\kappa_2,\kappa_3)}{\partial(\kappa_0,\kappa_1,\kappa_2,\kappa_3)^2} = \sum_{i=1}^{n} \frac{-(1-y_i)Q_i}{(1-Q_i)^2} \begin{pmatrix} 1 & g_i & e_i & g_i e_i \\ g_i & g_i & g_i e_i & g_i e_i \\ e_i & g_i e_i & e_i & g_i e_i \\ g_i e_i & g_i e_i & g_i e_i & g_i e_i \end{pmatrix}$$

where $Q_i = e^{\kappa_0+\kappa_1 g_i+\kappa_2 e_i+\kappa_3 g_i e_i}$. Let $Q = e^{\kappa_0+\kappa_1 G+\kappa_2 E+\kappa_3 GE}$. The expected information matrix is then given by

$$\begin{aligned} I &= E\left[\frac{(1-Y)Q}{(1-Q)^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}\right] \\ &= E\left[E\left[\frac{(1-Y)Q}{(1-Q)^2}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}|G,E\right]\right] \\ &= E\left[E\left[\frac{Q}{(1-Q)}\begin{pmatrix} 1 & G & E & GE \\ G & G & GE & GE \\ E & GE & E & GE \\ GE & GE & GE & GE \end{pmatrix}|G,E\right]\right] \end{aligned}$$

which we may write as

$$\frac{e^{\kappa_0}}{(1-e^{\kappa_0})}M_1\pi_{00} + \frac{e^{\kappa_0+\kappa_1}}{(1-e^{\kappa_0+\kappa_1})}M_2\pi_{10}$$

$$+\frac{e^{\kappa_0+\kappa_2}}{(1-e^{\kappa_0+\kappa_2})}M_3\pi_{01}+\frac{e^{\kappa_0+\kappa_1+\kappa_2+\kappa_3}}{(1-e^{\kappa_0+\kappa_1+\kappa_2+\kappa_3})}M_4\pi_{11}$$

where

$$M_1=\begin{pmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix},\qquad M_2=\begin{pmatrix}1&1&0&0\\1&1&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}$$

$$M_3=\begin{pmatrix}1&0&1&0\\0&0&0&0\\1&0&1&0\\0&0&0&0\end{pmatrix},\qquad M_4=\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$$

If we let $L^\dagger=\frac{e^{\kappa_0}}{(1-e^{\kappa_0})}\pi_{00}$, $F^\dagger=\frac{e^{\kappa_0+\kappa_1}}{(1-e^{\kappa_0+\kappa_1})}\pi_{10}$, $J^\dagger=\frac{e^{\kappa_0+\kappa_2}}{(1-e^{\kappa_0+\kappa_2})}\pi_{01}$, and $R^\dagger=\frac{e^{\kappa_0+\kappa_1+\kappa_2+\kappa_3}}{(1-e^{\kappa_0+\kappa_1+\kappa_2+\kappa_3})}\pi_{11}$ we then have

$$I=\begin{pmatrix}L^\dagger+F^\dagger+J^\dagger+R^\dagger & F^\dagger+R^\dagger & J^\dagger+R^\dagger & R^\dagger\\ F^\dagger+R^\dagger & F^\dagger+R^\dagger & R^\dagger & R^\dagger\\ J^\dagger+R^\dagger & R^\dagger & J^\dagger+R^\dagger & R^\dagger\\ R^\dagger & R^\dagger & R^\dagger & R^\dagger\end{pmatrix}$$

The inverse of this matrix is

$$I^{-1}=\begin{pmatrix}\frac{1}{L^\dagger} & -\frac{1}{L^\dagger} & -\frac{1}{L^\dagger} & \frac{1}{L^\dagger}\\ -\frac{1}{L^\dagger} & \frac{1}{L^\dagger}+\frac{1}{F^\dagger} & \frac{1}{L^\dagger} & -\frac{1}{L^\dagger}-\frac{1}{F^\dagger}\\ -\frac{1}{L^\dagger} & \frac{1}{L^\dagger} & \frac{1}{L^\dagger}+\frac{1}{J^\dagger} & -\frac{1}{L^\dagger}-\frac{1}{J^\dagger}\\ \frac{1}{L^\dagger} & -\frac{1}{L^\dagger}-\frac{1}{F^\dagger} & -\frac{1}{L^\dagger}-\frac{1}{J^\dagger} & \frac{1}{L^\dagger}+\frac{1}{F^\dagger}+\frac{1}{J^\dagger}+\frac{1}{R^\dagger}\end{pmatrix}$$

from which it follows $V_{mult(RR)}=\frac{1}{L^\dagger}+\frac{1}{F^\dagger}+\frac{1}{J^\dagger}+\frac{1}{R^\dagger}$.

The RERI from the log-linear model is given by

$$RERI=e^{\kappa_1+\kappa_2+\kappa_3}-e^{\kappa_1}-e^{\kappa_2}+1$$

Suppose we wish to use a Wald test for the null hypothesis $RERI=0$. The sample size required to detect a $RERI$ of magnitude $\eta=e^{\kappa_1+\kappa_2+\kappa_3}-e^{\kappa_1}-e^{\kappa_2}+1$ with significance level α and power β is

$$n=\frac{(Z_{1-\alpha/2}+Z_\beta)^2V_{RERI(RR)}}{\eta^2}$$

where $Z_{1-\alpha/2}$ and Z_β are the $(1-\alpha/2)$th and βth quantiles, respectively, of the standard normal distribution and where $V_{RERI(RR)}$ is the variance of $RERI=e^{\widehat{\kappa}_1+\widehat{\kappa}_2+\widehat{\kappa}_3}-e^{\widehat{\kappa}_1}-e^{\widehat{\kappa}_2}+1$ under the alternative. Likewise, to calculate the power for a given sample size we could use

$$Power=\Phi^{-1}\left\{-Z_{1-\alpha/2}+\eta\sqrt{(n/V_{RERI(RR)})}\right\}$$

Using an argument analogous to that in Section A.13.6 we have that

$$V^*_{RERI(RR)} = \left(\frac{1}{L^\dagger} + \frac{1}{R^\dagger}\right) e^{2(\kappa_1+\kappa_2+\kappa_3)} - \frac{2}{L^\dagger} e^{2\kappa_1+\kappa_2+\kappa_3} - \frac{2}{L^\dagger} e^{\kappa_1+2\kappa_2+\kappa_3} + \left(\frac{1}{L^\dagger} + \frac{1}{F^\dagger}\right) e^{2\kappa_1} + \left(\frac{1}{L^\dagger} + \frac{1}{J^\dagger}\right) e^{2\kappa_2} + \frac{2}{L^\dagger} e^{\kappa_1+\kappa_2}.$$

A.14. A UNIFICATION OF MEDIATION AND INTERACTION

A.14.1. A General Four-Way Decomposition

As in prior appendices on mediation, we will let A denote the exposure of interest, Y the outcome, and M a potential mediator, and let C denote a set of baseline covariates. We will suppose we want to compare two levels of the exposure, a and a^*. We let Y_a and M_a denote, respectively, the values of the outcome and mediator that would have been observed had the exposure A been set to level a. We let Y_{am} denote the value of the outcome that would have been observed had A been set to level a, and M to m. We will also later consider counterfactuals of the form $Y_{aM_{a^*}}$ which is the outcome Y that would have occurred if we fixed A to a and we fixed M to the level it would have taken if A had been a^*. We will also make some technical assumptions referred to as consistency and composition that are also needed to relate the observed data to counterfactual quantities. The consistency assumption in this context is that when $A = a$, the counterfactual outcomes Y_a and M_a are equal to the observed outcomes Y and M, respectively, and that when $A = a$ and $M = m$, the counterfactual outcome Y_{am} is equal to Y. The composition assumption is that $Y_a = Y_{aM_a}$. Further discussion of these assumptions is given elsewhere (VanderWeele and Vansteelandt, 2009). Here we will give a four-way decomposition, but we will no longer restrict attention to binary exposure and mediator and will consider an arbitrary exposure and mediator. We will assume we are comparing two exposure levels a and a^*. We give the general four-way decomposition result in Proposition 14.1.

Proposition 14.1 (VanderWeele, 2014):
For any level m^* of M we have

$$Y_a - Y_{a^*} = (Y_{am^*} - Y_{a^*m^*}) + \sum_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) 1(M_{a^*} = m) + \sum_m (Y_{am} - Y_{a^*m}) \{1(M_a = m) - 1(M_{a^*} = m)\} + (Y_{a^*M_a} - Y_{a^*M_{a^*}})$$

Proof:
We have that

$$\begin{aligned} Y_a - Y_{a^*} &= Y_{aM_a} - Y_{a^*M_{a^*}} \\ &= (Y_{aM_a} - Y_{a^*M_a}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \\ &= (Y_{aM_{a^*}} - Y_{a^*M_{a^*}}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \end{aligned}$$

$$
\begin{aligned}
&+(Y_{aM_a} - Y_{a^*M_a} - Y_{aM_{a^*}} + Y_{a^*M_{a^*}}) \\
&= (Y_{am^*} - Y_{a^*m^*}) + \{(Y_{aM_{a^*}} - Y_{a^*M_{a^*}}) - (Y_{am^*} - Y_{a^*m^*})\} \\
&\quad + (Y_{aM_a} - Y_{a^*M_a} - Y_{aM_{a^*}} + Y_{a^*M_{a^*}}) + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \\
&= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m \{(Y_{am} - Y_{a^*m}) - (Y_{am^*} - Y_{a^*m^*})\} 1(M_{a^*} = m) \\
&\quad + \sum\nolimits_m \{(Y_{am} - Y_{a^*m}) 1(M_a = m) - (Y_{am} - Y_{a^*m}) 1(M_{a^*} = m)\} \\
&\quad + (Y_{a^*M_a} - Y_{a^*M_{a^*}}) \\
&= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) 1(M_{a^*} = m) \\
&\quad + \sum\nolimits_m (Y_{am} - Y_{a^*m}) \{1(M_a = m) - 1(M_{a^*} = m)\} \\
&\quad + (Y_{a^*M_a} - Y_{a^*M_{a^*}}). \quad \blacksquare
\end{aligned}
$$

The four components of the decomposition in general form are thus

$$
\begin{aligned}
CDE(m^*) &:= (Y_{am^*} - Y_{a^*m^*}) \\
INT_{ref}(m^*) &:= \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}) 1(M_{a^*} = m) \\
INT_{med} &:= \sum\nolimits_m (Y_{am} - Y_{a^*m}) \{1(M_a = m) - 1(M_{a^*} = m)\} \\
PIE &:= (Y_{a^*M_a} - Y_{a^*M_{a^*}})
\end{aligned}
$$

The components are those due to neither mediation nor interaction ($CDE(m^*)$), due to just interaction ($INT_{ref}(m^*)$), due to both mediation and interaction (INT_{med}), and due to just mediation (PIE). Note we can also rewrite $INT_{med} = \sum_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})\{1(M_a = m) - 1(M_{a^*} = m)\}$ and we can rewrite $PIE = \sum_m (Y_{a^*m} - Y_{a^*m^*})\{1(M_a = m) - 1(M_{a^*} = m)\}$. Doing so with binary A and M and setting $a = 1, a^* = 0, m^* = 0$ gives us the decomposition in (14.1) in the text:

$$
\begin{aligned}
Y_1 - Y_0 = (Y_{10} - Y_{00}) &+ (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\
&+ (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0)
\end{aligned}
$$

The decomposition also has an empirical analogue given in the next Proposition.

Proposition 14.2 (VanderWeele, 2014):
For any level m^* of M we have

$$
\begin{aligned}
\mathbb{E}[Y|a,c] - \mathbb{E}[Y|a^*,c] = \{&\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\} \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a,m^*,c] \\
&+ \mathbb{E}[Y|a^*,m^*,c]\}\, dP(m|a^*,c) \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}\{dP(m|a,c)
\end{aligned}
$$

$$- dP(m|a^*,c)\}$$
$$+ \int \mathbb{E}[Y|a^*,m,c]\{dP(m|a,c) - dP(m|a^*,c)\}.$$

Proof:
We have that

$$\begin{aligned}
&\mathbb{E}[Y|a,c] - \mathbb{E}[Y|a^*,c] \\
&= \mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c] + \{\mathbb{E}[Y|a,c] - \mathbb{E}[Y|a,m^*,c]\} - \{\mathbb{E}[Y|a^*,c] \\
&-\mathbb{E}[Y|a^*,m^*,c]\} \\
&= \mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c] + \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a,m^*,c]\}\, dP(m|a,c) \\
&- \int \{\mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a^*,m^*,c]\} dP(m|a^*,c) \\
&= \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\} \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\} - \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\}\, dP(m|a,c) \\
&+ \int \{\mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
&= \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\} \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\} - \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\}\, dP(m|a^*,c) \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\} - \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\}\{dP(m|a,c) \\
&- dP(m|a^*,c)\} \\
&+ \int \{\mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
&= \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\} \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a,m^*,c] + \mathbb{E}[Y|a^*,m^*,c]\}\, dP(m|a^*,c) \\
&+ \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
&+ \int \mathbb{E}[Y|a^*,m,c]\{dP(m|a,c) - dP(m|a^*,c)\}. \quad \blacksquare
\end{aligned}$$

Note we can also rewrite the third term as $\int\{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\} - \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\}$ and the fourth term as $\int\{\mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a^*,m^*,c]\}\{dP(m|a,c) - dP(m|a^*,c)\}$. Doing so with binary A and M and setting $a = 1, a^* = 0, m^* = 0$ gives decomposition (14.1b) in the text:

$$\begin{aligned}
p_{a=1} - p_{a=0} &= (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\
&+(p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\
&+(p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \qquad \textbf{(A.14.5)}
\end{aligned}$$

Note that the decomposition in Proposition 14.2 is a property of the expectations and probabilities. It does not require confounding assumptions. However, to interpret the components as causal effects, confounding assumptions are required. We will begin our discussion of confounding by first considering nonparametric structural equations (Pearl, 2009). Consider the following four confounding assumptions: (A2.1) the effect the exposure A on the outcome Y is unconfounded conditional on C; (A2.2) the effect the mediator M on the outcome Y is unconfounded conditional on C; (A2.3) the effect the exposure A on the mediator M is unconfounded conditional on C; and (A2.4) none of the mediator–outcome confounders are themselves affected by the exposure. As in Section A.2, if we let $X \perp\!\!\!\perp Y|Z$ denote that X is independent of Y conditional on Z, then these four assumptions stated formally in terms of counterfactual independence are (A2.1) $Y_{am} \perp\!\!\!\perp A|C$, (A2.2) $Y_{am} \perp\!\!\!\perp M|\{A,C\}$, (A2.3) $M_a \perp\!\!\!\perp A|C$, and (A2.4) $Y_{am} \perp\!\!\!\perp M_{a^*}|C$.

Proposition 14.3 (VanderWeele, 2014):
Under assumptions (A2.1)–(A2.4) we have

$$\begin{aligned}
\mathbb{E}[CDE(m^*)|c] &= \{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\} \\
\mathbb{E}[INT_{ref}(m^*)|c] &= \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a,m^*,c] \\
&\quad + \mathbb{E}[Y|a^*,m^*,c]dP(m|a^*,c) \\
\mathbb{E}[INT_{med}|c] &= \int \{\mathbb{E}[Y|a,m,c] \\
&\quad - \mathbb{E}[Y|a^*,m,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
\mathbb{E}[PIE|c] &= \int \mathbb{E}[Y|a^*,m,c]\{dP(m|a,c) - dP(m|a^*,c)\}
\end{aligned}$$

Proof:
The first equality is established by Robins (1986), the fourth by Pearl (2001), the third by VanderWeele (2013b), and proofs are likewise given in Section A.2. For the second equality we have

$$\begin{aligned}
\mathbb{E}[INT_{ref}(m^*)|c] &= E\left[\sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})1(M_{a^*} = m)|c\right] \\
&= \int_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]dP(M_{a^*} = m|c) \\
&= \int_m \{\mathbb{E}[Y_{am}|c] - \mathbb{E}[Y_{a^*m}|c] - \mathbb{E}[Y_{am^*}|c] \\
&\quad + \mathbb{E}[Y_{a^*m^*}|c]\}dP(M_{a^*} = m|c) \\
&= \int_m \{\mathbb{E}[Y_{am}|a,m,c] - \mathbb{E}[Y_{a^*m}|a^*,m,c] - \mathbb{E}[Y_{am^*}|a,m^*,c] \\
&\quad + \mathbb{E}[Y_{a^*m^*}|a^*,m^*,c]\}dP(M_{a^*} = m|a^*,c)
\end{aligned}$$

$$= \int_m \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] - \mathbb{E}[Y|a,m^*,c] + \mathbb{E}[Y|a^*,m^*,c]\} dP(M=m|a^*,c)$$

where the second equality follows by assumption (A2.4) and the fourth by assumptions (A2.1)–(A2.3). In fact, the other three equalities in Proposition 14.3 can be established in much the same way. ■

We can also interpret the terms in the decomposition in Proposition 14.2 causally under assumptions (A14.3)–(A14.4) alone, though the causal interpretation is slightly weaker.

Proposition 14.4 (VanderWeele, 2014):
Under assumptions (A2.1)–(A2.3) we have

$$\begin{aligned}
&\{\mathbb{E}[Y|a,m^*,c] - \mathbb{E}[Y|a^*,m^*,c]\} \\
&= \mathbb{E}[Y_{am^*} - Y_{a^*m^*}|c] \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c] \\
&- \mathbb{E}[Y|a,m^*,c] + \mathbb{E}[Y|a^*,m^*,c]\} dP(m|a^*,c) \\
&= \int \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c] dP(M_{a^*}|c) \\
&\quad \int \{\mathbb{E}[Y|a,m,c] - \mathbb{E}[Y|a^*,m,c]\}\{dP(m|a,c) - dP(m|a^*,c)\} \\
&= \int \mathbb{E}[Y_{am} - Y_{a^*m}|c]\{dP(M_a|c) - dP(M_{a^*}|c) \\
&\quad \int \mathbb{E}[Y|a^*,m,c]\{dP(m|a,c) - dP(m|a^*,c)\} \\
&= \int \mathbb{E}[Y_{a^*m}|c]\{dP(M_a|c) - dP(M_{a^*}|c)\}
\end{aligned}$$

Proof:
The first equality is established by Robins (1986), the second in the final four lines of the proof of Proportion 14.3 above, the third in VanderWeele (2013a) and the fourth, using slightly different notation by Didelez, et al. (2006). ■

Note we can also rewrite the right-hand side of the third equality as $\int \{\mathbb{E}[Y_{am} - Y_{a^*m}|c] - \mathbb{E}[Y_{am^*} - Y_{a^*m^*}|c]\}\{dP(M_a|c) - dP(M_{a^*}|c)$ and the right-hand side of the fourth equality as $\int \{\mathbb{E}[Y_{a^*m} - Y_{a^*m^*}|c]\}\{dP(M_a|c) - dP(M_{a^*}|c)\}$. The right-hand side of the equalities in Proposition 14.4 are causal quantities but rather than directly taking population averages of the four components of the decomposition, the effect of A and M on Y are integrated over the distribution of M under different exposure settings. As discussed in Section A.7.2 and also further below in Appendix 14.4, these effects can be interpreted as randomized interventional analogues of the four components of the decomposition. They only require assumptions (A2.1)–(A2.3) for identification (i.e., they do not require the

more controversial cross-world independence assumption (A2.4)), but the causal interpretation of these randomized interventional analogues is somewhat weaker.

A.14.2. Continuous Outcomes and Linear Regression Models

For Y and M continuous, under assumptions (A2.1)–(A2.4) and correct specification of the regression models for Y and M, we have

$$\begin{aligned}\mathbb{E}[Y|a,m,c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ \mathbb{E}[M|a,c] &= \beta_0 + \beta_1 a + \beta_2' c\end{aligned}$$

VanderWeele and Vansteelandt (2009) and VanderWeele (2013b) and Section A.2.2 showed that the average controlled direct effect, the pure indirect effect, and the mediated interaction conditional on covariates $C = c$ were given by

$$\begin{aligned}\mathbb{E}[CDE(m^*)|c] &= (\theta_1 + \theta_3 m^*)(a - a^*) \\ \mathbb{E}[PIE|c] &= (\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*) \\ \mathbb{E}[INT_{med}|c] &= \theta_3\beta_1(a - a^*)(a - a^*)\end{aligned}$$

They also showed that the pure direct effect was given by $\mathbb{E}[PDE|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$. The reference interaction is then given by difference between the pure direct effect and the controlled direct effect:

$$\begin{aligned}\mathbb{E}[INT_{ref}(m^*)|c] &= \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*) - (\theta_1 + \theta_3 m^*)(a - a^*) \\ &= \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c - m^*)\}(a - a^*)\end{aligned}$$

Standard errors for these expressions could be derived using the delta method along the lines of the derivations in VanderWeele and Vansteelandt (2009) as in Section A.2.2 or by using bootstrapping.

For Y continuous and M binary, under assumptions (A2.1)–(A2.4) and correct specification of the regression models for Y and M:

$$\begin{aligned}\mathbb{E}[Y|a,m,c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ \text{logit}\{P(M = 1|a,c)\} &= \beta_0 + \beta_1 a + \beta_2' c\end{aligned}$$

Valeri and VanderWeele (2013) show that the average controlled direct effect and the average pure indirect effect are given by

$$\begin{aligned}\mathbb{E}[CDE(m^*)|c] &= (\theta_1 + \theta_3 m^*)(a - a^*) \\ \mathbb{E}[PIE|c] &= (\theta_2 + \theta_3 a^*) \\ &\quad \times \left\{ \frac{\exp[\beta_0 + \beta_1 a + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{\exp[\beta_0 + \beta_1 a^* + \beta_2' c]}{1 + \exp[\beta_0 + \beta_1 a^* + \beta_2' c]} \right\}\end{aligned}$$

The reference interaction is given by the difference between the pure direct effect and the controlled direct effect, which were both given by Valeri and VanderWeele

(2013):

$$\begin{aligned}\mathbb{E}[INT_{ref}(m^*)|c] &= \{\theta_1(a-a^*)\} + \{\theta_3(a-a^*)\}\frac{\exp[\beta_0+\beta_1 a^*+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a^*+\beta_2' c]} \\ &\quad -(\theta_1+\theta_3 m^*)(a-a^*) \\ &= \theta_3(a-a^*)\left(\frac{\exp[\beta_0+\beta_1 a^*+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a^*+\beta_2' c]} - m^*\right)\end{aligned}$$

The mediated interaction is given by the difference between the total indirect effect and the pure indirect effect, which were also both given by Valeri and VanderWeele (2013):

$$\begin{aligned}&\mathbb{E}[INT_{med}|c] \\ &= (\theta_2+\theta_3 a)\left\{\frac{\exp[\beta_0+\beta_1 a+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a+\beta_2' c]} - \frac{\exp[\beta_0+\beta_1 a^*+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a^*+\beta_2' c]}\right\} \\ &\quad -(\theta_2+\theta_3 a^*)\left\{\frac{\exp[\beta_0+\beta_1 a+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a+\beta_2' c]} - \frac{\exp[\beta_0+\beta_1 a^*+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a^*+\beta_2' c]}\right\} \\ &= \theta_3(a-a^*)\left\{\frac{\exp[\beta_0+\beta_1 a+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a+\beta_2' c]} - \frac{\exp[\beta_0+\beta_1 a^*+\beta_2' c]}{1+\exp[\beta_0+\beta_1 a^*+\beta_2' c]}\right\}.\end{aligned}$$

A.14.3. Decomposition on a Ratio Scale and Logistic Regression Models

From Proposition 14.1, we have

$$\begin{aligned}Y_a - Y_{a^*} &= (Y_{am^*} - Y_{a^*m^*}) + \sum\nolimits_m (Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})1(M_{a^*} = m) \\ &\quad + \sum\nolimits_m (Y_{am} - Y_{a^*m})\{1(M_a = m) - 1(M_{a^*} = m)\} \\ &\quad + (Y_{a^*M_a} - Y_{a^*M_{a^*}})\end{aligned}$$

Taking expectations conditional on $C = c$ gives

$$\begin{aligned}\mathbb{E}(Y_a - Y_{a^*}|c) &= \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) + \sum\nolimits_m \\ &\quad \mathbb{E}[(Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*})1(M_{a^*} = m)|c] \\ &\quad + \sum\nolimits_m \mathbb{E}[(Y_{am} - Y_{a^*m})\{1(M_a = m) - 1(M_{a^*} = m)\}|c] \\ &\quad + \mathbb{E}(Y_{a^*M_a} - Y_{a^*M_{a^*}}|c)\end{aligned}$$

Under assumption (A2.4), this is

$$\mathbb{E}(Y_a - Y_{a^*}|c) = \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) + \sum\nolimits_m \mathbb{E}(Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c)$$

$$
\begin{aligned}
&P(M_{a^*} = m|c) \\
&+ \sum\nolimits_m \mathbb{E}(Y_{am} - Y_{a^*m}|c)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\
&+ \mathbb{E}(Y_{a^*M_a} - Y_{a^*M_{a^*}}|c)
\end{aligned}
$$

and dividing by $\mathbb{E}(Y_{a^*}|c)$ gives

$$RR_c^{TE} - 1 = \kappa\left[RR_c^{CDE}(m^*) - 1\right] + \kappa RR_c^{INT_{ref}}(m^*) + \kappa RR_c^{INT_{med}} + (RR_c^{PIE} - 1)$$

where $RR_c^{TE} = \frac{\mathbb{E}(Y_a|c)}{\mathbb{E}(Y_{a^*}|c)}$, $\kappa = \frac{\mathbb{E}(Y_{a^*m^*}|c)}{\mathbb{E}(Y_{a^*}|c)}$, and

$$
\begin{aligned}
RR_c^{CDE}(m^*) &= \frac{\mathbb{E}(Y_{am^*}|c)}{\mathbb{E}(Y_{a^*m^*}|c)} \\
RR_c^{INT_{ref}}(m^*) &= \sum\nolimits_m RERI(a^*, m^*)P(M_{a^*} = m|c) \\
RR_c^{INT_{med}} &= \sum\nolimits_m RERI(a^*, m^*)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\
RR_c^{PIE} &= \frac{\mathbb{E}(Y_{a^*M_a}|c)}{\mathbb{E}(Y_{a^*M_{a^*}}|c)}
\end{aligned}
$$

with $RERI(a^*, m^*) = \left(\frac{\mathbb{E}(Y_{am}|c)}{\mathbb{E}(Y_{a^*m^*}|c)} - \frac{\mathbb{E}(Y_{a^*m}|c)}{\mathbb{E}(Y_{a^*m^*}|c)} - \frac{\mathbb{E}(Y_{am^*}|c)}{\mathbb{E}(Y_{a^*m^*}|c)} + 1\right)$. Under assumptions (A2.1)–(A2.3) we also have $\mathbb{E}(Y_a|c) = \mathbb{E}(Y|a,c)$, $\mathbb{E}\ (Y_{am}|c) = \sum_m \mathbb{E}[Y|a,m,c]\ P(m|a,c)$ and thus $P(M_a = m|c) = P(M = m|a,c)$ and thus the right-hand side of the equalities above would be identified from the data. VanderWeele (2013b) also showed that $\kappa RR_c^{INT_{med}} = \kappa \sum_m RERI(a^*, m^*)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} = \left(\frac{\mathbb{E}[Y_{aM_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{aM_{a^*}}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} - \frac{\mathbb{E}[Y_{a^*M_a}|c]}{\mathbb{E}[Y_{a^*M_{a^*}}|c]} + 1\right)$ and called this latter term $RERI_{mediated}$.

Note also under assumption (A2.4), $(RR_c^{PIE} - 1)$ can be rewritten as

$$
\begin{aligned}
(RR_c^{PIE} - 1) &= \left(\frac{\mathbb{E}(Y_{a^*M_a}|c)}{\mathbb{E}(Y_{a^*}|c)} - \frac{\mathbb{E}(Y_{a^*}|c)}{\mathbb{E}(Y_{a^*}|c)}\right) \\
&= \frac{\kappa}{\mathbb{E}(Y_{a^*m^*}|c)}\{\mathbb{E}(Y_{a^*M_a}|c) - \mathbb{E}(Y_{a^*}|c)\} \\
&= \frac{\kappa}{\mathbb{E}(Y_{a^*m^*}|c)} \sum\nolimits_m \{\mathbb{E}[Y_{a^*m}|c] - \mathbb{E}[Y_{a^*m^*}|c]\}\{P(M_a = m|c) \\
&\quad - P(M_{a^*} = m|c)\} \\
&= \kappa \sum\nolimits_m \left(\frac{\mathbb{E}(Y_{a^*m}|c)}{\mathbb{E}(Y_{a^*m^*}|c)} - 1\right)\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\
&= \kappa \sum\nolimits_m \frac{\mathbb{E}(Y_{a^*m}|c)}{\mathbb{E}(Y_{a^*m^*}|c)}\{P(M_a = m|c) - P(M_{a^*} = m|c)\}
\end{aligned}
$$

The proportion attributable to each of the four components is then obtained by simply dividing each of the four components in the display equation above by their

sum as in Table 14.2. A similar decomposition could likewise be carried out on an additive scale using hazard ratios.

By similar arguments to those above but applied to Propositions 14.2 and 14.4, if assumption (A2.4) did not hold but assumptions (A2.1)–(A2.3) all did hold, we would have that $(RR_c^{TE} - 1)$ decomposed into the product of κ and the sum of

$$RR_c^{CDE}(m^*) - 1 = \frac{\mathbb{E}[Y|a, m^*, c]}{\mathbb{E}[Y|a^*, m^*, c]} - 1$$

$$\int RERI(a^*, m^*)\, dP(M_{a^*}|c)$$

$$= \int \left\{ \frac{\mathbb{E}[Y|a, m, c]}{\mathbb{E}[Y|a^*, m^*, c]} - \frac{\mathbb{E}[Y|a^*, m, c]}{\mathbb{E}[Y|a^*, m^*, c]} - \frac{\mathbb{E}[Y|a, m^*, c]}{\mathbb{E}[Y|a^*, m^*, c]} + 1 \right\} \times dP(m|a^*, c)$$

$$\int RERI(a^*, m^*)\{dP(M_a|c) - dP(M_{a^*}|c)$$

$$= \int \left\{ \frac{\mathbb{E}[Y|a, m, c]}{\mathbb{E}[Y|a^*, m^*, c]} - \frac{\mathbb{E}[Y|a^*, m, c]}{\mathbb{E}[Y|a^*, m^*, c]} \right\} \{dP(m|a, c) - dP(m|a^*, c)\}$$

$$\int \frac{\mathbb{E}[Y_{a^*m}|c]}{\mathbb{E}[Y_{a^*m^*}|c]}\{dP(M_a|c) - dP(M_{a^*}|c)\} = \int \frac{\mathbb{E}[Y|a^*, m, c]}{\mathbb{E}[Y|a^*, m^*, c]}\{dP(m|a, c) - dP(m|a^*, c)\}.$$

Suppose Y were binary and M continuous, that assumptions (A2.1)–(A2.4) held, that the outcome is rare, and that the following regressions were correctly specified:

$$\text{logit}(P(Y = 1|a, m, c)) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\mathbb{E}[M|a, c] = \beta_0 + \beta_1 a + \beta_2' c$$

with M normally distribution conditional on (A, C) with variance σ^2. Suppose that the outcome is rare so that odds ratios approximate risk ratios. VanderWeele and Vansteelandt (2010) derived expressions for the controlled direct effect, the pure indirect effect, and the pure direct effect, all on the risk ratio scale. The total effect, controlled direct effect, and pure indirect effect were given approximately by

$$RR_c^{TE} \approx \exp[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + \tfrac{1}{2}\theta_3^2\sigma^2(a^2 - a^{*2})]$$
$$RR_c^{CDE}(m^*) \approx \exp[(\theta_1 + \theta_3 m^*)(a - a^*)]$$
$$RR_c^{PIE} \approx \exp[(\theta_2\beta_1 + \theta_3\beta_1 a^*)(a - a^*)]$$

where the approximations (here and below) hold to the extent that the outcome is rare. We have that $\kappa = \frac{\mathbb{E}(Y_{a^*m^*}|c)}{\mathbb{E}(Y_{a^*}|c)}$ is given by

$$\begin{aligned}\kappa &= \frac{\mathbb{E}(Y_{a^*m^*}|c)}{\mathbb{E}(Y_{a^*}|c)} = \frac{\mathbb{E}[Y|a^*,m^*,c]}{\int \mathbb{E}[Y|a^*,m,c]dP(m|a^*,c)} \\ &\approx \frac{\exp(\theta_0+\theta_1a^*+\theta_2m^*+\theta_3a^*m^*+\theta_4'c)}{\exp\{\theta_0+\theta_1a^*+\theta_4'c\}\int \exp\{(\theta_2+\theta_3a^*)m\}\, dP(m|a^*,c)} \\ &= \frac{\exp(\theta_2m^*+\theta_3a^*m^*)}{\exp\{(\theta_2+\theta_3a^*)(\beta_0+\beta_1a^*+\beta_2'c)+\frac{1}{2}(\theta_2+\theta_3a^*)^2\sigma^2\}} \\ &= e^{\theta_2m^*+\theta_3a^*m^*-(\theta_2+\theta_3a^*)(\beta_0+\beta_1a^*+\beta_2'c)-\frac{1}{2}(\theta_2+\theta_3a^*)^2\sigma^2}\end{aligned}$$

We have

$$\begin{aligned}&\int \frac{\mathbb{E}[Y|a,m,c]}{\mathbb{E}[Y|a^*,m^*,c]}\, dP(m|a^\dagger,c) \\ &\quad\approx \int \exp(\theta_1a+\theta_2m+\theta_3am-\theta_1a^*-\theta_2m^*-\theta_3a^*m^*)\, dP(m|a^\dagger,c) \\ &\quad= \exp\{\theta_1(a-a^*)-\theta_2m^*-\theta_3a^*m^*\}\int \exp\{(\theta_2+\theta_3a)m\}dP(m|a^\dagger,c) \\ &\quad= \exp\{\theta_1(a-a^*)-\theta_2m^*-\theta_3a^*m^*\}\exp\{(\theta_2+\theta_3a)(\beta_0+\beta_1a^\dagger+\beta_2'c) \\ &\qquad+\tfrac{1}{2}(\theta_2+\theta_3a)^2\sigma^2\} \\ &\quad= e^{\theta_1(a-a^*)-\theta_2m^*-\theta_3a^*m^*+(\theta_2+\theta_3a)(\beta_0+\beta_1a^\dagger+\beta_2'c)+\frac{1}{2}(\theta_2+\theta_3a)^2\sigma^2}\end{aligned}$$

The reference interaction is thus given by

$$\begin{aligned}RR_c^{INT_{ref}}(m^*) &= \int \left\{\frac{\mathbb{E}[Y|a,m,c]}{\mathbb{E}[Y|a^*,m^*,c]} - \frac{\mathbb{E}[Y|a^*,m,c]}{\mathbb{E}[Y|a^*,m^*,c]} - \frac{\mathbb{E}[Y|a,m^*,c]}{\mathbb{E}[Y|a^*,m^*,c]} + 1\right\} \\ &\qquad \times dP(m|a^*,c) \\ &= e^{\theta_1(a-a^*)-\theta_2m^*-\theta_3a^*m^*+(\theta_2+\theta_3a)(\beta_0+\beta_1a^*+\beta_2'c)+\frac{1}{2}(\theta_2+\theta_3a)^2\sigma^2} \\ &\quad - e^{-\theta_2m^*-\theta_3a^*m^*+(\theta_2+\theta_3a^*)(\beta_0+\beta_1a^*+\beta_2'c)+\frac{1}{2}(\theta_2+\theta_3a^*)^2\sigma^2} \\ &\quad - e^{(\theta_1+\theta_3m^*)(a-a^*)} + 1\end{aligned}$$

and the component due to the reference interaction $\kappa RR_c^{INT_{ref}}(m^*)$ by

$$\begin{aligned}&e^{\{\theta_1+\theta_3(\beta_0+\beta_1a^*+\beta_2'c+\theta_2\sigma^2)\}(a-a^*)+\frac{1}{2}\theta_3^2\sigma^2(a^2-a^{*2})} - 1 \\ &\quad - e^{\theta_1(a-a^*)+\theta_2m^*+\theta_3am^*-(\theta_2+\theta_3a^*)(\beta_0+\beta_1a^*+\beta_2'c)-\frac{1}{2}(\theta_2+\theta_3a^*)^2\sigma^2} \\ &\quad + e^{\theta_2m^*+\theta_3a^*m^*-(\theta_2+\theta_3a^*)(\beta_0+\beta_1a^*+\beta_2'c)-\frac{1}{2}(\theta_2+\theta_3a^*)^2\sigma^2}\end{aligned}$$

The mediated interaction is given by

$$RR_c^{INT_{med}} = \int \{\frac{\mathbb{E}[Y|a,m,c]}{\mathbb{E}[Y|a^*,m^*,c]} - \frac{\mathbb{E}[Y|a^*,m,c]}{\mathbb{E}[Y|a^*,m^*,c]}\}\{dP(m|a,c) - dP(m|a^*,c)\}$$
$$\approx e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a)(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2}$$
$$- e^{-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}$$
$$- e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2}$$
$$+ e^{-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}$$

and the component due to the mediated interaction $\kappa RR_c^{INT_{med}}$ by

$$e^{\{\theta_1+\theta_2\beta_1+\theta_3(\beta_0+\beta_1 a^*+\beta_1 a+\beta_2' c+\theta_2\sigma^2)\}(a-a^*)+\frac{1}{2}\theta_3^2\sigma^2(a^2-a^{*2})}$$
$$- e^{(\theta_2\beta_1+\theta_3\beta_1 a^*)(a-a^*)} - e^{\{\theta_1+\theta_3(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2\sigma^2)\}(a-a^*)+\frac{1}{2}\theta_3^2\sigma^2(a^2-a^{*2})} + 1$$

We also have that the component due to controlled direct effect is

$$\kappa\left[RR_c^{CDE}(m^*) - 1\right] = \kappa[e^{(\theta_1+\theta_3 m^*)(a-a^*)} - 1]$$
$$= e^{\theta_1(a-a^*)+\theta_2 m^*+\theta_3 a m^*-(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a^*+\beta_2' c)-\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}$$
$$- e^{\theta_2 m^*+\theta_3 a^* m^*-(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a^*+\beta_2' c)-\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}$$

and the component due to the pure indirect effect is

$$(RR_c^{PIE} - 1) = \kappa \int_m \frac{\mathbb{E}(Y_{a^* m}|c)}{\mathbb{E}(Y_{a^* m^*}|c)}\{dP(m|a,c) - dP(m|a^*,c)\}$$
$$= \kappa\{e^{-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}$$
$$- e^{-\theta_2 m^*-\theta_3 a^* m^*+(\theta_2+\theta_3 a^*)(\beta_0+\beta_1 a^*+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^*)^2\sigma^2}\}$$
$$= e^{(\theta_2\beta_1+\theta_3\beta_1 a^*)(a-a^*)} - 1$$

Standard errors for these various expressions could be derived using the delta method along the lines of the derivations in the Online Appendix of VanderWeele and Vansteelandt (2010) and in Section A.2.2 or by using bootstrapping.

Suppose both Y and M were binary, that assumptions (A2.1)–(A2.4) held, that the outcome was rare, and that the following regressions were correctly specified:

$$\text{logit}\{P(Y = 1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$\text{logit}\{P(M = 1|a,c)\} = \beta_0 + \beta_1 a + \beta_2' c$$

Valeri and VanderWeele (2013) show that the average total effect, controlled direct effect and the average pure indirect effect conditional on $C = c$ are given approximately by

$$RR_c^{TE} \approx \frac{\exp(\theta_1 a)\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a)\}}{\exp(\theta_1 a^*)\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)\}}$$

$$RR_c^{CDE}(m^*) \approx \exp\{(\theta_1+\theta_3 m)(a-a^*)\}$$

$$RR_c^{PIE} \approx \frac{\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^*)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)\}}$$

where the approximations (here and below) hold to the extent that the outcome is rare. We have that $\kappa = \frac{\mathbb{E}(Y_{a^*m^*}|c)}{\mathbb{E}(Y_{a^*}|c)}$ is given by

$$\begin{aligned}
\kappa &= \frac{\mathbb{E}(Y_{a^*m^*}|c)}{\mathbb{E}(Y_{a^*}|c)} = \frac{\mathbb{E}[Y|a^*,m^*,c]}{\int \mathbb{E}[Y|a^*,m,c]dP(m|a^*,c)} \\
&\approx \frac{\exp(\theta_0+\theta_1 a^*+\theta_2 m^*+\theta_3 a^* m^*+\theta_4' c)}{\exp\{\theta_0+\theta_1 a^*+\theta_4' c\}\int \exp\{(\theta_2+\theta_3 a^*)m\}dP(m|a^*,c)} \\
&= \frac{\exp(\theta_2 m^*+\theta_3 a^* m^*)}{\frac{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*)}{1+\exp(\beta_0+\beta_1 a^*+\beta_2' c)}} \\
&= \frac{e^{\theta_2 m^*+\theta_3 a^* m^*}\{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\}}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}}
\end{aligned}$$

We also have

$$\begin{aligned}
&\int \frac{\mathbb{E}[Y|a,m,c]}{\mathbb{E}[Y|a^*,m^*,c]}\, dP(m|a^\dagger,c) \\
&\approx \int \exp(\theta_1 a+\theta_2 m+\theta_3 am-\theta_1 a^*-\theta_2 m^*-\theta_3 a^* m^*)\, dP(m|a^\dagger,c) \\
&= \exp\{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*\}\int \exp\{(\theta_2+\theta_3 a)m\}\, dP(m|a^\dagger,c) \\
&= \frac{e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*}}{1+e^{\beta_0+\beta_1 a^\dagger+\beta_2' c}}(1+e^{\beta_0+\beta_1 a^\dagger+\beta_2' c+\theta_2+\theta_3 a}) \\
&\quad\times\frac{e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*}(1+e^{\beta_0+\beta_1 a^\dagger+\beta_2' c+\theta_2+\theta_3 a})}{1+e^{\beta_0+\beta_1 a^\dagger+\beta_2' c}}
\end{aligned}$$

The reference interaction is thus given by

$$\begin{aligned}
RR_c^{INT_{ref}}(m^*) &= \int\left\{\frac{\mathbb{E}[Y|a,m,c]}{\mathbb{E}[Y|a^*,m^*,c]}-\frac{\mathbb{E}[Y|a^*,m,c]}{\mathbb{E}[Y|a^*,m^*,c]}-\frac{\mathbb{E}[Y|a,m^*,c]}{\mathbb{E}[Y|a^*,m^*,c]}+1\right\} \\
&\quad dP(m|a^*,c) \\
&= \frac{e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*}(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a})}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}} \\
&\quad -\frac{e^{-\theta_2 m^*-\theta_3 a^* m^*}(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*})}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}} \\
&\quad -e^{(\theta_1+\theta_3 m^*)(a-a^*)}+1
\end{aligned}$$

and the component due to the reference interaction $\kappa RR_c^{INT_{ref}}(m^*)$ by

$$
\begin{aligned}
&= \frac{e^{\theta_1(a-a^*)}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a}\right)}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}} - 1 \\
&\quad - \frac{e^{\theta_1(a-a^*)+\theta_2 m^*+\theta_3 a m^*}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\right)}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}} e^{(\theta_1+\theta_3 m^*)(a-a^*)} \\
&\quad + \frac{e^{\theta_2 m^*+\theta_3 a^* m^*}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\right)}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}}
\end{aligned}
$$

The mediated interaction is given by

$$
\begin{aligned}
RR_c^{INT_{med}} &= \int \left\{ \frac{\mathbb{E}[Y|a,m,c]}{\mathbb{E}[Y|a^*,m^*,c]} - \frac{\mathbb{E}[Y|a^*,m,c]}{\mathbb{E}[Y|a^*,m^*,c]} \right\} \{dP(m|a,c) - dP(m|a^*,c)\} \\
&= \frac{e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*}\left(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a}\right)}{1+e^{\beta_0+\beta_1 a+\beta_2' c}} \\
&\quad - \frac{e^{-\theta_2 m^*-\theta_3 a^* m^*}\left(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^*}\right)}{1+e^{\beta_0+\beta_1 a+\beta_2' c}} \\
&\quad - \frac{e^{\theta_1(a-a^*)-\theta_2 m^*-\theta_3 a^* m^*}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a}\right)}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}} \\
&\quad + \frac{e^{-\theta_2 m^*-\theta_3 a^* m^*}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}\right)}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c}}
\end{aligned}
$$

and the component due to the mediated interaction $\kappa RR_c^{INT_{med}}$ by

$$
\begin{aligned}
&= \frac{e^{\theta_1(a-a^*)}\left(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a}\right)\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\right)}{\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}\right)\left(1+e^{\beta_0+\beta_1 a+\beta_2' c}\right)} \\
&\quad - \frac{\left(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^*}\right)\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\right)}{\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}\right)\left(1+e^{\beta_0+\beta_1 a+\beta_2' c}\right)} \\
&\quad - \frac{e^{\theta_1(a-a^*)}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a}\right)}{\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}\right)} + 1
\end{aligned}
$$

We also have that the component due to controlled direct effect is

$$
\begin{aligned}
\kappa\left[RR_c^{CDE}(m^*) - 1\right] &= \kappa\left[e^{(\theta_1+\theta_3 m^*)(a-a^*)} - 1\right] \\
&= \frac{e^{\theta_1(a-a^*)+\theta_2 m^*+\theta_3 a m^*}\left(1+e^{\beta_0+\beta_1 a^*+\beta_2' c}\right)}{1+e^{\beta_0+\beta_1 a^*+\beta_2' c+\theta_2+\theta_3 a^*}}
\end{aligned}
$$

$$-\frac{e^{\theta_2 m^* + \theta_3 a^* m^*}(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*}}$$

and the component due to the pure indirect effect is

$$\begin{aligned}
&\kappa \int_m \frac{\mathbb{E}(Y_{a^* m}|c)}{\mathbb{E}(Y_{a^* m^*}|c)} \{dP(m|a,c) - dP(m|a^*,c)\} \\
&= \kappa \left(\frac{e^{-\theta_2 m^* - \theta_3 a^* m^*}(1 + e^{\beta_0 + \beta_1 a + \beta_2' c + \theta_2 + \theta_3 a^*})}{1 + e^{\beta_0 + \beta_1 a + \beta_2' c}} \right. \\
&\quad \left. - \frac{e^{-\theta_2 m^* - \theta_3 a^* m^*}(1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*})}{1 + e^{\beta_0 + \beta_1 a^* + \beta_2' c}} \right) \\
&= \frac{\{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)\}\{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c + \theta_2 + \theta_3 a^*)\}}{\{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c)\}\{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 + \theta_3 a^*)\}} - 1
\end{aligned}$$

Standard errors for these expressions could be derived using the delta method along the lines of the derivations in the Online Appendix of Valeri and VanderWeele (2013) and Section A.2.2 or by using bootstrapping.

A.14.4. Decomposition in the Presence of an Exposure-Induced Mediator–Outcome Confounder

Consider a setting in which there is a variable L that is affected by exposure A and in turn affects both M and Y as in Figure 14.2. Although several of the components of the four-way decomposition are not identified in this setting, alternative effects that randomly set M to a value chosen from the distribution of a particular exposure level can be identified. The discussion here will give a randomized interventional interpretation to Proposition 14.4 in the text and extend that result to settings such as Figure 14.2 in which there is a mediator–outcome confounder affected by the exposure.

Let $G_{a|c}$ denote a random draw from the distribution of the mediator amongst those with exposure status a conditional on $C = c$. Let a and a^* be two values of the exposure, for example, for binary exposure we may have $a = 1$ and $a^* = 0$. As in VanderWeele (2013b), the effect $\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{aG_{a^*|c}}|c)$ is then the effect on the outcome of randomly assigning an individual who is given the exposure to a value of the mediator from the distribution of the mediator amongst those given exposure versus no exposure, conditional on covariates; this is a randomized interventional analogue of the pure indirect effect. Next consider the effect $\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^* G_{a^*|c}}|c)$; this is a direct effect comparing exposure versus no exposure with the mediator in both cases randomly drawn from the distribution of the population when given the absence of exposure, conditional on covariates; this is a randomized interventional analogue of the pure direct effect. Finally, the effect $\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^* G_{a^*|c}}|c)$ compares the expected outcome when having the exposure with the mediator randomly drawn from the distribution of the population

when given the exposure, conditional on covariates to the expected outcome when not having the exposure with the mediator randomly drawn from the distribution of the population when not exposed, conditional on covariates. With effects thus defined we have the decomposition $\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c) = \{\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{aG_{a^*|c}}|c)\} + \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\}$ so that the total effect decomposes into the sum of the effect through the mediator and the direct effect. These effects arise from randomly choosing for each individual a value of the mediator from the distribution of the mediator amongst all of those with a particular exposure.

We might further decompose this as follows:

$$\begin{aligned}\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c) = {} & \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} \\ & + \{\mathbb{E}(Y_{a^*G_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} \\ & + [\{\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a|c}}|c)\} - \{\mathbb{E}(Y_{aG_{a^*|c}}|c) \\ & - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\}]\end{aligned}$$

where the first term in the decomposition is the randomized intervention analogue of the pure direct effect, the second is the randomized intervention analogue of the pure indirect effect, and the third is the difference between the randomized intervention analogue of the total direct effect and the pure direct effect. As shown in VanderWeele (2013b) and Section A.7, this third term has the interpretation of an interaction. We have that

$$\begin{aligned}&\{\mathbb{E}(Y_{aG_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a|c}}|c)\} - \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} \\ &\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|G_{a|c} = m, c]P(G_{a|c} = m|c) \\ &\qquad - \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|G_{a^*|c} = m, c]P(G_{a^*|c} = m|c) \\ &\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]P(M_a = m|c) - \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|c]P(M_{a^*} = m|c) \\ &\quad = \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\}\end{aligned}$$

where m^* is an arbitrary value of M. We have the three-way decomposition given in VanderWeele (2013b) and Section A.7. Moreover, for the analogue of the pure direct effect we have

$$\begin{aligned}&\{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c)\} \\ &\quad = \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) + \{\mathbb{E}(Y_{aG_{a^*|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c) - \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c)\} \\ &\quad = \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) + \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m}|G_{a^*|c} = m, c]P(G_{a^*|c} = m|c) \\ &\qquad - \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) \\ &\quad = \mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) + \sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]P(M_{a^*} = m|c)\end{aligned}$$

that is, the analogue of the pure direct effect is the sum of a controlled direct effect and the reference interaction term, $\sum_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]P(M_{a^*} =$

$m|c$). We thus have a randomized interventional analogue of the four-way decomposition.

To identify these effects, the following conditions suffice: assumptions (A2.1) $Y_{am} \perp\!\!\!\perp A|C$ and (A2.3) $M_a \perp\!\!\!\perp A|C$ above, that conditional on C there is no unmeasured exposure–outcome or exposure–mediator confounding, along with an assumption that (A14.1) $Y_{am} \perp\!\!\!\perp M|\{A,C,L\}$—that is, that conditional on (A,C,L), there is no unmeasured confounding of the mediator–outcome relationship. These three assumptions would hold in the causal diagram in Figure 14.2. Under the three assumptions, each of these component are identified from data and it follows from the g-formula (Robins, 1986) that

$$\begin{aligned}\mathbb{E}(Y_{am^*} - Y_{a^*m^*}|c) &= \sum\nolimits_l \{\mathbb{E}[Y|a,l,m^*,c]P(l|a,c) \\ &\quad -\mathbb{E}[Y|a^*,l,m^*,c]P(l|a^*,c)\} \\ \mathbb{E}(Y_{a^*G_{a|c}}|c) - \mathbb{E}(Y_{a^*G_{a^*|c}}|c) &= \sum\nolimits_{l,m} \mathbb{E}[Y|a^*,l,m,c]P(l|a^*,c)\{P(m|a,c) \\ &\quad -P(m|a^*,c)\}\end{aligned}$$

$$\begin{aligned}&\sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_a = m|c) - P(M_{a^*} = m|c)\} \\ &= \sum\nolimits_{l,m} \{\mathbb{E}[Y|a,l,m,c]P(l|a,c) - \mathbb{E}[Y|a^*,l,m,c]P(l|a^*,c)\}\{P(m|a,c) - P(m|a^*,c)\}\end{aligned}$$

and

$$\begin{aligned}&\sum\nolimits_m \mathbb{E}[Y_{am} - Y_{a^*m} - Y_{am^*} + Y_{a^*m^*}|c]\{P(M_{a^*} = m|c)\} \\ &\quad = \sum\nolimits_{l,m} \{\mathbb{E}[Y|a,l,m,c]P(l|a,c) - \mathbb{E}[Y|a^*,l,m,c]P(l|a^*,c) \\ &\qquad -\mathbb{E}[Y|a,l,m^*,c]P(l|a,c) + \mathbb{E}[Y|a^*,l,m^*,c]P(l|a^*,c)\}P(m|a^*,c)\end{aligned}$$

Thus a randomized interventional analogue of the four-way decomposition holds and its components can be identified under assumptions (A2.1), (A14.1), and (A2.3). When Figure 4.1 is in fact the underlying causal diagram so the L can be chosen to be empty, then assumption (A14.1) simply becomes assumption (A2.2) in the text. And the identification results here simply reduce to those of Proposition 14.4 in the text. As in Proposition 14.4 in the text, the randomized interventional interpretation does not require the more controversial cross-world independence assumption, assumption (A2.4).

A.15. SOCIAL INTERACTIONS AND SPILLOVER EFFECTS

A.15.1. Notation and Definitions

Suppose that in a particular study there are K households indexed by $i = 1,\ldots,K$ in which there are two people under study per household (e.g., husband and wife) indexed by $j = 1,2$. Initially we will assume that the two persons are distinguishable from one another (e.g., $j = 1$ denotes the wife and $j = 2$ denotes the husband).

We let A_{ij} denote the exposure status for person j in household i. We let Y_{ij} denote the infection status of person j in household i after some suitable follow-up. We let $Y_{ij}(a_{i1}, a_{i2})$ denote the counterfactual outcome for person j in household i if the two people in that household i had (possibly contrary to fact) vaccine status of (a_{i1}, a_{i2}). Note that under this counterfactual or "potential outcomes" notation, the potential outcome for person 1, $Y_{i1}(a_{i1}, a_{i2})$, depends on the vaccine status of both persons. This allows for the possibility that the exposure status of one person affects the outcomes of another, sometimes referred to as interference or a spillover effect. Most literature in causal inference makes a "no-interference" assumption (Cox, 1958) that one person's outcome does not depend on the exposure of other people. This no-interference assumption is part of Rubin's so-called Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1986), the other part being the no-multiple-versions-of-treatment assumption discussed in Section A.7.1 In the current context the no-interference assumption would imply that $Y_{i1}(a_{i1}, a_{i2}) = Y_{i1}(a_{i1})$ and $Y_{i2}(a_{i1}, a_{i2}) = Y_{i2}(a_{i2})$ so that each person's outcome depends only on his or her own exposure status. We will allow for such interference here, but will assume that the vaccine status of people in one household do not affect the outcomes of those in other households. This assumption is sometimes referred to as "partial interference" (Sobel, 2006; Hudgens and Halloran, 2008).

A.15.2. Basic Spillover and Individual/Direct Effects

We can define the individual/direct effect for individual 1 while individual 2's vaccine fixed at a_{i2} as $\mathbb{E}[Y_{i1}(1, a_{i2}) - Y_{i1}(0, a_{i2})]$. We can define the spillover/indirect effect of individual 2's exposure on individual 1's outcome with individual 1's exposure fixed at a_{i1} by $\mathbb{E}[Y_{i1}(a_{i1}, 1) - Y_{i1}(a_{i1}, 0)]$. We can also define the "total" effect on individual 1 of giving both versus neither the exposure as $\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(0,0)]$. We also have that this "total effect" is equal to the sum of a spillover and a direct/individual effect as above. For example, $\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(0,0)] = \mathbb{E}[Y_{i1}(1,1) - Y_{i1}(1,0)] + \mathbb{E}[Y_{i1}(1,0) - Y_{i1}(0,0)]$. Analogous effect could likewise be defined for individual 2. Under the assumption that the vaccine/exposure of both individual is randomized, all of these effects are identified since:

$$\mathbb{E}[Y_{i1}(1, a_{i2}) - Y_{i1}(0, a_{i2})] = \mathbb{E}[Y_{i1}|A_{i1} = 1, A_{i2} = a_{i2}] - \mathbb{E}[Y_{i1}|A_{i1} = 0, A_{i2} = a_{i2}]$$
$$\mathbb{E}[Y_{i1}(a_{i1}, 1) - Y_{i1}(a_{i1}, 0)] = \mathbb{E}[Y_{i1}|A_{i1} = a_{i1}, A_{i2} = 1] - \mathbb{E}[Y_{i1}|A_{i1} = a_{i1}, A_{i2} = 0]$$
$$\mathbb{E}[Y_{i1}(1,1) - Y_{i1}(0,0)] = \mathbb{E}[Y_{i1}|A_{i1} = 1, A_{i2} = 1] - \mathbb{E}[Y_{i1}|A_{i1} = 0, A_{i2} = 0]$$

We could also average over the effects for individuals 1 and 2 (Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012). Similar definitions can be given when there are more than two individuals per cluster though the definitions can become somewhat more subtle (Hudgens and Halloran, 2008; VanderWeele and Tchetgen Tchetgen, 2011b; Tchetgen Tchetgen and VanderWeele, 2012) as described below.

A.15.3. Infectiousness Effect

We now consider a setting in which only person 1 is randomized to vaccination. We thus implicitly condition on $A_{i2} = 0$ throughout this section of the appendix. As discussed in the text, the crude estimator for the infectiousness effect (on the risk difference scale) might be taken as

$$\mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] - \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1] \tag{14.1a}$$

This is a comparison of the infection rates for individual 2 in the subgroup in which individual 1 was vaccinated and infected versus in the subgroup in which individual 1 was unvaccinated and infected. Although this is an appealing intuitive measure for trying to capture the extent to which the vaccine may render those infected less infectious, which may in turn prevent the second individual from being infected (i.e., the "infectiousness effect"), the measure is subject to selection bias. The vaccine status for individual 1, A_{i1}, is randomized, but conditioning on a variable that occurs after treatment (namely, the infection status of individual 1) in effect breaks randomization. We instead define the causal infectiousness effect as (VanderWeele and Tchetgen Tchetgen, 2011a; Halloran and Hudgens, 2012a,b):

$$\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \tag{14.2}$$

This contrast compares the infection status for individual 2 if individual 1 was vaccinated, $Y_{i2}(1,0)$, versus unvaccinated, $Y_{i2}(0,0)$, but only among the subset of households for whom individual 1 would have been infected irrespective of whether or not individual 1 was vaccinated, that is, $Y_{i1}(1,0) = Y_{i1}(0,0) = 1$. Such a subgroup is sometimes referred to as a principal stratum (Frangakis and Rubin, 2002). Because we are considering only the subset of households for whom individual 1 would have been infected irrespective of whether or not individual 1 was vaccinated, individual 2 is exposed to the infection of individual 1 and thus any effect of the vaccine ought to occur through changing the infectiousness. We might therefore take the contrast in (14.2) as a formal causal contrast corresponding to the "infectiousness effect" in the vaccine literature.

It is not identified from the data, but bounds can be obtained under the following assumptions. We first assume the vaccine will never be the cause of the infection; that is, there may be persons who would be infected irrespective of vaccination status or who would not be infected irrespective of vaccination status or who would be infected if unvaccinated and not infected if vaccinated, but there is no one who would be infected if vaccinated and uninfected if unvaccinated. Stated more formally, we assume (A15.1) for all i, $Y_{i1}(1,0) \leq Y_{i1}(0,0)$. Our second assumptions is that (A15.2) $\mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1] \leq \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1} = 1]$. Assumption (A15.2) states that the average infection rate for individual 2 if both individuals 1 and 2 were unvaccinated would be lower in the subgroup of households for which individual 1 would be infected and unvaccinated than in the subgroup of households for which individual 1 would be infected and vaccinated. The assumption is arguably reasonable insofar as the subgroup for which individual 1 was vaccinated and infected is likely less healthy (or the infection more

virulent) than the subgroup for which individual 1 was unvaccinated and infected; thus, under the scenario in which both people are unvaccinated, individual 2 is more likely to be infected in the first subgroup than in the second.

Proposition 15.1 (VanderWeele and Tchetgen Tchetgen, 2011a):
Under assumptions (A15.1) and (A15.2), the crude contrast in (15.1) is conservative for the causal infectiousness effect in (15.2) in that

$$\begin{aligned}&\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1]\\ &\quad \leq \mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] - \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1].\end{aligned}$$

Proof:
Under assumption (A15.1) we have that

$$\begin{aligned}&\mathbb{E}[Y_{i2}(1,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}(1,0)|A_{i1} = 1, Y_{i1}(1,0) = Y_{i1}(0,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}(1,0)|A_{i1} = 1, Y_{i1}(1,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1]\end{aligned}$$

where the first equality holds by randomization of A_{i1}, the second by assumption (A15.1), and the third follows because if $A_{i1} = 1, A_{i2} = 0$, then $Y_{i2}(1,0) = Y_{i2}$ and $Y_{i1}(1,0) = Y_{i1}$ (i.e., the "consistency" assumption). Also under assumption (A15.1), we have

$$\begin{aligned}&\mathbb{E}[Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1}(1,0) = Y_{i1}(0,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1}(1,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1}(1,0) = 1] + \{\mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1]\\ &\qquad - \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1]\}\\ &\quad = \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1} = 1] + \{\mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1]\\ &\qquad - \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1]\}\\ &\quad = \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1] + \{\mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1} = 1]\\ &\qquad - \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1]\}\end{aligned}$$

where the first equality holds by randomization of A_{i1} and the second by assumption (A15.1); the third follows by adding and subtracting, and the fourth holds because if $A_{i1} = 1, A_{i2} = 0$, then $Y_{i1}(1,0) = Y_{i1}$ and if $A_{i1} = 0, A_{i2} = 0$, then $Y_{i2}(0,0) = Y_{i2}$ (i.e., the "consistency" assumption). We thus have

$$\begin{aligned}&\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1]\\ &\quad = \mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] - \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1]\\ &\qquad + \{\mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1] - \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1} = 1]\}\end{aligned}$$

Under assumption (A15.2), we then have

$$\begin{aligned}&\mathbb{E}[Y_{i2}(1,0) - Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ &\quad \leq \mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] - \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1]\end{aligned}$$

This completes the proof. ■

Other measures of effect might also be of interest. For notational convenience in this section we define the following:

$$\begin{aligned}p_v &= \mathbb{E}[Y_{i2}(1,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ p_u &= \mathbb{E}[Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ p_1 &= \mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] \\ p_0 &= \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1]\end{aligned}$$

The causal infectiousness effect on the risk difference scale is then just $p_v - p_u$ and Proposition 15.1 simply states that $p_v - p_u \leq p_1 - p_0$. We might similarly be interested in the causal infectiousness effect on the risk ratio scale, p_v/p_u or on the odds ratio scale, $p_v(1-p_u)/\{p_u(1-p_v)\}$. We might further be interested in what one might refer to as the causal vaccine efficacy infectiousness effect, $1 - p_v/p_u$. Under assumptions (A15.1) and (A15.2), for each of these additional measures of effect, the crude estimator is also conservative for the true causal infectiousness effect.

Proposition 15.2 (VanderWeele and Tchetgen Tchetgen, 2011a):
Under assumptions (A15.1) and (A15.2) we have

$$\begin{gathered}p_v/p_u \leq p_1/p_0 \\ p_v(1-p_u)/\{p_u(1-p_v)\} \leq p_1(1-p_0)/\{p_0(1-p_1)\} \\ \text{and } 1 - p_v/p_u \geq 1 - p_1/p_0\end{gathered}$$

Proof:
It was shown in the proof of Proposition 15.1 that under assumption (A15.1) we have

$$p_v = \mathbb{E}[Y_{i2}(1,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] = \mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1] = p_1$$

and

$$\begin{aligned}&\mathbb{E}[Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \\ &\quad = \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1] + \{\mathbb{E}[Y_{i2}(0,0)|A_{i1} = 1, Y_{i1} = 1] \\ &\qquad - \mathbb{E}[Y_{i2}(0,0)|A_{i1} = 0, Y_{i1} = 1]\}\end{aligned}$$

so that under assumptions (A15.1) and (A15.2) we have

$$p_u = \mathbb{E}[Y_{i2}(0,0)|Y_{i1}(1,0) = Y_{i1}(0,0) = 1] \geq \mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1] = p_0$$

By Proposition 15.1, $p_v - p_u \leq p_1 - p_0$. Since $p_u \geq p_0$ we thus also have, by dividing, that $(p_v - p_u)/p_u \leq (p_1 - p_0)/p_0$, that is, $p_v/p_u - 1 \leq p_1/p_0 - 1$. Multiplying this

inequality by -1 gives the result for the vaccine efficacy measure. Moreover, from $p_v/p_u - 1 \leq p_1/p_0 - 1$ we immediately have $p_v/p_u \leq p_1/p_0$, which gives the result for the risk ratio measure. Because, under assumption (A15.1), $p_v = p_1$ we thus also have $p_v/\{p_u(1-p_v)\} \leq p_1/\{p_0(1-p_v)\}$, and, because, under assumptions (A15.1) and (A15.2), $p_u \geq p_0$ and so $(1-p_u) \leq (1-p_0)$, we also have $p_v(1-p_u)/\{p_u(1-p_v)\} \leq p_1(1-p_0)/\{p_0(1-p_1)\}$, which gives the result for the odds ratio measure. ■

A.15.4. Contagion Versus Infectiousness Effects

Here we give a formal statement of the identification assumption (A15.3)–(A15.6) in the text, provide nonparametric empirical expressions for the contagion and unconditional infectiousness effects when they are identified, and derive closed-form expressions for these when logistic or log-linear regression models are used to model the probabilities of infection and provide a sensitivity analysis technique when the identification assumptions are violated (cf. VanderWeele et al., 2012d). Most of this is accomplished by noting an analytic relation between the contagion and infectiousness effects defined in the text and natural direct and indirect effects in mediation analysis from Chapter 2. In mediation analysis, interest lies in assessing the extent to which the effect of an exposure A on outcome Y is mediated by some intermediate M. If we take the exposure as person 1's vaccine status, the mediator as person 1's infection status, and the outcome as person 2's infection status, then the contagion and unconditional infectiousness effects defined in this section correspond to the "total" natural direct effect and the "pure" natural indirect effect in mediation.

We will let $j = 1$ denote the individual who may or may not be vaccinated and $j = 2$ the individual who is always unvaccinated. We will assume that only person 1, not person 2, can be infected from outside the household; person 2 can be infected only by person 1. Thus if $Y_{i1}(a_{i1}, a_{i2}) = 0$, then $Y_{i2}(a_{i1}, a_{i2}) = 0$. We assume that in addition to potentially intervening to give person 1 the vaccine, we could also, at least hypothetically, intervene to infect or not infect person 1. We let $Y_{i2}(a_{i1}, a_{i2}, y_{i1})$ denote the infection status of person 2 if we would set the vaccine status of person 1 and person 2 to a_{i1} and a_{i2} and set the infection status of person 1 to y_{i1}.

The assumption that individual 2 is always unvaccinated allows a simplified notation. Counterfactuals $Y_{i1}(a_{i1}, a_{i2})$ and $Y_{i2}(a_{i1}, a_{i2})$ can be written as $Y_{i1}(a_{i1}) := Y_{i1}(a_{i1}, 0)$ and $Y_{i2}(a_{i1}) := Y_{i2}(a_{i1}, 0)$. We are still assuming interference/spillover in that the vaccine of person 1 affects the outcome of person 2. This simple setting in which person 2 always remains unvaccinated also allows us to rewrite the counterfactual $Y_{i2}(a_{i1}, a_{i2}, y_{i1})$ as $Y_{i2}(a_{i1}, y_{i1}) := Y_{i2}(a_{i1}, 0, y_{i1})$. We thus consider counterfactuals of the form $Y_{i1}(a_{i1})$, $Y_{i2}(a_{i1})$, and $Y_{i2}(a_{i1}, y_{i1})$. The direct effect of person 1's vaccine on person 1's outcome is $\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)]$; the indirect effect of person 1's vaccine on person 2's outcome is simply $\mathbb{E}[Y_{i2}(1) - Y_{i2}(0)]$. The contagion effect is defined as

$$\mathbb{E}[Y_{i2}(0, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(0))]$$

and the (unconditional) infectiousness effect is defined as

$$\mathbb{E}[Y_{i2}(1, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(1))]$$

We can decompose an indirect effect into a contagion and an infectiousness effect by taking the indirect effect and adding and subtracting the term $\mathbb{E}[Y_{i2}(0, Y_{i1}(1)]]$:

$$\begin{aligned}\mathbb{E}[Y_{i2}(1) - Y_{i2}(0)] &= \mathbb{E}[Y_{i2}(1, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(0))] \\ &= \mathbb{E}[Y_{i2}(1, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(1))] \\ &\quad + \mathbb{E}[Y_{i2}(0, Y_{i1}(1)) - Y_{i2}(0, Y_{i1}(0))]\end{aligned}$$

The indirect effect on the risk ratio and odds-ratio scale could be defined as $\frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]}$ or $\frac{\mathbb{E}[Y_{i2}(1)]/\{1-\mathbb{E}[Y_{i2}(1)]\}}{\mathbb{E}[Y_{i2}(0)]/\{1-\mathbb{E}[Y_{i2}(0)]\}}$. The decomposition for the risk ratio is

$$\frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]} = \frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]} \times \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}$$

Similar decomposition holds for odds-ratio measures. Similar definitions and a somewhat analogous decomposition holds with a vaccine efficacy measure. The vaccine efficacy measure for the indirect effect would be defined as

$$VE_{indirect} = 1 - \frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]}$$

We might likewise define vaccine efficacy for the contagion effect and infectiousness effect as

$$VE_{cont} = 1 - \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}$$

$$VE_{\text{inf}} = 1 - \frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}$$

Some algebra gives

$$\begin{aligned}1 - \frac{\mathbb{E}[Y_{i2}(1)]}{\mathbb{E}[Y_{i2}(0)]} &= \left(1 - \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}\right) \\ &+ \frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]} \left(1 - \frac{\mathbb{E}[Y_{i2}(1, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}\right)\end{aligned}$$

and we thus have

$$VE_{indirect} = VE_{cont} + \left(\frac{\mathbb{E}[Y_{i2}(0, Y_{i1}(1))]}{\mathbb{E}[Y_{i2}(0, Y_{i1}(0))]}\right) VE_{\text{inf}}$$

In counterfactual notation, identification assumptions (A15.3)–(A15.6) in the text can be formally stated as

(A15.3) $Y_{i2}(a_{i1}, y_{i1}) \perp\!\!\!\perp A_{i1} | C_i$

(A15.4) $Y_{i2}(a_{i1},y_{i1}) \perp\!\!\!\perp Y_{i1}|(C_i,A_{i1})$
(A15.5) $Y_{i1}(a_{i1}) \perp\!\!\!\perp A_{i1}|C_i$
(A15.6) $Y_{i2}(a_{i1},y_{i1}) \perp\!\!\!\perp Y_{i1}(a^*_{i1})|C_i$

where a^*_{i1} is simply a different value of A_{i1} than a_{i1}. Drawing on the analogy from mediation analysis, the interpretation of (A15.3)–(15.6) above is essentially that given in the text.

Assumptions (A15.3) and (15.5) will hold if A_{i1}, the vaccine status of person 1, is randomized. Assumptions (A15.3) and (A15.5) are substantial and would have be to determined on subject matter grounds. Under assumptions (A15.3)–(A15.5), the contagion and infectiousness effects are identified from the vaccine trial data. To see this, note that

$$\begin{aligned}
\mathbb{E}[Y_{i2}(a_{i1},Y_{i1}(a^*_{i1}))|c] &= \sum\nolimits_{y_1} \mathbb{E}[Y_{i2}(a_{i1},y_1)|Y_{i1}(a^*_{i1})=y_1,c]P(Y_{i1}(a^*_{i1})=y_1|c) \\
&= \sum\nolimits_{y_1} \mathbb{E}[Y_{i2}(a_{i1},y_1)|c]P(Y_{i1}(a^*_{i1})=y_1|c) \\
&= \sum\nolimits_{y_1} \mathbb{E}[Y_{i2}(a_{i1},y_1)|a_{i1},c]P(Y_{i1}(a^*_{i1})=y_1|a^*_{i1},c) \\
&= \sum\nolimits_{y_1} \mathbb{E}[Y_{i2}(a_{i1},y_1)|a_{i1},y_1,c]P(Y_{i1}(a^*_{i1})=y_1|a^*_{i1},c) \\
&= \sum\nolimits_{y_1} \mathbb{E}[Y_{i2}|a_{i1},y_1,c]P(Y_{i1}=y_1|a^*_{i1},c)
\end{aligned}$$

where the first equality holds by iterated expectations, the second by assumption (A15.6), the third by assumptions (A15.3) and (A15.5), and the fourth by assumption (A15.4). The final expression is given in terms of the observed data. If we first let $a_{i1}=0,a^*_{i1}=1$ and then $a_{i1}=0,a^*_{i1}=0$, the contagion effect conditional on C_i is given by

$$\begin{aligned}
&\mathbb{E}[Y_{i2}(0,Y_{i1}(1)) - Y_{i2}(0,Y_{i1}(0))|c] \\
&\quad = \sum\nolimits_{y_1} \mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]\{P(Y_{i1}=y_1|A_{i1}=1,c) - P(Y_{i1}=y_1|A_{i1}=0,c)\}
\end{aligned}$$

If we first let $a_{i1}=1,a^*_{i1}=1$ and then $a_{i1}=0,a^*_{i1}=1$, the infectiousness effect is given by

$$\begin{aligned}
&\mathbb{E}[Y_{i2}(1,Y_{i1}(1)) - Y_{i2}(0,Y_{i1}(1)|c] \\
&\quad = \sum\nolimits_{y_1} \{\mathbb{E}[Y_{i2}|A_{i1}=1,y_1,c] - \mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]\}P(Y_{i1}=y_1|A_{i1}=1,c)
\end{aligned}$$

The contagion effect then contrasts the observed expectation $\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]$ as standardized by the distribution of the infection status of person 1 among the households with person 1 vaccinated versus unvaccinated. The infectiousness effect is the observed expectation contrast $\mathbb{E}[Y_{i2}|A_{i1}=1,y_1,c] - \mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]$ standardized by the distribution of the infection status of person 1 among the households with person 1 vaccinated.

Likewise on a risk ratio scale the contagion effect is given by

$$\frac{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))|c]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(0))|c]} = \frac{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]P(Y_{i1}=y_1|A_{i1}=1,c)}{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]P(Y_{i1}=y_1|A_{i1}=0,c)}$$

and the infectiousness effect is given by

$$\frac{\mathbb{E}[Y_{i2}(1,Y_{i1}(1))|c]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))|c]} = \frac{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=1,y_1,c]P(Y_{i1}=y_1|A_{i1}=1,c)}{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]P(Y_{i1}=y_1|A_{i1}=1,c)}$$

Note that under the assumption that person 2 cannot be infected from outside the household we have $\mathbb{E}[Y_{i2}|a_{i1},Y_{i1}=0,c]=0$, and thus the empirical expressions above simplify considerably. When fitting statistical models for $\mathbb{E}[Y_{i2}|A_{i1}=1,y_1,c]$ and $P(Y_{i1}=y_1|A_{i1}=1,c)$, the restriction $\mathbb{E}[Y_{i2}|a_{i1},Y_{i1}=0,c]=0$ should be taken into account (Ogburn and VanderWeele, 2014b). This point was neglected in VanderWeele et al. (2012d).

Suppose the two models

$$\begin{aligned}\text{logit}\{P(Y_1=1|a_1,c)\} &= \beta_0+\beta_1 a_1+\beta_2' c\\ \text{logit}\{P(Y_2=1|a_1,Y_{i1}=1,c)\} &= \theta_0+\theta_1 a_1+\theta_4' c\end{aligned}$$

were fit to the data and that the infection outcome Y_2 for person 2 is sufficiently rare so that odds ratios approximate risk ratios (and the logit link approximated a log-link). Note that here we are modeling $\text{logit}\{P(Y_2=1|a_1,Y_{i1}=1,c)\}$; we do not need to model $\mathbb{E}[Y_{i2}|a_{i1},Y_{i1}=0,c]$ since $\mathbb{E}[Y_{i2}|a_{i1},Y_{i1}=0,c]=0$ and thus we can ignore terms such as $\theta_2 y_1$ and $\theta_3 a_1 y_1$, as we had in Chapter 2, in this model. Using these models for the conditional predicted probabilities for Y_1 and Y_2 gives, for the contagion effect,

$$\begin{aligned}\frac{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))|c]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(0))|c]} &= \frac{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]P(Y_{i1}=y_1|A_{i1}=1,c)}{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]P(Y_{i1}=y_1|A_{i1}=0,c)}\\ &\approx \frac{e^{\theta_0+\theta_4' c}\frac{e^{\beta_0+\beta_1+\beta_2' c}}{1+e^{\beta_0+\beta_1+\beta_2' c}}+0}{e^{\theta_0+\theta_4' c}\frac{e^{\beta_0+\beta_2' c}}{1+e^{\beta_0+\beta_2' c}}+0}\\ &= \frac{(1+e^{\beta_0+\beta_2' c})(e^{\beta_0+\beta_1+\beta_2' c}+1)}{(1+e^{\beta_0+\beta_1+\beta_2' c})(e^{\beta_0+\beta_2' c}+1)}\end{aligned}$$

and for the infectiousness effect

$$\begin{aligned}\frac{\mathbb{E}[Y_{i2}(1,Y_{i1}(1))|c]}{\mathbb{E}[Y_{i2}(0,Y_{i1}(1))|c]} &= \frac{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=1,y_1,c]P(Y_{i1}=y_1|A_{i1}=1,c)}{\sum_{y_1}\mathbb{E}[Y_{i2}|A_{i1}=0,y_1,c]P(Y_{i1}=y_1|A_{i1}=1,c)}\\ &\approx \frac{e^{\theta_0+\theta_1+\theta_4' c}\frac{e^{\beta_0+\beta_1+\beta_2' c}}{1+e^{\beta_0+\beta_1+\beta_2' c}}+0}{e^{\theta_0+\theta_4' c}\frac{e^{\beta_0+\beta_1+\beta_2' c}}{1+e^{\beta_0+\beta_1+\beta_2' c}}+0}\end{aligned}$$

$$= e^{\theta_1}.$$

If the infection outcome for person 2 is not rare, then the results above will hold if the logistic regression model for Y_2 is replaced by a log-linear model but the model for Y_1 is kept as a logistic model. No rare outcome assumption or log-linear model is needed for Y_1. Standard errors and confidence intervals for these expressions can be obtained via the delta method as in Chapter 2. The formal analytic relation between natural direct and indirect effects and the contagion and unconditional infectiousness effects also allows us to adapt sensitivity analysis techniques for natural direct and indirect effects in Chapter 3 to apply to contagion and infectiousness effects as in the text.

A few further technical comments merit attention. In Section 15.3 we gave an alternative definition of the infectiousness effect on a ratio scale as $\mathbb{E}[Y_{i2}(1, Y_{i1}(1))|Y_{i1}(1) = Y_{i1}(0) = 1]/\mathbb{E}[Y_{i2}(0, Y_{i1}(0))|Y_{i1}(1) = Y_{i1}(0) = 1]$—that is, the effect of the vaccine of person 1 on the infection status of person 2 in the principal stratum in which person 1 would be infected irrespective of vaccine status. This infectiousness effect is "conditional" in the sense that it is conditional on the subgroup for which person 1 would be infected irrespective of vaccine status, whereas the infectiousness effect in the text is an unconditional infectiousness effect—it averages over also those households for whom person 1 is uninfected. Yet another definition of an infectiousness effect could be given as $\mathbb{E}[Y_{i2}(1,1)|c]/\mathbb{E}[Y_{i2}(0,1)|c]$—that is, the effect of the vaccine of person 1 on the outcome of person 2, intervening to set person 1's infection status to positive. This effect is analogous to the "controlled direct effect" in the mediation literature considered in Part I of this book. Under assumptions (A15.3) and (A15.4) above, it is identified by $\mathbb{E}[Y_{i2}|A_{i1} = 1, Y_{i1} = 1, C_i = c]/\mathbb{E}[Y_{i2}|A_{i1} = 0, Y_{i1} = 1, C_i = c]$. It is a marginal effect insofar as it is for the entire population for which $C_i = c$; however it is "conditional on infection" in the sense that it considers a hypothetical contrast in which, in all households, person 1 is infected. Under the logistic regression models given above (assuming rare outcome or using a log-linear model rather than logistic model for Y_{i2}), this would be e^{θ_1}.

A.15.5. Tests for Specific Forms of Interference Using Causal Interactions

Here we show that the relationships between causal interaction and specific forms of interference extend to settings with multiple persons per cluster (cf. VanderWeele et al., 2012f). For causal interactions with three binary exposures of interest, let $D_{a_1a_2a_3}$ denote the counterfactual outcome for a person for some binary variable D if A_1, A_2, and A_3 had been set, possibly contrary to fact, to a_1, a_2, and a_3 respectively. We say that there is a three-way sufficient-cause interaction between A_1, A_2, and A_3 if there is a person such that $D_{111} = 1$ and $D_{110} = D_{101} = D_{011} = 0$. VanderWeele and Robins (2008) noted that if we let $p_{a_1a_2a_3} = \mathbb{E}(D|A_1 = a_1, A_2 = a_2, A_3 = a_3)$, then if the effects of A_1, A_2 and A_3 on D are unconfounded and if

$$p_{111} - p_{110} - p_{101} - p_{011} > 0$$

then a sufficient cause interaction must be present between A_1, A_2, and A_3. If the effects of A_1, A_2, and A_3 on D are positive monotonic and unconfounded, then any of the following three conditions imply the presence of a three-way sufficient cause interaction (VanderWeele and Robins, 2008; VanderWeele and Richardson, 2012):

$$p_{111} - p_{110} - p_{101} - p_{011} + p_{100} + p_{010} > 0$$
$$p_{111} - p_{110} - p_{101} - p_{011} + p_{100} + p_{001} > 0$$
$$p_{111} - p_{110} - p_{101} - p_{011} + p_{010} + p_{001} > 0$$

If just two of the exposures, say A_1 and A_2, have positive monotonic effects on the outcome, then the following condition suffices:

$$p_{111} - p_{110} - p_{101} - p_{011} + p_{001} > 0$$

Consider now instead the context of interference in which there are three people per cluster and A_{i1}, A_{i2}, and A_{i3} denote the exposures received by persons 1, 2, and 3 respectively in cluster i, and let $Y_{ij}(a_{i1}, a_{i2}, a_{i3})$ denote the counterfactual outcome for person j in cluster i if A_{i1}, A_{i2}, and A_{i3} had been set to a_{i1}, a_{i2}, and a_{i3} respectively. We could then define $D_i(a_{i1}, a_{i2}, a_{i3})$ as some function of $Y_{i1}(a_{i1}, a_{i2}, a_{i3})$, $Y_{i2}(a_{i1}, a_{i2}, a_{i3})$, and $Y_{i3}(a_{i1}, a_{i2}, a_{i3})$. For example, $D_i(a_{i1}, a_{i2}, a_{i3})$ might simply be whether person 1 has the outcome when $A_{i1} = a_{i1}, A_{i2} = a_{i2}, A_{i3} = a_{i3}$ i.e. $D_i(a_{i1}, a_{i2}, a_{i3}) = Y_{i1}(a_{i1}, a_{i2}, a_{i3})$. Alternatively $D_i(a_{i1}, a_{i2}, a_{i3})$ could be taken as an indicator of all three persons having the outcome or at least two having the outcome, or the first and second persons but not the third, and so on. In any case, regardless how $D_i(a_{i1}, a_{i2}, a_{i3})$ is defined, if we let $p_{a_1 a_2 a_3} = P(D = 1 | A_1 = a_1, A_2 = a_2, A_3 = a_3)$, we could test whether there are clusters such that $D_i(a_{i1}, a_{i2}, a_{i3}) = 1$ whenever all three people have the exposure present, but not when just two of the three have the exposure, by testing the condition the first of the inequalities above, that is, $p_{111} - p_{110} - p_{101} - p_{011} > 0$. For example, if we let $p_{a_1 a_2 a_3} = P(Y_1 = 1 | A_1 = a_1, A_2 = a_2, A_3 = a_3)$ and found that $p_{111} - p_{110} - p_{101} - p_{011} > 0$ held, then we could conclude that there were clusters in which the first person would have the outcome if all three people had the exposure but not if just two of three did.

As before, once $D_i(a_{i1}, a_{i2}, a_{i3})$ is defined, if we thought that the exposure for all three people had a positive monotonic effects on the outcome D_i, as defined, then we could test any of the three conditions in the second set of inequalities above instead of the first condition $p_{111} - p_{110} - p_{101} - p_{011} > 0$; if it were thought that the exposures of persons 1 and 2 had positive monotonic effects on the outcome, then one could test the condition $p_{111} - p_{110} - p_{101} - p_{011} + p_{001} > 0$ instead of condition $p_{111} - p_{110} - p_{101} - p_{011} > 0$ to test for the particular form of interference or pattern of outcome responses under question. As with two people per cluster, when there are three persons per cluster but they are indistinguishable from one another if A_i denotes the number who received the exposure in cluster or household i (i.e., $A_i = 0$, 1, 2, or 3) and we let $p_a = P(D = 1 | A = a)$, then if the exposure is randomized for each person with the same probability, the conditions above can be

rewritten respectively as

$$p_3 - 3p_2 > 0$$

or

$$p_3 - 3p_2 + 2p_1 > 0$$

or

$$p_3 - 3p_2 + p_1 > 0$$

Analogues of the epistatic/singular interaction for three binary exposures are given in VanderWeele and Richardson (2012) and could likewise be applied to test for corresponding forms of interference. VanderWeele and Richardson (2012) also consider both sufficient cause and epistatic/singular *n*-way interactions, and these could likewise be used to test for various forms of interference when there are *n* persons per cluster. The entire theory of causal interaction for *n*-way exposures maps onto tests for specific forms of interference among clusters with *n* persons.

Likewise the theory for causal interactions for two multivalued exposures maps onto tests for specific forms of interference for multivalued exposures between two people in a cluster. Theory has been developed for (a) sufficient-cause interactions for two exposures each with three levels and (b) epistatic/singular interactions for two exposures each with three levels (VanderWeele, 2010c,e). For example, for two exposures, A_1 and A_2, each with three levels, with $D_{a_1a_2}$ denoting the counterfactual outcome setting A_1 to a_1 and A_2 to a_2 and $p_{a_1a_2} = P(D = 1|A_1 = a_1, A_2 = a_2)$, then if the effects of A_1 and A_2 on D are unconfounded, there will be people such that $D_{a_1a_2} = 1$ if and only if $a_1 = a_2 = 2$ if

$$p_{22} - p_{21} - p_{20} - p_{12} - p_{11} - p_{10} - p_{02} - p_{01} - p_{00} > 0$$

Considerably weaker conditions can be tested under monotonicity assumptions or if sufficient-cause interactions, rather than epistatic/singular interactions, are in view. However, as before, these various empirical tests all have analogues for detecting various forms of interference between clusters with two people when the exposures have three levels (e.g., no vaccine, low-dose vaccine, high-dose vaccine). Here again, $Y_{ij}(a_{i1}, a_{i2})$ would denote the counterfactual outcome for person j in cluster i if A_{i1} and A_{i2} had been set to a_{i1} and a_{i2} respectively, $D_{ij}(a_{i1}, a_{i2},)$ could be defined as some function of $Y_{i1}(a_{i1}, a_{i2})$ and $Y_{i2}(a_{i1}, a_{i2})$, and the tests for either sufficient-cause or epistatic/singular interactions for categorical or ordinal variables could once again essentially be applied directly either with or without monotonicity assumptions. Conditions in which one exposure has two levels and the other has three are also available (VanderWeele, 2010c,e). Once we consider exposures with three or more levels, the number of possible forms of causal interaction (and forms of interference) and conditions that may be tested under varying monotonicity assumptions increases considerably; the interested reader can find these tests for causal interactions elsewhere (VanderWeele, 2010c,e). Analogous extensions are conceivable for multivalued exposures in clusters with an arbitrary number of people. We see that the close correspondence between causal interaction and forms of interference thus extends considerably further: The entire theory of causal interaction for *n*-way interactions between exposures (VanderWeele and Richardson,

2012) and also for multivalued exposures (VanderWeele, 2010ce) maps onto tests for specific forms interference.

A.15.6. Inferential Challenges with Many Individuals per Cluster

Here we will present the definitions for total, direct, indirect, and overall causal effects in the presence of interference under two-stage randomization. Suppose there are $N \geq 1$ groups of individuals, or blocks of units. For $i = 1, \ldots, N$, let n_i denote the number of individuals in group i and let $A_i = (A_{i1}, \ldots, A_{in_i})$ denote the treatments those n_i individuals receive. Assume A_{ij} is a dichotomous random variable having values 0 or 1 such that A_i can take on 2^{n_i} possible values. Let $A_{i(j)}$ denote the $n_i - 1$ subvector of A_i with the jth entry deleted. The vector A_i will be referred to as an intervention or treatment *program*, to distinguish it from the individual treatment A_{ij}. Let a_i and a_{ij} denote possible values of A_i and A_{ij}. Define R^j to be the set of vectors of possible treatment programs of length j, for $j = 1, 2, \ldots, n_i$.

Denote the potential outcome of individual j in group i under treatment a_i as $Y_{ij}(a_i)$. Denote $\mathbf{Y}_i(a_i)$ as the vector of such outcomes under treatment a_i for group i. The notation $Y_{ij}(a_i)$ allows for the possibility that the potential outcome for the individual j may depend on another individual's treatment assignment in group i; that is, there may be interference between individuals within a group. The $Y_{ij}(a_i)$ potential responses can be assumed fixed, as in this section, since they do not depend on the realized random assignment of treatments A_i, whereas the observed responses $Y_{ij}(A_i)$ do depend on A_i and thus are random variables. In subsequent sections we also consider potential outcomes $\mathbf{Y}_i(a_i)$ that are independent and identically distributed across blocks. We assume that if there is more than one group of individuals, then interference can occur within each group but not across groups, called a partial interference assumption by Sobel (2006). This will be a reasonable assumption, provided that the groups are sufficiently separate (e.g., in space or time). The form of the interference within groups is assumed unknown and can be of arbitrary form.

Hudgens and Halloran (2008) proposed a two-stage randomization scheme: The first stage is at the group level, whereas the second stage is at the individual level within groups. Let ψ and ϕ denote parameterizations that govern the distribution of A_i for $i = 1, \ldots, N$. Corresponding to the first stage of randomization, let $\mathbf{S} \equiv (S_1, \ldots, S_N)$ denote the group assignments with $S_i = 1$ if the group is assigned to ψ and 0 if assigned to ψ. Let ν denote the parameterization that governs the distribution of $\mathbf{S}$ and let $C \equiv \sum_i S_i$ denote the number of groups assigned ψ. Following Sobel (2006), Hudgens and Halloran (2008) focused on a mixed group and mixed individual assignment strategy whereby a fixed number of groups were allocated to ψ, and within each group, a fixed number of individuals were allocated to treatment versus control. VanderWeele and Tchetgen Tchetgen (2011b) and Tchetgen Tchetgen and VanderWeele (2012) consider in addition what we call a simple randomization scheme whereby treatment is randomly assigned to different individuals within group i according to a Bernoulli probability mass function. For example, under the mixed allocation scheme, ψ could be that exactly half the

individuals of each group receive the treatment, and the other half the control. Under the simple allocation scheme, ψ could be that each individual in each group assigned to ψ will be assigned treatment with probability 0.5. In the mixed scheme, there is some dependence among the treatment assignments of the individuals. The causal estimands defined below have the same form under either randomization scheme, though the different randomization schemes result in subtle differences of interpretation.

Causal estimands are typically defined in terms of averages of potential outcomes that are identifiable from observable random variables. Following this approach, the potential outcomes for individual j in group i under $a_{ij} = a$ can be written

$$Y_{ij}(\mathbf{a}_{i(j)}, a_{ij} = a)$$

for $a = 0, 1$. Because the potential outcome depends on $a_{i(j)}$, following Sobel (2006), Hudgens and Halloran (2008) defined the *individual average potential outcome* for individual j in group i under $a_{ij} = a$ by

$$\overline{Y}_{ij}(a;\psi) \equiv \sum_{\omega \in R^{n_i - 1}} Y_{ij}(\mathbf{a}_{i(j)} = \omega, a_{ij} = a)\Pr_{\psi}(\mathbf{A}_{i(j)} = \omega | A_{ij} = a)$$

In other words, under the mixed allocation strategy of Hudgens and Halloran (2008) and Sobel (2006), the individual average potential outcome is the conditional expectation of $Y_{ij}(A_i)$ given $A_{ij} = 1$ under assignment strategy ψ. In contrast, under the simple allocation strategy of VanderWeele and Tchetgen Tchetgen (2011b), the potential outcomes are averaged over the unconditional distribution of A_i. Averaging over individuals, define the *group average potential outcome* under treatment assignment a as $\overline{Y}_i(a;\psi) \equiv \sum_{j=1}^{n_i} \overline{Y}_{ij}(a;\psi)/n_i$. Finally, averaging over groups, define the *population average potential outcome* under treatment assignment a as $\overline{Y}(a;\psi) \equiv \sum_{i=1}^{N} \overline{Y}_i(a;\psi)/N$. The causal estimands are defined in terms of the individual, group, and population average potential outcomes.

Halloran and Struchiner (1991) defined the direct effect of a treatment on an individual as the difference between (a) the potential outcome for that individual given treatment and (b) the potential outcome for that individual without treatment, all other things being equal. Formally, following Halloran and Struchiner (1995), define the *individual direct causal effects* of treatment 0 compared to treatment 1 for the individual j in group i by

$$CE^{D}_{ij}(\mathbf{a}_{i(j)}) \equiv Y_{ij}(\mathbf{a}_{i(j)}, A_{ij} = 1) - Y_{ij}(\mathbf{a}_{i(j)}, A_{ij} = 0)$$

Hudgens and Halloran (2008) defined the *individual average direct causal effect* for the jth individual in the ith group by

$$\overline{CE}^{D}_{ij}(\psi) \equiv \overline{Y}_{ij}(1;\psi) - \overline{Y}_{ij}(0;\psi)$$

that is, the difference in individual average potential outcomes when $a_{ij} = 1$ and when $a_{ij} = 0$ under ψ. Following Hudgens and Halloran (2008), define the *group average direct causal effect* by $\overline{CE}^{D}_{i}(\psi) \equiv \overline{Y}_i(1;\psi) - \overline{Y}_i(0;\psi) = \sum_{j=1}^{n_i} \overline{CE}^{D}_{ij}(\psi)/n_i$

and define the *population average direct causal effect* by $\overline{CE}^D(\psi) \equiv \overline{Y}(1;\psi) - \overline{Y}(0;\psi) = \sum_{i=1}^{N} \overline{CE}_i^D(\psi)/N$.

In contrast to direct effects, an indirect effect describes the effect on an individual of the treatment received by others in the group. In particular, Halloran and Struchiner (1991) defined the indirect effect of a treatment on an individual as the difference between the potential outcomes for that individual without treatment when the group (i) receives an intervention program and (ii) receives the benchmark program of no intervention. Sobel (2006) refers to the indirect effect as the spillover effect. Similar to Halloran and Struchiner (1995), define the *individual indirect causal effects* of treatment program a compared with a' on individual j in group i by

$$CE_{ij}^I(\mathbf{a}_{i(j)}, \mathbf{a}'_{i(j)}) \equiv Y_i(\mathbf{a}_{i(j)}, a_{ij} = 0) - Y_i(\mathbf{a}'_{i(j)}, a'_{ij} = 0)$$

where a' is another n_i-dimensional vector of treatment random variables. (Note that a' does not denote the transpose of a). Similar to direct effects, Hudgens and Halloran (2008) defined the *individual average indirect causal effect* by $\overline{CE}_{ij}^I(\phi,\psi) \equiv \overline{Y}_{ij}(0;\phi) - \overline{Y}_{ij}(0;\psi)$. Clearly if $\psi = \phi$, then $\overline{CE}_{ij}^I(\phi,\psi) = 0$; that is, there will be no individual average indirect causal effects. Finally, Hudgens and Halloran (2008) defined the *group average indirect causal effect* as $\overline{CE}_i^I(\phi,\psi) \equiv \overline{Y}_i(0;\phi) - \overline{Y}_i(0;\psi) = \sum_{j=1}^{n_i} \overline{CE}_{ij}^I(\phi,\psi)/n_i$ and defined the *population average indirect causal effect* as $\overline{CE}^I(\phi,\psi) \equiv \overline{Y}(0;\phi) - \overline{Y}(0;\psi) = \sum_{i=1}^{N} \overline{CE}_i^I(\phi,\psi)/N$.

Total effects describe both the direct and indirect effects of a particular treatment assignment on an individual. Halloran and Struchiner (1991) define the total effect of a treatment on an individual as the difference between the potential outcomes for that individual (i) with treatment when the group receives an intervention program and (ii) without treatment when the group receives no intervention. Following Halloran and Struchiner (1995), define the *individual total causal effects* for individual i in group j as

$$CE_{ij}^T(\mathbf{a}_{i(j)}, \mathbf{a}'_{i(j)}) \equiv Y_{ij}(\mathbf{a}_{i(j)}, a_{ij} = 1) - Y_{ij}(\mathbf{a}'_{i(j)}, a'_{ij} = 0)$$

Hudgens and Halloran (2008) defined the *individual average total causal effect* by $\overline{CE}_{ij}^T(\phi,\psi) \equiv \overline{Y}_{ij}(1;\phi) - \overline{Y}_{ij}(0;\psi)$, the *group average total causal effect* by $\overline{CE}_i^T(\phi,\psi) \equiv \overline{Y}_i(1;\phi) - \overline{Y}_i(0;\psi) = \sum_{j=1}^{n_i} \overline{CE}_{ij}^T(\phi,\psi)/n_i$, and the *population average total causal effect* by $\overline{CE}^T(\phi,\psi) \equiv \overline{Y}(1;\phi) - \overline{Y}(0;\psi) = \sum_{i=1}^{N} \overline{CE}_i^T(\phi,\psi)/N$. It follows by simple addition and subtraction that a total effect is the sum of the direct and indirect effects at the individual, individual average, group average, and population average levels. For example, $\overline{CE}^T(\phi,\psi) = \overline{Y}(1;\phi) - \overline{Y}(0;\psi) = \overline{Y}(1;\phi) - \overline{Y}(0;\phi) + \overline{Y}(0;\phi) - \overline{Y}(0;\psi) = \overline{CE}_i^D(\phi) + \overline{CE}^I(\phi,\psi)$.

Halloran and Struchiner (1991) defined the overall causal effect to be the average effect of an intervention program relative to no intervention. Hudgens and Halloran (2008) defined the *individual overall causal effect* of treatment a_i compared to treatment a'_i for individual j in group i by $CE_{ij}^O(a_i, a'_i) \equiv Y_{ij}(a_i) - Y_{ij}(a'_i)$. Similarly,

for the comparison of ϕ to ψ, they defined the *individual average overall causal effect* by $\overline{CE}_{ij}^{O}(\phi,\psi) \equiv \overline{Y}_{ij}(\phi) - \overline{Y}_{ij}(\psi)$, the *group overall causal effect* by $\overline{CE}_{i}^{O}(\phi,\psi) \equiv \overline{Y}_{i}(\phi) - \overline{Y}_{i}(\psi)$, and the *population overall causal effect*by $\overline{CE}^{O}(\phi,\psi) \equiv \overline{Y}(\phi) - \overline{Y}(\psi)$. VanderWeele and Tchetgen Tchetgen (2011b) showed that the overall effect decomposes into the sum of an indirect effect and a contrast of two direct effects on the individual average, group average, and population average levels. For example, $\overline{CE}^{O}(\phi,\psi) = \overline{CE}^{I}(\phi,\psi) + \{\overline{CE}_{i}^{D}(\phi)\Pr_{\phi}(A_{ij}=1) - \overline{CE}_{i}^{D}(\psi)\Pr_{\psi}(A_{ij}=1)\}$.

The estimands defined above simplify under the assumption of no interference between individuals within a group since the potential outcomes of the jth individual in group i can be written as $Y_{ij}(1)$ and $Y_{ij}(0)$. In turn, the individual direct causal effect is no longer dependent on the treatment assignment vector $a_{i(j)}$ and simply equals $Y_{ij}(1) - Y_{ij}(0)$. The corresponding group average direct causal effect becomes $\sum_{j=1}^{n_i}\{Y_{ij}(1) - Y_{ij}(0)\}/n_i$, that is, the usual average causal effect estimand. The individual indirect causal effect equals zero for all individuals assuming no interference. Similarly, the individual total causal effect equals the individual direct causal effect. Likewise, at the group average level, under the no interference assumption the indirect causal effect is zero and the direct causal effect equals the total causal effect.

Assuming the two-stage randomization and mixed allocation strategy, Hudgens and Halloran (2008) give unbiased estimators for the population average direct, indirect, total, and overall effects. They provide variance estimates under the assumption of stratified interference, that is, if it matters only how many people are allocated to treatment, not exactly which ones. Tchetgen Tchetgen and VanderWeele (2012) provide conservative variance estimators under more general assumptions and give finite sample confidence intervals for direct, indirect, total, and overall effects without the assumption of stratified interference. VanderWeele and Tchetgen Tchetgen (2011b) provide unbiased estimators under the simple allocation strategy. Liu and Hudgens (2013) develop large sample randomization inference for the different causal effects in the presence of interference when the effects are assumed homogeneous across blocks.

A.15.7. Spillover Effects and Observational Data

We assume the counterfactual outcome of each person depends on the exposure received by the other persons in the same cluster only through some known function g (e.g., the mean of the other exposures of the other individuals, or some other summary measure) so that the potential outcome for person j in cluster i could be written as $Y_{ij}\left(a,g\right)$ where a is the individual i's own exposure and g is the summary measure of all the other individuals. Let C_{ij} denote the covariates for individual j in cluster i. Let H_{ij} be some known function (possibly a vector) of all covariates C_{ij} for all individuals in cluster i other than individual j. Consider the assumption

$$\mathbb{E}[Y\left(a,g\right)|A=a,G=g,C=c,H=h] = \mathbb{E}[Y\left(a,g\right)|C=c,H=h]$$

If this held, we would have

$$\mathbb{E}[Y\left(a,g\right)|C=c,H=h]=\mathbb{E}[Y|A=a,G=g,C=c,H=h]$$

From this, one could obtain conditional individual/direct, spillover/indirect and total effects, namely,

$$\begin{aligned}&\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y_{ij}(a^*,g)|c,h]\\&\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y_{ij}(a,g^*)|c,h]\\&\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y_{ij}(a^*,g^*)|c,h]\end{aligned}$$

Marginal effects, involving counterfactuals of the form $\mathbb{E}[Y\left(a,g\right)]$, could be obtained by averaging over the distributions of C and H.

Suppose now that we have unmeasured confounding by one or more unmeasured confounders U_{ij} and let V_{ij} denote some function (possibly the entire vector) of U_{ij} for all individuals in cluster i other than individual j. Suppose that conditional on observed C,H and unobserved U,V we have

$$\begin{aligned}&\mathbb{E}[Y\left(a,g\right)|A=a,G=g,C=c,H=h,U=u,V=v]\\&\quad=\mathbb{E}[Y\left(a,g\right)|C=c,H=h,U=u,V=v]\end{aligned}$$

Without data on U_{ij}, causal effects are not identified. However, we can still employ sensitivity analysis. Let B denote the difference between the causal effect and the biased estimand; that is, $B=\{\mathbb{E}[Y|a,g,c,h]-\mathbb{E}[Y|a^*,g^*,c,h]\}-\{\mathbb{E}[Y(a,g)|c,h]-\mathbb{E}[Y(a^*,g^*)|c,h]\}$ denotes this difference. We then have the following:

Proposition 15.3 (VanderWeele et al., 2014b):
Suppose

$$\begin{aligned}&\mathbb{E}[Y\left(a,g\right)|A=a,G=g,C=c,H=h,U=u,V=v]\\&\quad=\mathbb{E}[Y\left(a,g\right)|C=c,H=h,U=u,V=v]\end{aligned}$$

then if u' and v' denote arbitrary reference values for U and V, respectively, we have

$$\begin{aligned}B=&\sum\nolimits_{u,v}\{\mathbb{E}(Y|a,g,c,h,u,v)-\mathbb{E}(Y|a,g,c,h,u',v')\}\{P(u,v|a,g,c,h)-P(u,v|c,h)\}\\&-\sum\nolimits_{u,v}\{\mathbb{E}(Y|a^*,g^*,c,h,u,v)-\mathbb{E}(Y|a^*,g^*,c,h,u',v')\}\{P(u,v|a^*,g^*,c,h)\\&-P(u,v|c,h)\}\end{aligned}$$

If there is a single unmeasured confounder U and V is scalar and if the effects of U and V are additive in the sense that $\mathbb{E}(Y|a,g,c,h,u,v)-\mathbb{E}(Y|a,g,c,h,u',v')=\lambda(u-u')+\tau(v-v')$ for $(A,G)=(a,g)$ and $(A,G)=(a^*,g^*)$, then

$$B=\lambda\{\mathbb{E}[U|a,g,c,h]-\mathbb{E}[U|a^*,g^*,c,h]\}+\tau\{\mathbb{E}[V|a,g,c,h]-\mathbb{E}[V|a^*,g^*,c,h]\}$$

Proof:
If in Proposition 3.1, we take the exposure, measured covariates, and unmeasured covariates as (A,G), (L,H) and (U,V), then we have that

$$\begin{aligned}
B &= \mathbb{E}[Y|a,g,l,h] - \mathbb{E}[Y|a^*,g^*,l,h] - \mathbb{E}[Y(a,g)|l,h] - \mathbb{E}[Y(a^*,g^*)|l,h] \\
&= \mathbb{E}[Y|a,g,l,h] - \mathbb{E}[Y|a^*,g^*,l,h] - \sum\nolimits_{u,v}\{\mathbb{E}[Y|a,g,l,h,u,v] \\
&\quad - \mathbb{E}[Y|a^*,g^*,l,h,u,v]\}P(u,v) \\
&= \sum\nolimits_{u,v}\{\mathbb{E}(Y|a,g,c,h,u,v) - \mathbb{E}(Y|a,g,c,h,u',v')\}\{P(u,v|a,g,c,h) - P(u,v|c,h)\} \\
&\quad - \sum\nolimits_{u,v}\{\mathbb{E}(Y|a^*,g^*,c,h,u,v) - \mathbb{E}(Y|a^*,g^*,c,h,u',v')\}\{P(u,v|a^*,g^*,c,h) \\
&\quad - P(u,v|c,h)\}
\end{aligned}$$

If $\mathbb{E}(Y|a,g,l,h,u,v) - \mathbb{E}(Y|a,g,l,h,u^*,v^*) = \lambda(u-u^*) + \tau(v-v^*)$, then we have

$$\begin{aligned}
B &= \sum\nolimits_{u,v}\{\mathbb{E}(Y|a,g,l,h,u,v) - \mathbb{E}(Y|a,g,l,h,u',v')\}\{P(u,v|a,g,l,h) - P(u,v|l,h)\} \\
&\quad - \sum\nolimits_{u,v}\{\mathbb{E}(Y|a^*,g^*,l,h,u,v) - \mathbb{E}(Y|a^*,g^*,l,h,u',v')\}\{P(u,v|a^*,g^*,l,h) \\
&\quad - P(u,v|l,h)\} \\
&= \sum\nolimits_{u,v}\{\lambda(u-u') + \tau(v-v')\}\{P(u,v|a,g,l,h) - P(u,v|l,h)\} \\
&\quad - \sum\nolimits_{u,v}\{\lambda(u-u') + \tau(v-v')\}\{P(u,v|a^*,g^*,l,h) \\
&\quad - P(u,v|l,h)\} \\
&= \sum\nolimits_{u,v}(\lambda u + \tau v)P(u,v|a,g,l,h) - \sum\nolimits_{u,v}(\lambda u + \tau v)P(u,v|a^*,g^*,l,h) \\
&= \lambda\{\mathbb{E}[U|a,g,l,h] - \mathbb{E}[U|a^*,g^*,l,h]\} + \tau\{\mathbb{E}[V|a,g,l,h] - \mathbb{E}[V|a^*,g^*,l,h]\}
\end{aligned}$$

This completes the proof. ■

To obtain the bias factor B, one could thus specify the effect of the unmeasured confounders U and V on the outcome, $\mathbb{E}(Y|a,g,c,h,u,v) - \mathbb{E}(Y|a,g,c,h,u',v')$, $(A,G) = (a,g)$ and $(A,G) = (a^*,g^*)$, and also how the distribution of U and V differs when $(A,G) = (a,g)$ versus (a^*,g^*), that is, $P(u,v|a,g,c,h)$ and $P(u,v|a^*,g^*,c,h)$. One can use these sensitivity analysis parameters to calculate the bias factor in and then subtract the bias factor B from the estimate of the causal effect using the observed data $\mathbb{E}[Y|a,g,c,h] - \mathbb{E}[Y|a^*,g^*,c,h]$ to obtain a corrected effect estimate for $\mathbb{E}[Y(a,g)|c,h] - \mathbb{E}[Y(a^*,g^*)|c,h]$. Note that the expression for the bias factor in the Proposition makes no assumption beyond the assumption that control for observed (C,H) and unobserved (U,V) would suffice to control for confounding of the effect of (A,G) on Y; it allows for multiple unmeasured confounders. However, the use of the bias formula requires specifying a large number of parameters: $\mathbb{E}(Y|a,g,c,h,u,v) - \mathbb{E}(Y|a,g,c,h,u',v')$ for every value of u,v and the distribution $P(u,v|a,g,c,h)$ for both $(A,G) = (a,g)$ and $(A,G) = (a^*,g^*)$. The simplified approach under scalar U and V and additivity requires specifying far fewer parameters. Hong and Raudenbush (2006) use a similar approach

to express the simplified bias formula $B = \lambda\{\mathbb{E}[U|a,g,c,h] - \mathbb{E}[U|a^*,g^*,c,h]\} + \tau\{\mathbb{E}[V|a,g,c,h] - \mathbb{E}[V|a^*,g^*,c,h]\}$ but do not provide a derivation and do not articulate the assumptions needed for the use of the formula.

REFERENCES

Acute Respiratory Distress Syndrome Network. (2000). Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine*, **342**:1301–1308.

Ahsan, H., Chen, Y., Parvez, F., Zablotska, L., Argos, M., Hussain, I., Momotaj, H., Levy, D., Cheng, Z., Slavkovich, V., van Geen, A., Howe, G.R., and Graziano, J. H. (2006). Arsenic exposure from drinking water and risk of premalignant skin lesions in Bangladesh: Baseline results from the Health Effects of Arsenic Longitudinal Study. *American Journal of Epidemiology*, **163**:1138–1148.

Ai, C., and Norton E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, **80**:123–129.

Aiken, L. S., and West. S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. SAGE Publications.

Albert, J. M. (2012). Distribution-free mediation analysis for nonlinear models with confounding. *Epidemiology*, **23**:879–888.

Albert, J. M., and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, **67**:1028–1038.

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene–environment interactions. *American Journal of Epidemiology*, **154**:687–693.

Almirall, D., Ten Have, T., and Murphy, S. A. (2010) Structural nested mean models for assessing time-varying effect moderation. *Biometrics*, **66**:131–139.

Alwin, D. F., and Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, **40**:37–47.

Amos, C. I., Wu, X., Broderick, P., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics*, **40**(5):616–622.

An, W. (2011). *Peer Effects on Adolescent Smoking and Social Network-Based Interventions*. Ph.D. dissertation, Department of Sociology, Harvard University.

Ananth, C. V., and VanderWeele, T. J. (2011). Placental abruption and perinatal mortality with preterm delivery as a mediator: Disentangling direct and indirect effects. *American Journal of Epidemiology*, **174**:99–108.

Andersson, T., Alfredsson, L., Kallberg, H., Zdravkovic, S., and Ahlbom, A. (2005). Calculating measures of biological interaction. *European Journal of Epidemiology*, **20**:575–579.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, **91**:444–472.

Aronow, P. M., and Samii, C. (2013) Estimating average causal effects under general interference. arXiv:1305.6156v1 at arxiv.org/pdf.

Assmann, S. F., Hosmer, D. W., Lemeshow, S., and Mundt, K. A. (1996). Confidence intervals for measures of interaction. *Epidemiology*, **7**:286–290.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 357–363.

Balke, A., and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92**:1172–1176.

Baron, R. M., and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**:1173–1182.

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge, UK: Cambridge University Press.

Baum, L., and Frampton, P. H. (2007). Turnaround in cyclic cosmology. *Physical Review Letters*, **98**:071301.

Bennett, W. P., Alavanja, M. C. R., Blomeke, B., Vähäkangas, K. H., Castrén, K., Welsh, J. A., Bowman, E. D., Khan, M. A., Flieder, D. B., and Harris, C. C. (1999). Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. *Journal of the National Cancer Institute*, **91**:2009–2014.

Berzuini, C., and Dawid, A. P. (2013). Deep determinism and the assessment of mechanistic interaction. *Biostatistics*, **14**:502–513.

Bhavnani, D., Goldstick, J. E., Cevallos, W., Trueba, G., and Eisenberg, J. N. S. (2012). Synergistic effects between rotavirus and coinfecting pathogens on diarrheal disease: Evidence from a community-based study in northwestern Ecuador. *American Journal of Epidemiology*, **176**:387–395.

Blot, W. J., and Day, N. E. (1979). Synergism and interaction: Are they equivalent? *American Journal of Epidemiology*, **110**:99–100.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons, 1989.

Bonetti, M., and Gelber, R. D. (2000). A graphical method to assess treatment–covariate interactions using the cox model on subsets of the data. *Statistics in Medicine*, **19**:2595–2609.

Bonetti, M., and Gelber, R. D. (2005). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, **5**:465–481.

Botto, L. D., and Khoury, M. J. (2001). Facing the challenge of gene–environment interaction: The two-by-four table beyond. *American Journal of Epidemiology*, **153**:106–1020.

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics*, **10**:478–486.

Bross, I. D. (1966). Spurious effects from an extraneous variable. *Journal of Chronic Diseases*, **19**:637–647.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer.

Cacioppo, J. T., Fowler, J. H., and Christakis, N. A. (2009). Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology*, **97**:977–991.

Caffo, B., Chen, S., Stewart, W., Bolla, K., Yousem, D., Davatzikos, C., and Schwartz, B. S. (2008). Are brain volumes based on magnetic resonance imaging nediators of the associations of cumulative lead dose with cognitive function? *American Jouranl of Epidemiology*, **167**:429–437.

Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, **12**:270–282.

Cai, Z., Kuroki, M., Pearl, J., and Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometics*, **64**:695–701.

Canonicao, M., Olie, V., Carcaillon, L., Tubert-Bitter, P., and Scarabin, P.-Y. (2008). Synergism between non-O blood group and oral estrogen in the risk of vennous thromboemolism among postmenopausal women. The ESTHER Study. *Thrombosis and Haemostasis*, **99**:246–248.

Carrington, P. J., Scott, J., and Wasserman, S. (2005). *Model and Methods in Social Network Analysis*. New York: Cambridge University Press.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models*, 2nd ed. Boca Raton, FL: Chapman and Hall.

Cayley, A. (1853). Note on a question in the theory of probabilities. *London, Edinburgh and Dublin Philosophical Magazine*, **VI**:259.

Chang, V. W., and Lauderdale, D. S. (2005). Income disparities in body mass index and obesity in the United States, 1971–2002. *Archives of Internal Medicine*, **165**:2122–2128.

Chatterjee, N., and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene–environment independence in case–control studies. *Biometrika*, **92**:399–418.

Chatterjee, N., Kalaylioglu, Z., Moleshi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene–gene and gene–environment interactions. *American Journal of Human Genetics*, **79**:1002–1016.

Chen, H., Geng, Z., and Jia, J. (2007). Criteria for surrogate end points. *Journal of the Royal Statistical Society, Series B*, **69**:919–932.

Chen, Y., Graziano, J. H., Parvez, F., Hussain, I., Momotaj, H., van Geen, A., Howe, G. R., and Ahsan, H. (2006). Modification of risk of arsenic-induced skin lesions by sunlight exposure, smoking, and occupational exposures in Bangladesh. *Epidemiology*, **17**(4):459–467.

Cheng, J., and Small, D. S. (2006). Bounds on causal effects in three-arm trials with noncompliance. *Journal of the Royal Statistical Society, Series B*, **68**:815–836.

Cheng, K. F. and Lin, W. J. (2009). The effects of misclassification in studies of gene–environment interactions. *Human Heredity*, **67**:77–87.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**:367–405.

Chiba, Y. (2010). Bias analysis for the principal stratum direct effect in the presence of confounded intermediate variables. *Journal of Biometrics and Biostatistics*, **1**:101.

Chiba, Y., and Taguri, M. (2013). Conditional and unconditional infectiousness effects in vaccine trials. *Epidemiology*, **24**:336–337.

Chiba, Y., and VanderWeele, T. J. (2011). A simple sensitivity analysis technique for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology*, **173**:745–751.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, **357**:370–379.

Christakis, N. A., and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, **358**:2249–2258.

Christakis, N. A., and Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*, **32**:556–577.

Chu, H., Nie, L., and Cole, S. R. (2011). Estimating the relative excess risk due to interaction: A Bayesian approach. *Epidemiology*, **22**:242–248.

Clarke, S., and Leibniz, G. (1717). *The Leibniz–Clarke Correspondence*. H. G. Alexander, editor. Manchester: Manchester University Press, 1956.

Cochran, W.G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, **5**:171–176.

Cohen-Cole, E., and Fletcher. J. M. (2008). Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis. *British Medical Journal*, **337**:a2533.

Cole, D. A., and Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, **112**:558–577.

Cole, S. R., and Hernán, M. A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology*, **31**:163–165.

Cole, S. R., Platt, R. W., Schisterman, E. F., et al. (2010). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, **39**:417–420.

Collins, J., Hall, N. and Paul, L. A. (2004). Counterfactual and causation: History, problems and prospects. In: J. Collins, N. Hall, and L. A. Paul, editors. *Causation and Counterfactuals*. Cambridge, MA: MIT Press, pp. 1–58.

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *The Review of Economics and Statistics*, **94**:260–272.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society, Series B*, **59**:55–95.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**:2463–2468.

Cordell, H. J. (2009). Detecting gene–gene interaction that underlie human diseases. *Nature Reviews Genetics*, **10**:392–404.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, L. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22**:173–203.

Cox, D. R. (1958). *The Planning of Experiments*. New York: John Wiley & Sons.

Craig, W. L., and Sinclair, J. D. (2009). The kalām cosmological argument. In: W. L. Craig and J. P. Moreland, editors. *The Blackwell Companion to Natural Theology*. London: Blackwell, pp. 101–201.

Craig, W. L., and Smith, Q. (1993). *Theism, Atheism, and Big Bang Cosmology*. New York: Oxford University Press.

Cuzick, J., Sasieni, P., Myles, J., and Tyrer, J. (2007). Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *Journal of the Royal Statistical Society, Series B*, **69**:565–588.

Dai, J., Logsdon, B., Huang, Y., et al. (2012). Simultaneous testing for marginal genetic association and gene–environment interaction in genome-wide association studies. *American Journal of Epidemiology*, **176**:164–173.

Daniel, R.M., De Stavola, B.L., Cousens, S.N. (2011). gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, **11**(4):479–517.

Datta, S., Halloran, M. E., and Longini, I. M. (1999). Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual versus household. *Biometrics*, **55**:792–798.

Davey, K., and Clifton, R. (2001). Insufficient reason in the "New Cosmological Argument." *Religious Studies*, **37**:485–490.

Davey Smith, G., and Ebrahim, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, **32**:1–22.

Davey Smith, G. and Ebrahim, S. (2004). Mendelian randomization: Prospects, potentials, and limitations. *International Journal of Epidemiology*; **33**:30–42.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, **60**: 685–700.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon.

Dawber, T. R. (1980). *The Framingham Study: The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press.

de González, A. B, and Cox, D. R. (2007). Interpretation of interaction: A review. *Annals of Applied Statistics*, **1**:371–385.

Deeks, J. J. and Altman, D. G. (2003). Effect measures for met-analysis of trials with binary outcomes. In: M. Egger, G. Davey Smith, and D. G. Altman, editors. *Systematic Reviews in Health Care: Meta-analysis in Context*. London: BMJ Publishing Group, 313–335.

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine*, **27**:36–46.

Didelez, V., and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, **16**:309–330.

Didelez, V., Dawid, A. P., and Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, **25**:22–40.

Draper, N., and Smith H. (1981). *Applied Regression Analysis*, 2nd edn. New York: John Wiley & Sons.

Duncan, O. D. (1966). Path analysis: sociological examples. *American Journal of Sociology*, **72**:1–16.

Egleston, B., Sharfstein, D. O, and MacKenzie, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, **65**:497–504.

Elliott, M. R., Raghunathan, T. E., and Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes, *Biostatistics*, **11**:353–372.

Elliott, P., Chambers, J. C., Zhang, W., et al. (2009). Genetic loci assicated with C-reactive protein levels and risk of coronary heart disease. *JAMA*, **302**:37–48.

Emsley, R., Dunn, G., and White, I. R. (2010). Mediation and moderation of exposure effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research*, **19**:237–270.

Emsley, R. A., Liu, H., Dunn G., Valeri, L., and VanderWeele T. J. (2014). PARAMED: A command to perform causal mediation analysis using parametric models. Technical Report.

Emlsey, R., and Dunn, G. (2012). Evaluation of potential mediator in randomised trials. In: C. Berzuini, P. Dawid, and L. Bernardinelli, editors. *Causality: Statistical Perspectives and Applications*. New York: John Wiley & and Sons.

Emsley, R., and VanderWeele, T. J. (2014). Mediation and sensitivity analysis using two or more trials. Technical Report.

Engels, E. A., Schmid, C. H., Terrin, N., et al. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine*, **19**:1707–1728.

Evre, S., Bowes, J., Diogo, D., et al. (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics*, **44**:1336–1340.

Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M., and Castelli. W. P. (1975). The Framingham Offspring Study: Design and preliminary data. *Preventive Medicine*, **4**:518–525.

Figueiredo, J. C., Knight, J. A., Briollais, L., Andrulis, I. L., and Ozcelik, H. (2004). Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario Site of the Breast Cancer Family Registry. *Cancer Epidemiology, Biomarkers and Prevention*, **13**:583–591.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52**:399–433.

Fisher, R. A. (1958). Lung cancer and cigarettes. *Nature*, **182**(4628):108.

Flanders, D. (2006). Sufficient-component cause and potential outcome models. *European Journal of Epidemiology* **21**, 847–853.

Flanders, W. D., and Khoury, M. J. (1990). Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. *Epidemiology*, **1**:239–246.

Fleming, T. R. and DeMets, D. L. (1996). Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine*, **125**:606–613.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine tirals. *Biometrics*, **62**:1161–1169.

Foppa, I., and Spiegelman, D. (1997). Power and sample size calculations for case–control studies of gene-environment interactions with a polytomous exposure variable. *American Journal of Epidemiology*, **146**:596–604.

Fowler, J. H., and Christakis, N. A. (2008). Estimating peer effects on health in social networks. *Journal of Health Economics*, **27**:1386–1391.

Frangakis, C. E., and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58**:21–29.

Frangakis, C. E., Rubin, D. B., An, M. W., and MacKenzie, E. (2007). Principal stratification designs to estimate input data missing due to death (with discussion). *Biometrics*, **63**:641–662.

Freedman, L., Graubard, B., and Schatzkin, A. (1992). Statistcal validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**:167–178.

Freedman L. S., and Schatzkin A. (1992). Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology* **136**:1148–1159.

Fritz, M. S., and MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. Psychological Science **18**:233–289.

Gail, M. H., and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**(2):361–372.

Gail, M. H., Wacholder, S., and Lubin, J. H. (1988). Indirect corrections for confounding under multiplicative and additive risk models. *American Journal of Industrial Medicine*, **13**:119–130.

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics*, **1**:231–246.

Gale, R. M., and Pruss, A. R. (1999). A new cosmological argument. Reprinted in: R. M. Gale and A. R. Pruss, *The Existence of God*. Burlington, VT: Ashgate, 2003, pp. 365–380.

Gale, R. M., and Pruss, A. R. (2002). A response to Oppy, and to Davey and Clifton. *Religious Studies*, **38**:89–99.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., and Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, **28**:1108–1130.

Garcia-Closas, M., and Lubin, J. H. (1999). Power and sample size calculations in case–control studies of gene–environment interactions: Comments on different approaches. *American Journal of Epidemiology*, **149**:689–692.

Garcia-Closas, M., Thompson, W. D., and Robins, J. M. (1998). Differential misclassifcation and the assessment of gene–enviornment interactions in case–control studies. *American Journal of Epidemiology*, **147**:426–433.

Garcia-Closas, M., Couch, F. J., Lindstrom, S., et al. (2013). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature Genetics*, **45**:392–398.

Gauderman, W. J. (2002a). Sample size requirements for association studies of gene–gene interaction. *American Journal of Epidemiology*, **155**:478–484.

Gauderman, W. J. (2002b). Sample size requirements for matched case–control studies of gene–environment interaction. *Statistics in Medicine*, **21**:35–50.

Gayan, J., et al. (2008). A method for detecting epistasis in genome-wide studies using case–control multi-locus association analysis. *BMC Genomics*, **9**:360.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Soceity, Series B*, **69**:199–216.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, **64**:1146–1154.

Gilbert, P. B., Bosch, R., and Hudgens, M. G. (2003), Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, **59**:531–541.

Glennan, S. (2009). Causation and explanation. In: H. Beebee, C. Hitchcock, and P. Menzies, editors. *Oxford Handbook of Causation*. Oxford: Oxford University Press, pp. 315–325.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Teng, C. M., and Zhang, J. (2010). Actual causation: A stone soup essay. *Synthese* **175**:169–192.

Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012). Credible Mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, **175**:332–339.

Glynn, A. N. (2012). The product and difference fallacies for indirect effects. *American Journal of Political Science*. **56**:257–269.

Graham, B. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica*, **76**:643–660.

Graham, B. S., Imbens, G. W., and Ridder, G. (2014). Complementarity and aggregate implications of assortative matching: A nonparametric analysis. *Quantitative Economics*, **5**:29–66.

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, **112**:564–569.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: A review and study of power. *Statistics in Medicine*, **2**:243–251.

Greenland, S. (2003). The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of American Statistical Association*, **98**:47–54.

Greenland, S. (2005). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society Series A*, **168**:267–308.

Greenland, S. (2009). Interactions in epidemiology: Relevance, identification and estimation. *Epidemiology*, **20**:14–17.

Greenland, S., and Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, **31**:1030–1037.

Greenland, S., and Maldonado, G. (1994). The interpretation of multiplicative-model parameters as standardized parameters. *Statistics in Medicine*, **13**:989–999.

Greenland, S., and Poole, C. (1988). Invariants and noninvariants in the concept of interdependent effects. *Scandinavian Journal of Work, Environment and Health*, **14**:125–129.

Greenland, S., Lash, T. L., and Rothman, K. J. (2008). Concepts of interaction. In: K. J. Rothman, S. Greenland, and T. L. Lash, editors. *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott Williams and Wilkins, Chapter 5.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, **14**:29–46.

Greiner, D., and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *The Review of Economics and Statistics*, **93**:775–785.

Goetgeluk, S., Vansteelandt, S., and Goetghebeur, E. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B*, **70**:1049–1066.

Hacker, P. M. S. (1996). *Wittgenstein: Mind and Will. An Analytical Commentary on the Philosophical Investigations*, Volume 4. Oxford: Blackwell.

Hafeman, D. (2008). A sufficient cause based approach to the assessment of mediation. *European Journal of Epidemiology*, **23**:711–721.

Hafeman, D. M. (2009). "Proportion explained": A causal interpretation for standard measures of indirect effect? *American Journal of Epidemiology*, **170**:1443–1448.

Hafeman, D. M. (2011). Confounding of indirect effects: A sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. *American Journal of Epidemiology*, **174**:710–717.

Hafeman, D. M., and Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology*, **38**:838–845.

Hafeman, D. M., and VanderWeele, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, **22**:753–764.

Hall, N., and Paul, L. A. (2003). Causation and preemption. In: P. Clark and K. Hawley, editors. *Philosophy of Science Today*. Oxford: Oxford University Press, pp. 100–129.

Halloran, M. E., and Hudgens, M. G. (2012a). Causal inference for vaccine effects on infectiousness. *International Journal of Biostatistics*, **8**:(2) Article 6, DOI: 10.2202/1557-4679.1354.

Halloran, M. E., and Hudgens, M. G. (2012b). Comparing bounds for vaccine effects on infectiousness. *Epidemiology*, **23**:931–932.

Halloran, M. E., and Struchiner, C. J. (1991). Study designs for dependent happenings. *Epidemiology*, **2**:331–338

Halloran, M. E., and Struchiner, C. J. (1995). Causal inference for infectious diseases. *Epidemiology*, **6**:142–151.

Halpern, J. Y., and Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal of the Philosophy of Science* **56**:843–887.

Han, S. S., Rosenberg, P. S., Garcia-Closas, M., Figueroa, J. D., Silverman, D., Chanock, S. J., Rothman, N., and Chatterjee, N. (2012). Likelihood ratio test for detecting gene (G)–environment (E) interactions under an additive risk model exploiting G–E independence for case–control data. *American Journal of Epidemiology*, **176**:1060–1067.

Hayden, D., Pauler, D. K., and Schoenfeld, D. (2005). An estimator for treatment comparisons amongst survivors in randomized trials. *Biometrics*, **61**:305–310.

Heckman, J. J., and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*. **96**:4730–4734.

Hernán, M. A. (2005). Invited commentary: Hypothetical interventions to define causal effects: Afterthought or prerequisite? *American Journal of Epidemiology*, **162**:618–620.

Hernán, M. A., and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, **17**:360–372.

Hernán, M. A., and Robins, J. M. (2015). *Causal Inference*. Chapman Hall, in press.

Hernán, M. A., and VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, **22**:368–377.

Hernán, M. A., Brumback B., and Robins J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, **21**:1689–1709.

Hernán, M. A., Hernández-Diaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, **15**:615–625.

Hernández-Díaz, S., Schisterman, E. F., and Hernán, M. A. (2006). The birth weight "paradox" uncovered? *American Journal of Epidemiology*, **164**:1115–1120.

Hicks, R., and Tingley, D. (2011). Causal mediation analysis. *Stata Journal*, **11**:605–619.

Hiddleston, E. (2005). Causal powers. *British Journal of the Philosophy of Science*, **56**: 27–59.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, **58**:295–300.

Hilt, B., Langård, S., Lund-Larsen, P. G. and Lien, J. T. (1986). Previous asbestos exposure and smoking habits in the county of Telemark, Norway—A cross-sectional population study. *Scandinavian Journal of Work, Environment and Health*, **12**:561–566.

Hindoff, L. A., Sethupathy, P., Junkins, H. A., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences U.S.A.*, **106**:9362–9367.

Hoffmann, K., Heidemann, C., Weikert, C., Schulze, M. B., and Boeing, H. (2006). Estimating the proportion of disease due to classes of sufficient causes. *American Journal of Epidemiology*, **163**:76–83.

Hoffmann, T. J., Lange, C., Vansteelandt, S., and Laird, N. M. (2009). Gene–environment interaction tests for dichotomous traits in trios and sibships. *Genetic Epidemiology*, **33**:691–699.

Hogan, M. D., Kupper, L. L., Most, B. M., and Haseman, J. K. (1978). Alternatives to Rothmanís approach for assessing syerngism (or antagonism) in cohort studies. *American Journal of Epidemiology*, **108**:60–67.

Holland, P. W. (1988). Causal inference, path analysis, and recursive strutural equations models. In: C. C. Clogg, editor. *Sociological Methodology*. Washington, DC: American Sociological Association, pp. 449–484.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proceedings of the American Statistical Association, Biometrics Section, Alexandria, VA: American Statistical Association*, pp. 2401–2415.

Hong, G., and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, **101**:901–910.

Hosmer, D. W., Lemeshow, S. (1992). Confidence interval estimation of interaction. *Epidemiology*, **3**:452–456.

Huang, Y., and Gilbert, P. B. (2011). Comparing biomarkers as principal surrogate endpoints. *Biometrics*, **67**:1442–1451.

Hudgens, M. G., and Halloran, M. E. (2006). Causal vaccine effects on binary post-infection outcomes. *Journal of the American Statistical Association*, **101**:51–64.

Hudgens, M. G., and Halloran, M. E. (2008). Towards causal inference with interference. *Journal of the American Statistical Association*, **103**:832–842, 2008.

Hudgens, M. G., Hoering, A., and Self, S. G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine*, **22**:2281–2298.

Hume, D. (1739). *A Treatise of Human Nature*.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Reprinted, 1958, LaSalle, IL: Open Court Press.

Hung, R. J., McKay, J. D., Gaborieau, V., et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, **452**(7187):633–637.

Hwang S.-J., Beaty, T. H., Liang, K.-Y., Coresh, J., and Khoury, M. J. (1994). Minimum sample size esitimation to detect gene–environment interaction in case-control designs. *American Journal of Epidemiology*, **140**:1029–1037.

Hyman, H. H. (1955). *Survey Design and Analysis: Principles, Cases and Procedures.* Glencoe, IL: Free Press.

Imai, K. (2008). Sharp bounds on causal effects in randomized experiments with "truncation-by-death." *Statistics and Probability Letters*, **78**:144–149.

Imai, K., and Yamamoto, T. (2012). Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Analysis*, **21**:141–171.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, **15**(4):309–334.

Imai, K., Keele L.,Tingley, D., and Yamamoto, T. (2010b). Causal mediation analysis using R, In: H. D. Vinod, editor. *Advances in Social Science Research Using R*. New York: Springer, pp. 129–154.

Imai, K., Keele, L., and Yamamoto, T. (2010c). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, **25**:51–71.

Imai, K., Tingley, D., and Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms (with discussion). *Journal of the Royal Statistical Society, Series A*, **176**:5–51.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2014). Practical implications of theoretical results for causal mediation analysis. *Psychological Methods*, in press.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, **93**:126–132.

Imbens, G. W., and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, **25**:305–327.

Imbens, G., and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press, in press.

Jaccard, J. J. (2001). *Interaction Effects in Logistic Regression*. Thousand Oaks, CA: SAGE Publications.

Jaccard, J. J., and Turrisi, R. (2003). *Interaction Effects in Multiple Regression*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

James, L. R., and Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, **69**:307–321.

Jemiai, Y., Rotnitzsky, A., Shepherd, B. E., and Gilbert, P. B. (2007). Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. *Journal of the Royal Statistical Society, Series B*, **69**:879–902.

Jiang, Z. and VanderWeele, T. J. (2014). When is the difference method conservative for mediation? *American Journal of Epidemiology*, in press.

Jiang, Z., and VanderWeele, T. J. (2015a). Causal mediation analysis in the presence of a mismeasured outcome. *Epidemiology*, in press.

Jiang, Z., and VanderWeele, T. J. (2015b). Causal mediation analysis in the presence of a misclassified binary exposure. Technical Report.

Jo, B., Stuart, E. A., MacKinnon, D. P., and Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, **46**:425–452.

Joffe, M. M., and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics*, **65**:530–538.

Joffe, M., Small, D., and Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*, **22**:74–97.

Ju, C., and Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *Journal of the Royal Statistical Society: Series B*, **72**:129–142.

Judd, C. M. and Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, **5**:602–619.

Katan, M. B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, **1**:507–508.

Kaufman, J. S. (2008). Epidemiologic analysis of racial/ethnic disparities: Some fundamental issues and a cautionary example. *Social Science and Medicine*, **66**:1659–1669.

Kaufman, J. S., MacLehose, R. F., and Kaufman, S. (2004). A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives and Innovations*, **1**:4.

Kaufman, S., Kaufman, J. S., MacLehose, R. F., Greenland, S., and Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine*, **24**:1683–1702.

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference*, **139**:3473–3487.

Kenny, A. J. P. (1976). *Will, Freedom and Power*. Oxford: Blackwell Publishers.

Kenny, D. A., and Judd, C. M. (2014). Power anomalies in testing mediation. Psychological Science, **25**:334–339.

Khoury, M. J., and Wacholder, S. (2009). From Genome-wide association studies to gene–environment-wide interaction studies—Challenges and opportunities. *American Journal of Epidemiology*, **169**:227–230.

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of American Statistical Association*, **50**:1168–1194.

Knol, M. J., and VanderWeele, T. J. (2012). Guidelines for presenting analyses of effect modification and interaction. *International Journal of Epidemiology*, **41**:514–520.

Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E., and Geerlings, M. I. (2007). Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *International Journal of Epidemiology*, **36**:1111–1118.

Knol, M. J., Vandenbroucke, J. P., Scott, P., and Egger, M. (2008). What do case–control studies estimate? Survey of methods and assumptions in published case–control research. *American Journal of Epidemiology*, **168**:1073–1081.

Knol, M. J., Egger, M., Scott, P., Geerlings, M. I., and Vandenbroucke, J. P. (2009). When one depends on the other: reporting of interaction in case–control and cohort studies. *Epidemiology*, **20**:161–166.

Knol, M. J., VanderWeele, T. J., Groenwold, R. H. H., Klungel, O. H., Rovers, M. M., and Grobbee, D. E. (2011). Estimating measures of interaction on an additive scale for preventive exposures. *European Journal of Epidemiology*, **26**:433–438.

Knol, M. J., le Cessie, S., Algra, A., Vandenbroucke, J. P., and Groenwold, R. H. H. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: Alternatives to logistic regression. *Canadian Medical Association Journal*, **184**:895–899.

Kotelchuck, M. (1994). An evaluation of the Kessner adequacy of prenatal care index and a proposed adequacy of prenatal care utilization index. *American Journal of Public Health*, **84**:1414–1420.

Kraemer, H. C., Kiernan, M., Essex, M., and Kupfer, D. J. (2008). How and why the criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, **27**(2 Suppl):S101–S108.

Kraft, P. (2004). Multiple comparisons in studies of gene × gene and gene × environment interaction. *American Journal of Human Genetics*, **74**:582–585.

Kraft, P., Yen, Y.C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene–environment interaction to detect disease susceptibility loci. *Human Heredity*, **63**:111–119.

Kremer, M., and Levy, D. (2008). Peer effects and alcohol use among college students. *Journal of Economic Perspectives*, **22**:189–206.

Kuroki, M., and Miyakawa, M. (1999). Identifiability criteria for causal effects of joint interventions. *Journal of the Japan Statistical Society*, **29**:105–117.

Kroenke, K., West, S. L., Swindle, R., Gilsenan, A., Eckert, G. J., Dolor, R., Stang, P., Zhou, X. H., Hays, R., and Weinberger, M. (2001). Similar effectiveness of paroxetine, fluoxetine, and sertraline in primary care: A randomized trial. *JAMA*, **286**(23): 2947–2955.

Kuss, O., Schmidt-Pokrzywniak, A., and Stang, A. (2010). Confidence intervals for the interaction contrast ratio. *Epidemiology*, **21**:273–274.

Kuyvenhoven, J. P., Veenendaal, R. A., and Vandenbroucke, J. P. (1999). Peptic ulcer bleeding: Interaction between non-steroidal anti-inflammatory drugs, *Helicobacter pylori* infection, and the ABO blood group system. *Scandinavian Journal of Gastroenterology*, **34**:1082–1086.

Lake, S., and Laird, N. (2004). Tests of gene–environment interaction for case–parent triads with general environmental exposures. *Annals of Human Genetics*, **68**:55–64.

Lange, T., and Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology*, **22**:575–581.

Lange, T., Vansteelandt, S., and Bekaert, M. (2012). A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, **176**:190–195.

Lange, T. Rasmussen, M., and Thygessen, L. C. (2014). Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, **179**: 513–518.

Lango Allen, H., Estrada, K., Lettre, G., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**:832–838.

Lauritzen, S. L. (2004). Discussion on causality. *Scandinavian Journal of Statistics*, **31**:189–192.

Lawlor, D. A. (2011). Biological interaction: Time to drop the term? *Epidemiology*, **22**:148–150.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, **27**:1133– 1163.

le Cessie, S., Debeij, J., Rosendaal, F. R., Cannegieter, S. C., and Vandenbroucke J. (2012). Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, **23**:551–560.

Le Marchand, L., Derby, K. S., Murphy, S. E., et al. (2008). Smokers with the CHRNA lung cancer-associated variants are exposed to higher levels of nicotine equivalents and a carcinogenic tobacco-specific nitrosamine. *Cancer Research*, **68**:9137–9140.

Leibniz, G. (1714). *Monadology*. In: P. P Wiener, editor. *Leibniz Slections*. New York: Charles Scribner's Sons, 1951.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Li, J., and Chan, I. S. (2006). Detecting qualitative interactions in clinical trials: An extension of range test. *Journal of Biopharmaceutical Statistics*, **16**:831–841.

Li, R., and Chambless, L. (2007). Test for additive interaction in proportional hazards models. *Annals of Epidemiology*, **17**:227–236.

Li, Y., et al. (2010a). Genetic variants and risk of lung cancer in never smokers: A genome-wide association study. *Lancet Oncology*, **11**:321–330.

Li, Y. Taylor, J. M. G., and Elliott, M. R. (2010b). A Bayesian approach to surrogacy assessment using principal stratification in clincial trials. *Biometrics*, **66**, 523–531.

Lin, D. Y., Fleming, T. R., and DeGruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, **16**:1515–1527.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, **54**:948–963.

Lindström, S., Yen, Y.-C., Spiegelman, D., and Kraft, P. (2009). The impact of gene–environment dependence and misclassification in genetic association studies incorporating gene–environment interactions. *Human Heredity*, **68**, 171–181.

Lipton, P. (2009). Causation and explanation. In: H. Beebee, C. Hitchcock, and P. Menzies, editors. *Oxford Handbook of Causation*. Oxford: Oxford University Press, pp. 619–631.

Little, R. J., Long, Q., and Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*, **65**:640–649.

Liu, L., and Hudgens, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statististical Association*, **109**:288–301.

Lundberg, M., Fredlund, P., Hallqvist, J., Diderichsen, F. (1996). A SAS program calculating three measures of interaction with confidence intervals. *Epidemiology*, **7**:655–656.

Luo, X., Small, D. S. Li, C. R., and Rosenbaum, P. R. (2012). Inference with interference between units in an fMRI experiment of motor inhibition. *Journal of the American Statististical Association*, **107**:530–541.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social-network analyses. *Statistics, Politics and Policy*, **2**(1):1–26, Article 2.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, **67**:1–25.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, **2**:245–255.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. New York: Erlbaum.

MacKinnon, D. P., and Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, **17**:144–158.

MacKinnon, D. P., Warsi, G., and Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, **30**:41–62.

MacMahon, B., and Pugh, T. F. (1968). Causes and entities of disease. In: D. W. Clark and B. MacMahon, editors. *Preventive Medicine*. Boston: Little, Brown and Company, pp. 11–18.

Maity, A., Carroll, R. J., Mammen, E., and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene–environment interactions, *Journal of the Royal Statistical Society, Series B*, **71**:75–96.

Maldonado, G., and Greenland, S. (1993). Interpreting model coefficients when the true model form is unknown. *Epidemiology*, **4**:310–318.

Mann, J., McDermott, S., Griffith, M., Hardin, J., and Gregg, A. (2011). Uncovering the complex relationship between pre-eclampsia, preterm birth and cerebral palsy. *Paediatric and Perinatal Epidemiology*, **25**:100–110.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, **80**:319–323.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, **60**:531–542.

Manski, C. F. (1997). Monotone treatment response. *Econometrica*, **65**:1311–1334.

Manski, C. F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, **14**:115–136.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer.

Manski, C. F. (2013). Identification of treatment response with social interactions. *Econometrical Journal*, **16**:S1–S23.

Martinussen, T., and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Berlin: Springer.

Martinussen, T., Vansteelandt, S., Gerster, M., and von Bornemann Hjelmborg, J. (2011). Estimation of direct effects for survival data by using the Aalen additive hazards model. *Journal of the Royal Statistical Society, Series B*, **73**:773–788.

Mattei, A., and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society, Series B*, **73**:729–752.

McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, **26**:2331–2347.

Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science*, **71**:820–832.

Miettinen, O. S. (1982). Causal and preventive interdependence: elementary principles. *Scandinavian Journal of Work Environment and Health*, **8**:159–168.

Mill, J. S. (1911). *A System of Logic: Ratiocinative and Inductive*. London: Longmans, Green.

Millar, E. V., Watt, J. P., Bronsdon, M. A., Dallas, J., Reid, R., Santosham, M., and O'Brien, K. L. (2008). Indirect effect of 7-valent pneumococcal conjugate vaccine on pneumococcal colonization among unvaccinated household members. *Clinical Infectious Diseases*, **47**:989–996.

Miller, D. P., Liu, G., De Vivo, I., et al. (2002). Combinations of the variant genotypes of GSTP1, GSTM1, and p53 are associated with an increased lung cancer risk. *Cancer Research*, **62**:2819–2823.

Moore, J. H., and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *American Journal of Human Genetics*, **85**(3):309–320.

Moore, T. (1995). *Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster*. New York: Simon and Schuster.

Moorman, X. Y., Calingaret, B., Palmieri, R. T., Iversen, E. S., Bentley, R. C., Halabi, S., Berchuck, A., and Schildkraut, J. M. (2008). Hormonal risk factors for ovarian cancer in premenopausal and postmenopausal women. *American Journal of Epidemiology*, **167**:1059–1069.

Morgan, S. L., and Winship, C. (2007). *Counterfactuals and Causal Inference*. Cambridge, UK: Cambridge University Press.

Morris, A. P., Voight, B. F., Teslovich, T. M, et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, **44**:981–990.

Morrow, G. R., Hickok, J. T., Roscoe, J. A., Raubertas, R. F., Andrews, P. L. R., Flynn, P. J., Hynes, H. E., Banerjee, T. K., Kirshner, J. J., and King, D. K. (2003). Differential effects of paroxetine on fatigue and depression: A randomized, double-bind trial from the University of Rochester Cancer Center Community Clinical Oncology Program. *Journal of Clinical Oncology*, **21**:4635–4641.

Mukherjee, B., and Chatterjee, N. (2008). Exploiting gene–environment independence for analysis of case–control studies: An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. *Biometrics*, **64**:685–694.

Mukherjee, B., Ahn, J., Gruber, S. B., Chatterjee, N. (2012). Testing gene–environment interaction in large-scale case–control association studies: Possible choices and comparisons. *American Journal of Epidemiology*, **175**:177–190.

Muller, D., Judd, C. M., and Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, **89**: 852–863.

Mumford, S. L., Schisterman, E. F., Siega-Riz, A. M., Gaskins, A. J., and VanderWeele, T. J. (2011). Effect of dietary fiber intake on lipoprotein cholesterol levels independent of estradiol in healthy premenopausal women. *American Journal of Epidemiology*, **173**:145–156.

Murcray, C. E., Lewinger J. P., and Gauderman, W. J. (2009). Gene–environment interaction in genome-wide association studies. *American Journal of Epidemiology*, **169**:219–226.

Musser, G. (2004). Four keys to cosmology. *Scientific American*, **290**(2):42–43.

Muthén, B. (2012). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Technical Report.

Nandi, A. Glymour, M. M., Kawachi, I., and VanderWeele, T. J. (2012). Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes and stroke. *Epidemiology*, **23**:223–232.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed, translators) in *Statistical Science*, **5**:463–472.

Neyman, J. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society, II* **2**, 107–154.

Nie, L., Chu, H., Li, F., and Cole, S. R. (2010). Relative excess risk due to interaction: Resampling-based confidence intervals. *Epidemiology*, **21**:552–556.

Noel, H., and Nyhan, B. (2011). The "unfriending" problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Social Networks*, **33**:211–218.

Norton, E. C., Wang, H., and Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, **4**:154–167.

Novick, L. R., and Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, **111**:455–485.

Ogburn, E. L. and VanderWeele, T. J. (2012). Analytic results on the bias due to nondifferential misclassification of a binary mediator. *American Journal of Epidemiology*, **176**:555–561.

Ogburn, E. L., and VanderWeele, T. J. (2014a). Causal diagrams for interference and contagion. *Statistical Science*, in press.

Ogburn, E. L., and VanderWeele, T. J. (2014b). Vaccines, contagion, and social networks. Technical Report.

Oppy, G. (2000). On "A new cosmological argument." *Religious Studies*, **36**:345–353.

Page, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, **5**(3): 215–244.

Patel, J. K., and Campbell, B. (1996). *Handbook of the Normal Distribution*. New York: Marcel Dekker.

Pan, G., and Wolfe, D. A. (1997). Test for qualitative interaction of clinical significance. *Statistics in Medicine*, **16**:1645–1652.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, **82**:669–710.

Pearl, J. (2001). Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann, pp. 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press.

Pearl, J. (2011). Principal stratification—A goal or a tool? *International Journal of Biostatistics*, **7**(1):Article 20.

Pearl, J. (2012). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, **13**:426–436.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, in press.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, **17**:276–284.

Petersen, M. L., Deeks, S. G., Martin, J. N., and van der Laan, M. J. (2007). History-adjusted marginal structural models for estimating time-varying effect modification. *American Journal of Epidemiology*, **166**:985–993.

Peto, R. (1982). Statistical aspects of cancer trials. In: K. E. Halnan, editor. *Treatment of Cancer*. London: Chapman and Hall, pp. 867–871.

Phillips, P.C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**:855–867.

Piantadosi, S., and Gail, M. H. (1993). A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, **12**:1239–1248.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Statistics in Medicine*, **13**:153–162.

Pierce, B. L., and Ahsan, H. (2010). Case-only genome-wide interaction study of disease risk, prognosis and treatment. *Genetic Epidemiology*, **34**:7–15.

Pierce, B. L. and VanderWeele, T. J. (2012). The effect of non-differential measurement error on bias, precision, and power in Mendelian randomization studies. *International Journal of Epidemiology*, **41**:1383–1393.

Politis, D. N., and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, **22**:2031–2050.

Poole, C. (2010). On the origin of risk relativism. *Epidemiology*, **21**:3–9.

Preacher, K. J., and Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers* **36**:717–731.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, **42**(1):185–227.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, **8**:431–440.

Préziosi, M.-P., and Halloran, M. E. (2003). Effects of pertussis vaccination on transmission: Vaccine efficacy for infectiousness. *Vaccine*, **21**:1853–1861.

Psillos, S. (2002). *Causation and Explanation*. Montreal: McGill-Queen's University Press.

Psillos, S. (2009). Regularity theories. In: H. Beebee, C. Hitchcock, and P. Menzies, editors., *Oxford Handbook of Causation*. Oxford: Oxford University Press, pp. 619–631.

Ramsahai, R. R. (2013). Probabilistic causality and detecting collections of interdependence patterns. *Journal of the Royal Statistical Society Series B*, **75**:705–723.

Ramsey, F. P. (1928). Universal of law and of fact. In: D. H. Mellor, editor. *Foundations: Essays*. New York: Columbia University Press.

Reichenbach, B. (2013). Cosmological argument. In: Edward N. Zalta, editor. *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition). Accessed February 14, 2014: http://plato.stanford.edu/archives/spr2013/entries/cosmological-argument.

Reinisch, J., Sanders, S., Mortensen, E., and Rubin D. B. (1995). In-utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association*, **274**:1518–1525.

Richardson, D. B., and Kaufman, J. S. (2009). Estimation of the relative excess risk due to interaction and associated confidence bounds. *American Journal of Epidemiology*, **169**:756–60.

Richardson, T. S., Evans, R. J., and Robins, J. M. (2011). Transparent parametrizations of models for potential outcomes. In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors. Oxford: Oxford University Press *Bayesian Statistics* **9**, 569–610.

Richardson, T. S. and Robins, J. M. (2013). Single World Intervention Graphs: A Primer. Second UAI Workshop on Causal Structure Learning, Bellevue, Washington.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**:1393–1512.

Robins, J. M. (1999a). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: C. Glymour, and G. F. Cooper, editors. *Computation, Causation, and Discovery*. Menlo Park, CA/Cambridge, MA: AAAI Press/The MIT Press, pp. 349–405.

Robins, J. M. (1999b). Marginal structural models versus structural nested models as tools for causal inference. In: M. E. Halloran and D. Berry, editors. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. IMA Volume 116. New York: Springer-Verlag, pp. 95–134.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In: P. Green, N. L. Hjort, and S. Richardson, editors. *Highly Structured Stochastic Systems*, New York: Oxford University Press, pp. 70–81.

Robins, J. M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**:143–155.

Robins, J. M., and Greenland, S. (2000). Comment on "Causal inference without counterfactuals." *Journal of the American Statistical Association*, **95**:477–482.

Robins, J. M., and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In: P. Shrout, editor. *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. Oxford University Press.

Robins J. M., Hernán M. A., and Brumback B. (2000a). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**:550–560.

Robins, J. M., Scharfstein, D., and Rotnitzky, A. (2000b). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: E. Halloran, and D. Berry, editors. *Statistical Models for Epidemiology, the Environment, and Clinical Trials*. New York: Springer-Verlag, pp. 1–95.

Robins, J. M., Hernán, M. A., and Rotnitzky, A. (2007a). Effect modification by time-varying covariates. *American Journal of Epidemiology*, **166**:994–1002.

Robins, J. M., Rotnitzky, A., and Vansteelandt, S. (2007b). Discussion of "Principal stratification designs to estimate input data missing due to death." *Biometrics*, **63**:650–654.

Robins, J.M., VanderWeele, T.J., and Gill, R. (2015). A proof of Bell's inequality in quantum mechanics using causal interactions. *Scandinavian Journal of Statistics*, in press.

Robinson, L., and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Rev*iew, **59**:227–240.

Rosenbaum, P. R. (2002). *Observational Studies*. 2nd edn. New York: Springer-Verlag.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, **102**:191–200.

Rosenbaum, P. R., and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B*, **45**:212–218.

Rosenbaum, P. R., and Rubin, D. B. (1983b). The central role of propensity score in observational studies for causal effects. *Biometrika*, **70**:41–55.

Rosenbaum, P. R., and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**:33–38.

Rothman, K. J. (1974). Synergy and antagonism in cause-e§ect relationships. *American Journal of Epidemiology*, **99**:385–388.

Rothman K. J. (1976). Causes. *American Journal of Epidemiology*, **104**:587–592.

Rothman, K. J. (1978). Estimation versus detection in the assessment of synergy. *American Journal of Epidemiology*, **108**:9–11.

Rothman, K. J. (1986). *Modern Epidemiology*. 1st edition Little, Brown and Company, Boston.

Rothman, K. J., and Greenland, S. editors (1998). *Modern Epidemiology*, 2nd edn. Philadelphia: Lippincott.

Rothman, K. J., Greenland, S., and Walker, A. M. (1980). Concepts of interaction. *American Journal of Epidemiology*, **112**:467–470.

Rothman, K. J., Greenland, S., and Lash T. L., editors. (2008). *Modern Epidemiology*, 3rd edn. Philadelphia: Lippincott.

Rowe, W. L. (1997). Cosmological arguments. In: P. L. Quinn and C. Taliaferro, editors. *A Companion to the Philosophy of Religion.* Oxford: Blackwell Publishers, pp. 331–337.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**:688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, **6**:34–58.

Rubin, D. B. (1980). Comment on: "Randomization analysis of experimental data: The Fisher randomization test by D. Basu." *Journal of the American Statistical Association*, **75**:591–593.

Rubin, D. B. (1986). Which if's have causal answers? *Journal of the American Statistical Association*, **81**:961–962.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistcal Science*, **5**:472–480.

Rubin, D. B. (2004). Direct and indirect effects via potential outcomes. *Scandinavian Journal of Statistics*, **31**:161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decision. *Journal of the American Statistical Association*, **100**:322–331.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death (with discussion). *Statistical Science*, **21**:299–321.

Rubin, D. B. (2010). Direct and indirect causal effects: A helpful distinction? International Conference on Applied Statistics 2010, September 19. http://videolectures.net/as2010_rubin_dic/. Accessed: November 11, 2013.

Rubin, D. B. (2011). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, **2**:808–840.

Rubin, D. B. (2013). Discussion of: Experimental designs for identifying causal mechanisms (Imai, K., Tingley, D., and Yamamoto, T.). *Journal of the Royal Statistical Society, Series A*, **176**:45.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics*, **116**:681–704.

Sanchez-Vaznaugh, E. V., Kawachi, I., Subramanian, S. V., Sanchez, B. N., and Acevedo-Garcia, D. (2009). Do socioeconomic gradients in body mass index vary by race/ethnicity, gender, and birthplace? *American Journal of Epidemiology*, **169**:1102–1112.

Saracci, R. (1980). Interaction and synergism. *American Journal of Epidemiology*, **112**:465–466.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association*, **94**:1096–1120.

Schlesselman, J. J. (1978). Assessing effects of confounding variables. *American Journal of Epidemiology*, **108**:3–8.

Schunkkert, H., König, I. R., Kathiresan, S., et al. (2011). Large-scale association analysis indentifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, **43**:333–338.

Schwartz, B. S., Stewart, W. F., Bolla, K. I., et al. (2000). Past adult lead exposure is associated with longitudinal decline in cognitive function. *Neurology*, **55**:1144–1150.

Scott, R. A., Lagou, V., Welch, R. P., et al. (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genetics*, **44**:991–1005.

Sen, M., and Wasow, O. (2013). How and when to make causal claims based on race or ethnicity. Technical Report.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shalizi, C. R., Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, **40**:211–239.

Shepherd, B. E., Gilbert, P. B., Jemiai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*, **62**:332–342.

Shepherd, B. E., Gilbert, P. B., and Lumley, T. (2007). Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. *Journal of the American Statistical Association*, **102**:573–582.

Shpitser, I., and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, **9**:91941–1979.

Shpitser, I., and VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *International Journal of Biostatistics*, **7**, Article **16**:1–24.

Siemiatycki, J., and Thomas, D. C. (1981). Biological models and statistical interactions: An example from multistage carcinogenesis. *International Journal of Epidemiology*, **10**:383–387.

Silvapulle, M. J. (2001). Tests against qualitative interaction: Exact critical values and robust tests. *Biometrics*, **57**:1157–1165.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, **28**:558–571.

Skrondal, A. (2003). Interaction as departure from additivity in case–control studies: A cautionary note. *American Journal of Epidemiology*, **158**:251–258.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, **31**:361–395.

Snijders, T. A. B. (2005). Models for longitudinal network data. In: P. J. Carrington, J. Scott, and S. S. Wasserman, editors. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, Chapter 11.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In: S. Leinhart, editor. *Sociological Methodology*. San Francisco: Jossey-Bass, pp. 290-312.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, **101**:1398–1407.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, **33**:230–251.

Song, X., and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics*, **60**:874–883.

Speliotes, E. K., Willer, C. J., Berndt, S. I., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, **42**:937–948.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Steglich, C. E., Snijders, T. A. and Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, **40**:329–393.

Stern, M. C., Johnson, L. R., Bell, D. A., and Taylor, J. A. (2002a). XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. *Cancer Epidemiology, Biomarkers and Prevention*, **11**:1004–1011.

Stern, M. C., Umbach, D. M., Lunn, R. M., and Taylor, J. A. (2002b). DNA repair gene XCRR3 codon 241 polymorphism, its interaction with smoking and XRCC1 polymorphisms, and bladder cancer risk. *Cancer Epidemiology, Biomarkers and Prevention*, **11**:939–943.

Sterne, J. A., and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, **54**:1046–1055.

Stewart, W. F., Schwartz, B. S., Davatzikos, C., et al. (2006). Past adult lead exposure is linked to neurodegeneration measured by brain MRI. *Neurology*, **66**:1476–1484.

Strain, P. S., Shores, R. E., and Kerr, M. M. (1976). An experimental analysis of "spillover" effects on the social interaction of behaviorally handicapped preschool children. *Journal of Applied Behavior Analysis*, **9**:31–40.

Strong, V., Waters, R., Hibberd, C., Murray, G., Wall, L., Walker, J., McHugh, G., Walker, A., and Sharpe. M. (2008). Management of depression for people with cancer (SMaRT oncology 1): A randomised trial. *Lancet*, **372**(9632):40–48.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, **162**:279–289.

Susser, M. (1973). *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology*. New York: Oxford University Press.

Suzuki, E. and VanderWeele, T. J. (2014). Compositional epistasis: An epidemiologic perspective. In: J. Moore, and S. Williams, editors. *Epistasis and Genetic Architecture*. Springer, in press.

Suzuki, E., Yamamoto, E., and Tsuda, T. (2011). Identification of operating mediation and mechanism in the sufficient-component cause framework. *European Journal of Epidemiology*, **26**:347–57.

Suzuki, E., Evans, D., Chaix, B., and VanderWeele, T. J. (2014). On the "proportion eliminated" for risk differences versus excess relative risks. *Epidemiology*, **25**:309–310.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, **101**:1607–1618.

Taylor, J. M. G., Wang, Y., and Thiebaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate markers. *Biometrics*, **61**:1101–1111.

Tchetgen Tchetgen, E. J. (2010). On the interpretation, robustness, and power of varieties of case-only tests of gene–environment interaction. *American Journal of Epidemiology*, **172**:1335–1338.

Tchetgen Tchetgen, E. J. (2011). On causal mediation analysis with a survival outcome. *International Journal of Biostatistics*, **7**:Article 33, 1–38.

Tchetgen Tchetgen, E. J. (2013). A note on formulae for causal mediation analysis in an odds ratio context. *Epidemiologic Methods*, **2**:21–32.

Tchetgen Tchetgen, E. J., and Kraft, P. (2011). On the robustness of tests of genetic associations incorporating gene–environment interaction when the environmental exposure is misspecified. *Epidemiology*, **22**:257–261.

Tchetgen Tchetgen, E. J., and Lin, S. H. (2013). Robust estimation of pure/natural direct effects with mediator measurement error. Technical Report.

Tchetgen Tchetgen, E. J., and Robins, J. M. (2010). The semi-parametric case-only estimator. *Biometrics*, **66**:1138–1144.

Tchetgen Tchetgen, E. J., and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, **40**:1816–1845.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research—Special Issue on Causal Inference*, **21**:55–75.

Tchetgen Tchetgen, E. J., and VanderWeele, T. J. (2014a). Robustness of measures of interaction to unmeasured confounding. Harvard University Technical Report.

Tchetgen Tchetgen, E. J., and VanderWeele, T. J. (2014b). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, **25**:282–291.

Tchetgen Tchetgen, E. J., Glymour, M. M., Weuve, J., and Shpitser, I. (2012). To weight or not to weight? On the relation between inverse-probability weighting and principal stratification for truncation by death. *Epidemiology*, **23**:644–646.

Tchetgen Tchetgen, E. J., Walter, S., and Glymour, M. M. (2013). Commentary: Building an evidence base for mendelian randomization studies: Assessing the validity and strength of proposed genetic instrumental variables. *International Journal of Epidemiology*, **42**:328–331.

Tein, J.-Y., MacKinnon, D. P. (2003). Estimating mediated effects with survival data. In: H. Yanai, A. O. Rikkyo, K. Shigemasu, Y. Kano, and J. J. Meulman, editors. *New Developments on Psychometrics*. Springer-Verlag Tokyo, Tokyo pp. 405–412.

Ten Have, T. R., and Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research*, **21**:77–107.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, **63**:926–934.

Thomas, D. (2010). Gene–environment-wide association studies: Emerging approaches. *Nature Reviews Genetics*, **11**:259–272.

Thomas, D. C., and Conti, D. V. (2004). Commentary: The concept of "Mendelian Randomization." *International Journal of Epidemiology*, **33**:21–25.

Thomas, W. (1991). Effect modification and the limits of biological inference from epidemiologic data. *Journal of Clinical Epidemiology*, **44**:221–232.

Thompson, W. D. (1991). Effect modification and the limits of biologic inference from epidemiologic data. *Journal of Clinical Epidemiology*, **44**:221–232.

Thorgeirsson, T. E., Geller, F., Sulem, P., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, **452**(7187): 638–642.

Thorgeirsson, T. E., and Stefansson, K. (2010). Commentary: Gene–environment interactions and smoking-related cancers. *International Journal of Epidemiology*, **39**(2):577–579.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, **59**(5):1–38.

Truong, T., Hung, R. J., Amos, C. I., et al. (2010). Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: A pooled analysis from the International Lung Cancer Consortium. *Journal of the National Cancer Institute*, **102**: 959–971.

Umbach, D., and Weinberg, C. (2000). The use of case-parent triads to study joint effects of genotype and exposure. *American Journal of Human Genetics*, **66**:251–261.

Valente, T. W., and Davis. R. L. (1999). Accelerating the diffusion of innovations using opinion leaders. *The Annals of the American Academy of Political and Social Science*, **566**:55–67.

Valente, T. W., Hoffman, B. R., Ritt-Olson, A., Lichtman, K., and Anderson Johnson, C. (2003). Effects of a social-network method for group assignment strategies on peer-led tobacco prevention programs in schools. *American Journal of Public Health*, **93**:1837–1843.

Valeri, L., and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, **18**:137–150.

Valeri, L., and VanderWeele, T. J. (2014). The estimation of direct and indirect causal effects in the presence of a misclassified binary mediator. *Biostatistics*, **15**:498–512.

Valeri, L., Lin, X., and VanderWeele, T. J. (2014). Mediation analysis when the mediator is measured with error and the outcome follows a generalized linear model. *Statistics in Medicine*, in press.

Valeri, L. and VanderWeele, T.J. (2015). SAS macro for causal mediation analysis with survival data. *Epidemiology*, in press.

van der Laan, M. J., and Petersen, M. L. (2008). Direct effect models. *International Journal of Biostatistics*, **4**:Article 23.

Vandenbroucke, J. P., Koster, T., Briët, E., Reitsma, P. H., Bertina, R. M., and Rosendaal, F. R. (1994). Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet*, **344**:1453–1457.

Vandenbroucke, J. P., von Elm, E., Altman, D. G., et al. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *Epidemiology*, **18**:805–835.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, **78**:2957–2962.

VanderWeele, T. J. (2009a). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, **20**:18–26.

VanderWeele, T. J. (2009b). Concerning the consistency assumption in causal inference. *Epidemiology*, **20**:880–883.

VanderWeele, T. J. (2009c). On the distinction between interaction and effect modification. *Epidemiology*, **20**:863–871.

VanderWeele, T. J. (2009d). Sufficient cause interactions and statistical interactions. *Epidemiology*, **20**:6–13.

VanderWeele, T. J. (2009e). Criteria for the characterization of token causation. *Logic and Philosophy of Science*, 115–127.

VanderWeele, T. J. (2009f). Mediation and mechanism. *European Journal of Epidemiology*, **24**:217–224.

VanderWeele, T. J. (2010a). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, **21**:540–551.

VanderWeele, T. J. (2010b). Empirical tests for compositional epistasis. *Nature Reviews Genetics*, **11**:166.

VanderWeele, T. J. (2010c). Epistatic interactions. *Statistical Applications in Genetics and Molecular Biology*, **9**:Article 1, 1–22.

VanderWeele, T. J. (2010d). Response to "On the definition of effect modification," by E. Shahar and D. J. Shahar. *Epidemiology*, **21**:587–588.

VanderWeele, T. J. (2010e). Sufficient cause interactions for categorical and ordinal exposures with three levels. *Biometrika*, **97**:647–659.

VanderWeele, T. J. (2010f). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods and Research*, **38**: 515–544.
VanderWeele, T. J. (2011a). Controlled direct and mediated effects: Definition, identification and bounds. *Scandinavian Journal of Statistics*, **38**:551–563.
VanderWeele, T. J. (2011b). Causal mediation analysis with survival data. *Epidemiology*, **22**:575–581.
VanderWeele, T. J. (2011c). Subtleties of explanatory language: what is meant by "mediation"? *European Journal of Epidemiology*, **26**:343–346.
VanderWeele, T. J. (2011d). Principal stratification—Uses and limitations. *International Journal of Biostatistics*, 7, Article **28**:1–14.
VanderWeele, T. J. (2011e). A word and that to which it once referred: assessing "biologic" interaction. *Epidemiology*, **22**:612–613.
VanderWeele, T. J. (2011f). Causal interactions in the proportional hazards model. *Epidemiology*, **22**:713–717.
VanderWeele, T. J. (2011g). Sample size and power calculations for case-only interaction studies: Formulas for common test statistics. *Epidemiology*, **22**:873–874.
VanderWeele, T. J. (2011h). Sensitivity analysis for contagion effects in social networks. *Sociological Methods and Research*, **40**:240–255.
VanderWeele, T. J. (2012a). Mediation analysis with multiple versions of the mediator. *Epidemiology*, **23**:454–463.
VanderWeele, T. J. (2012b). Interaction tests under exposure misclassification. *Biometrika*, **99**:502–508.
VanderWeele, T. J. (2012c). Sample size and power calculations for additive interactions. *Epidemiologic Methods*, **1**:159–188.
VanderWeele, T. J. (2012d). The sufficient cause framework in statistics, philosophy and the biomedical and social sciences. In: C. Berzuini, P. Dawid, and L. Bernardinelli, editors. *Causality: Statistical Perspectives and Applications.* Hoboken, NJ: John Wiley. pp. 180–191.
VanderWeele, T.J. (2012e). Structural equation modeling in epidemiologic analysis. *American Journal of Epidemiology*, **176**:608–612.
VanderWeele, T. J. (2013a). Policy-relevant proportions for direct effects. *Epidemiology*, **24**:175–176.
VanderWeele, T. J. (2013b). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, **24**:224–232.
VanderWeele, T. J. (2013c). Unmeasured confounding and hazard scales: Sensitivity analysis for total, direct and indirect effects. *European Journal of Epidemiology*, **28**:113–117.
VanderWeele, T. J. (2013d). Surrogate measures and consistent surrogates (with discussion). *Biometrics*, **69**:561–681.
VanderWeele, T. J. (2013e). Reconsidering the denominator of the attributable proportion for additive interaction. *European Journal of Epidemiology*, **28**:779–784.
VanderWeele, T. J. (2014). A unification of mediation and interaction: a four-way decomposition. *Epidemiology*, **25**:749–761.
VanderWeele, T. J., and An, W. (2013). Social networks and causal inference. In: S. L. Morgan, editor. *Handbook of Causal Analysis for Social Research*. New York: Springer, Chapter 17, pp. 353–37.
VanderWeele, T. J., and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. *Epidemiology*, **22**:42–52.

VanderWeele, T. J., and Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator–outcome confounders. *Epidemiology, Biostatistics, and Public Health*, **11**(2):1–16. DOI:10.2427/9027.

VanderWeele, T. J., and Hernán, M. A. (2006). From counterfactuals to sufficient causes, and vice versa. *European Journal of Epidemiology*, **21**:855–858.

VanderWeele, T.J. and Hernán, M. A. (2012). Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for non-manipulable exposures with application to the effects of race and sex. In: C. Berzuini, P. Dawid, and L. Bernardinelli, editors. *Causal Inference: Statistical Perspectives and Applications*. Hoboken, NJ: John Wiley & Sons, pp. 101–113.

VanderWeele, T. J., and Hernán, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, **1**:1–20.

VanderWeele, T. J. and Hernández-Díaz, S. (2011). Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? *Paediatric and Perinatal Epidemiology*, **25**:111–115.

VanderWeele, T. J., and Knol, M. J. (2011a). The interpretation of subgroup analyses in randomized trials: Heterogeneity versus secondary interventions. *Annals of Internal Medicine*, **154**:680–683.

VanderWeele, T. J. and Knol, M. J. (2011b). Remarks on antagonism. *American Journal of Epidemiology*, **173**:1140–1147.

VanderWeele, T. J., and Knol, M. J. (2014). A tutorial on interaction. *Epidemiologic Methods*, Published Ahead of Print May 27, 2014, DOI 10.1515/em-2013-0005.

VanderWeele, T. J. and Richardson, T. S. (2012). General theory for interactions in sufficient cause models with dichotomous exposures. *Annals of Statistics*, **40**:2128–2161.

VanderWeele, T. J., and Robins, J. M. (2007a). Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology*, **166**:1096–1104.

VanderWeele, T. J., and Robins, J. M. (2007b). The identification of synergism in the sufficient-component cause framework. *Epidemiology*, **18**:329–339.

VanderWeele, T. J., and Robins, J. M. (2007c). Four types of effect modification—A classification based on directed acyclic graphs. *Epidemiology*, **18**:561–568.

VanderWeele, T. J., and Robins, J. M. (2008). Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*, **95**:49–61.

VanderWeele, T. J., and Robins, J. M. (2009a). Minimal sufficient causation and directed acyclic graphs. *Annals of Statistics*, **37**:1437–1465.

VanderWeele, T. J., and Robins, J. M. (2009b). The properties of monotonic effects on directed acyclic graphs. *Journal of Machine Learning Research—Special Topic on Causality*, **10**:699–718.

VanderWeele, T. J., and Robins, J. M. (2010). Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society, Series B*, **72**:111–127.

VanderWeele, T. J., and Robins, J. M. (2012). Stochastic counterfactuals and stochastic sufficient causes. *Statistica Sinica*, **22**:379–392.

VanderWeele, T. J., and Robinson, W. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, **25**:473–484.

VanderWeele, T. J., and Tchetgen Tchetgen, E. J. (2011a). Effect partitioning under interference for two-stage randomized vaccine trials. *Statistics and Probability Letters*, **81**:861–869.

VanderWeele, T. J., and Tchetgen Tchetgen, E. J. (2011b). Bounding the infectiousness effect in vaccine trials. *Epidemiology*, **22**:686–693.
VanderWeele, T. J., and Tchetgen Tchetgen, E. J. (2014). Attributing effects to interactions. *Epidemiology*, **25**:711–722.
VanderWeele, T. J., and Vansteelandt S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, **2**(4):457–468.
VanderWeele, T. J., and Vansteelandt S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, **172**(12):1339–1348.
VanderWeele, T. J., and Vansteelandt, S. (2011). A weighting approach to causal effects and additive interaction in case–control studies: Marginal structural linear odds models. *American Journal of Epidemiology*, **174**:1197–1203.
VanderWeele, T.J., and Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic Methods*, **2**:95–115.
VanderWeele, T. J., Hernán, M. A., and Robins, J. M. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology*, **19**:720–728.
VanderWeele, T. J., Lantos, J. D., Siddique, J., and Lauderdale, D. S (2009). A comparison of four prenatal care indices in birth outcome models: comparable results for predicting small-for-gestational-age outcome but different results for preterm birth or infant mortality. *Journal of Clinical Epidemiology*, **62**:438–445.
VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2010). Marginal structural models for sufficient cause interactions. *American Journal of Epidemiology*, **171**:506–514.
VanderWeele, T. J., Chen, Y., and Ahsan, H. (2011). Inference for causal interactions for continuous exposures under dichotomization. *Biometrics*, **67**:1414–1421.
VanderWeele, T. J., Asomaning, K., Tchetgen Tchetgen, E. J., Han, Y., Spitz, M. R., Shete, S., Wu, X., Gaborieau, V., Wang, Y., McLaughlin, J., Hung, R. J., Brennan, P., Amos, C. I., Christiani, D. C. and Lin, X. (2012a). Genetic variants on 15q25.1, smoking and lung cancer: An assessment of mediation and interaction. *American Journal of Epidemiology*, **175**:1013–1020.
VanderWeele, T. J., Mukherjee, B., and Chen, J. (2012b). Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine*, **31**:2552–2564.
VanderWeele, T. J., Ogburn, E. L., and Tchetgen Tchetgen, E. J. (2012c). Why and when "flawed" social network analyses still yield valid tests of no contagion. *Statistics, Politics, and Policy*, **3**:Article 4, 1–11.
VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Halloran, M. E. (2012d). Components of the indirect effect in vaccine trials: Identification of contagion and infectiousness effects. *Epidemiology*, **23**:285–292.
VanderWeele, T. J., Valeri, L., and Ogburn, E. L. (2012e). The role of misclassification and measurement error in mediation analyses. *Epidemiology*, **23**:561–564.
VanderWeele, T. J., Vandenbroucke, J. P., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012f). A mapping between interactions and interference: Implications for vaccine trials. *Epidemiology*, **23**:285–292.
VanderWeele, T. J., Hong, G., Jones, S., and Brown, J. (2013a). Mediation and spillover effects in group-randomized trials: A case study of the 4R's educational intervention. *Journal of the American Statistical Association*, in press.
VanderWeele, T. J., Ko, Y-.A., and Mukherjee, B. (2013b). Environmental confounding in gene–environment interaction studies. *American Journal of Epidemiology*, in press.
VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelus, M., and Kraft, P. (2014a). Methodological challenges in Mendelian randomization. *Epidemiology*, **25**:427–435.

VanderWeele, T. J., Tchetgen Tchetgen E. J., and Halloran, M. E. (2014b). Interference and sensitivity analysis. *Statistical Science*, in press.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014c). Methods for effect decomposition in the presence of an exposure-induced mediator–outcome confounder. *Epidemiology*, in press. **25**:300–306.

Vansteelandt, S. (2009). Estimating direct effects in cohort and case–control studies. *Epidemiology*, **20**:851–860.

Vansteelandt, S., and VanderWeele, T. J. (2012). Natural direct and indirect effects on the exposed: Effect decomposition under weaker assumptions. *Biometrics*, **68**:1019–1027.

Vansteelandt, S., Bekaert, M., and Lange, T. (2012a). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, **1**:131–158.

Vansteelandt, S., VanderWeele, T. J., and Robins, J. M. (2012b). Semiparametric inference for sufficient cause interactions. *Journal of the Royal Statistical Society, Series B*, **74**:223–244.

Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J., and Robins, J. M., (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, **103**:1693–1704.

Vimaleswaran, K. S., Berry, D. J., Lu, C., et al. (2013). Causal relationship between obesity and vitamin D status: Bi-directional Mendelian randomization analysis of multiple cohorts. *PLoS Medicine*, **10**(2):e1001383. doi: 10.1371/journal.pmed.1001383.

Vittinghoff, E., Sen, S., and McCulloch, C. E. (2009). Sample size calculations for evaluating mediation. *Statistics in Medicine*, **28**:541–557.

Walter S. D., and Holford, T. R. (1978). Additive, multiplicative, and other models for disease risks. *American Journal of Epidemiology*, **108**:341–346.

Wang, W., and Albert, J. M. (2012). Estimation of mediation effects for zero-inflated regression models. *Statistics in Medicine*, **31**:3118–3132.

Weinberg, C. R. (1986). Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. *American Journal of Epidemiology*, **123**:162–173.

Weinberg, C. R., Shi, M., and Umbach, D. M. (2011). A sibling-augmented case-only approach for assessing multiplicative gene–environment interactions. *American Journal of Epidemiology*, **174**:1183–1189.

Wilcox, A. J. (1993). Birthweight and perinatal mortality: The effect of maternal smoking. *American Journal of Epidemiology*, **137**:1098–1104.

Wing, R. R., and Jeffery, R. W. (1999). Benefits of recruiting participants with friends and increasing social support for weight loss and maintenance. *Journal of Consulting and Clinical Psychology*, **67**:132–138.

Wittgenstein, L. (1953). *Philosophical Investigations*, G. E. M. Anscombe, translator. New York: Macmillan.

Wolfson, J., and Gilbert, P. (2010). Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics*, **66**:1153–1161.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 2002.

Wright, R. W. (1988). Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review*. **73**:1001–1077.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**:557–585.

Xu, W. H., Dai, Q., Xiang, Y. B., Long, J. R., Ruan, Z. X., Cheng, J. R., Zheng, W., and Shu, X. O. (2007). Interaction of soy food and tea consumption with CYP19A1 Genetic polymorphisms in the development of endometrial cancer. *American Journal of Epidemiology*, **166**:1420–1430.

Yamamoto, T. (2014). Identification and estimation of causal mediation effects with treatment noncompliance. Technical Report.

Yanagawa, T. (1984). Case–control studies: Assessing the effect of a confounding factor. *Biometrika*, **71**:191–194.

Yang, C.-X., Matsuo, K., Ito, H., Kirose, K., Wakal, K., Saito, T., Shinoda, M., Hatooka, S., Mizutani, K., and Tajima, K. (2005). Esophageal cancer risk by ALDH2 and ADH2 polymorphisms and alcohol consumption: exploration of gene–environment and gene–gene interactions. *Asian Pacific Journal of Cancer Prevention*, **6**:256–262.

Yang, Q., Khoury, M. J. and Flanders, W. D. (1997). Sample size requirements in case-only designs to detect gene–environment interaction. *American Journal of Epidemiology*, **146**:713–719.

Yang, Q., Khoury, M. J., Sun, F., and Flanders, W. D. (1999). Case-only design to measure gene–gene interaction. *Epidemiology*, **10**:167–170.

Yelland, L. N., Salter, A. B., and Ryan, P. (2011). Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. *International Journal of Biostatistics*, **7**(1):1–31.

Yerushalmy, J. (1971). The relationship of parents' cigarette smoking to outcome of pregnancy—Implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology*, **93**:443–456.

Zhang, J. L., and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death." *Journal of Educational and Behavioral Statistics*, **28**:353–368.

Zhang, L., Mukherjee, B., Ghosh, M., Gruber, S., and Moreno, V. (2008). Accounting for error due to misclassifjcation of exposures in case–control studies of gene–environment interaction. *Statistics in Medicine*, **27**:2756–2783.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*. **108**:527–539.

Zheng, W., and van der Laan, M. J. (2012). Targeted maximum likelihood estimation of natural direct effects. *The International Journal of Biostatistics*, **8**(1):1–40.

Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology*, **168**:212–224.

INDEX

"f." indicates material in figures. "n." indicates material in footnotes. "t." indicates material in tables.