

**International Series in
Operations Research & Management Science**

Louis Anthony Cox Jr.
Douglas A. Popken · Richard X. Sun

Causal Analytics for Applied Risk Analysis



 Springer

International Series in Operations Research & Management Science

Volume 270

Series Editor

Camille C. Price
Stephen F. Austin State University, TX, USA

Associate Editor

Joe Zhu
Worcester Polytechnic Institute, MA, USA

Founding Editor

Frederick S. Hillier
Stanford University, CA, USA

More information about this series at <http://www.springer.com/series/6161>

Louis Anthony Cox Jr. • Douglas A. Popken
Richard X. Sun

Causal Analytics for Applied Risk Analysis



Springer

Louis Anthony Cox Jr.
Cox Associates
Denver, CO, USA

Douglas A. Popken
Cox Associates
Littleton, CO, USA

Richard X. Sun
Cox Associates
East Brunswick, NJ, USA

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-3-319-78240-9 ISBN 978-3-319-78242-3 (eBook)
<https://doi.org/10.1007/978-3-319-78242-3>

Library of Congress Control Number: 2018937974

© Springer International Publishing AG, part of Springer Nature 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To
Christine and Emeline*

Preface

Individual, group, organizational, and public policy decisions are often disconcertingly ineffective. They produce unintended and unwanted consequences, or fail to produce intended ones, even after large expenditures of hope, time, and resources. The difficulty of achieving unambiguous successes, in which costly actions or policies produce large, clear net benefits that even those who initially doubted find compelling after the fact, has been noted in areas as varied as personal financial decisions, corporate business decisions, engineering infrastructure decisions, non-profit initiatives for poverty disruption or delinquency prevention, public health efforts to curb the emergence of antibiotic-resistant superbugs, and regulation of pollutants to improve public and occupational health.

This book is about how to make more effective decisions—that is, decisions that are more likely to cause preferred outcomes and to avoid undesirable ones—by understanding and fixing what so often goes wrong. We believe that the most common reason for disappointing results from well-intended policies and actions is inadequate understanding of the causal relationships between actions and probabilities of outcomes. Actions guided by traditional statistical analyses of association patterns in observational data, such as regression modeling or epidemiological estimates of relative risk ratios, usually cannot be relied on to achieve their objectives because these traditional methods of analysis are usually not adequate for determining how changing some variables will change others. But that is what decision makers must know to make well-informed choices about what changes to implement. This book is therefore devoted to causal analytics methods that can provide answers to the crucial causal question of how changing decision variables—the things that a decision maker or policy maker can control or choose—changes probabilities of various outcomes. It presents and illustrates models, algorithms, principles, and software for deriving causal models from data and for using them to optimize decisions, evaluate effects of policies or interventions, make probabilistic predictions of the values of as-yet unobserved quantities from available data, and identify the most likely explanations for observed outcomes, including surprises and anomalies.

The first two chapters survey modern analytics methods, focusing mainly on techniques useful for decision, risk, and policy analysis. They emphasize how causal models are used throughout the rest of risk analytics in detecting and describing meaningful and useful patterns in data; predicting outcome probabilities if different courses of action are followed; identifying and prescribing a best course of action for making preferred outcomes more probable; evaluating the effects of current or past policies and interventions; and learning from experience, either individually or collaboratively, how to make choices that increase the probabilities of preferred outcomes. Chapter 2 also introduces the Causal Analytics Toolkit (CAT), a free in-browser set of analytics software tools available at <http://cox-associates.com/CloudCAT>, to allow readers to perform the analyses described or to apply modern analytics methods to their own data sets. Chapters 3 through 11 illustrate the application of causal analytics and risk analytics to practical risk analysis challenges, mainly related to public and occupational health risks from pathogens in food or from pollutants in air. Chapters 12 through 15 turn to broader questions of how to improve risk management decision-making by individuals, groups, organizations, institutions, and multi-generation societies with different cultures and norms for cooperation. They examine organizational learning, social risk management, and intergenerational collaboration and justice in managing risks and hazards.

Throughout the book, our main focus is on introducing and illustrating practical methods of causal modeling and analytics that practitioners can apply to improve understanding of how choices affect probabilities of consequences and, based on this understanding, to recommend choices that are more likely to accomplish their intended objectives. We believe that the analytics and big data revolutions now underway will become much more valuable as methods and software for causal analytics become more widely used to better understand how actions and policies affect outcomes.

Denver, CO, USA
Littleton, CO, USA
East Brunswick, NJ, USA

Louis Anthony Cox Jr.
Douglas A. Popken
Richard X. Sun

Acknowledgments

This book has grown out of efforts over the past decade to understand and explain how to use data and algorithms to determine as accurately, objectively, and reproducibly as possible the effects caused by changes in decisions, actions, or policies. This quest has been inspired, encouraged, and supported by many people and organizations. It is a pleasure to thank them.

Chapters 1 and 2 are based largely on a pedagogical approach developed to teach nonspecialists about information-based causal analytics methods quickly, as part of professional development and academic graduate courses taught by Tony Cox in 2017. These included a course on Decision Analysis at the University of Colorado at Denver and professional courses at the annual meetings of the Society for Benefit Cost Analysis (SBCA), the Society for Epidemiological Research (SER), and the American Industrial Hygiene Association (AIHce). Bruce Copley, Dennis Devlin, Dale Drysdale, Susan Dudley, Gary Kochenberger, and Deborah Kellogg believed in and enthusiastically supported development of a teaching approach and course materials that sought to make key ideas and methods of modern causal analytics accessible to a wider audience. We thank them.

The approach taken in those courses and in Chap. 2 of this book emphasizes the concepts and principles behind current causal analytics algorithms using a minimum of specialist jargon and mathematical notation, and then makes algorithms themselves readily available through software that can be used without learning the underlying R or Python languages and packages. Course materials are available at these links:

- <https://www.aiha.org/events/AIHce2017/Documents/PDC%20Handouts%202017/PDC%20604%20Handout.pdf>
- <http://cox-associates.com/CausalAnalytics/>

The Causal Analytics Toolkit (CAT) software and an explanation of its goals are available at these links:

- <http://cox-associates.com/CloudCAT>

- <https://regulatorystudies.columbian.gwu.edu/causal-analytics-toolkit-cat-assessing-potential-causal-relations-data>

Initial funding for development of CAT was provided by the George Washington University Regulatory Studies Center. Subsequent development of its Predictive Analytics Toolkit (PAT) module, discussed in Chap. 2, and a port from an Excel add-in version to a cloud-based version were supported in part by the American Chemistry Council. We thank Susan Dudley of the GWU Regulatory Studies Center and Rick Becker of the American Chemistry Council for their support and vision in making free, high-quality analytics software available to interested users via CAT.

The applications, ideas, and principles in Chaps. 3–15 are based mainly on recent journal articles. Material from the following articles has been used with the kind permission of Wiley-Blackwell, the publishers of *Risk Analysis: An International Journal*.

- Cox LA Jr, Popken DA. [Quantitative assessment of human MRSA risks from swine](#). *Risk Analysis*. 2014 Sep;34(9):1639–50 (Chap. 6)
- Cox LA Jr. [Overcoming learning-aversion in evaluating and managing uncertain risks](#). *Risk Analysis*. 2015 Oct; 35(10) (Chap. 12). (Thanks to Jim Hammitt and Lisa Robinson for a fascinating workshop at the Harvard Center for Risk Analysis that stimulated this work.)
- Paté-Cornell E, Cox LA Jr. [Improving risk management: from lame excuses to principled practice](#). *Risk Analysis*. 2014 Jul;34(7):1228–39. (Chap. 13)

Material from the following articles has been used with the kind permission of their publishers:

- Popken DA, Cox LA Jr. [Quantifying human health risks caused by Toxoplasmosis from open system production of swine](#). *Human and Ecological Risk Assessment*. 2015 Oct 3; 21(7): 1717–1735. (Chap. 7)
- Cox LA Jr., Popken DA, Kaplan AM, Plunkett LM, Becker RA. [How well can in vitro data predict in vivo effects of chemicals? Rodent carcinogenicity as a case study](#). *Regulatory Toxicology and Pharmacology*. 2016 Jun;77:54–64. (Chap. 8)
- Cox, LA Jr, Popken DA. [Has reducing PM2.5 and ozone caused reduced mortality rates in the United States?](#) *Annals of Epidemiology*. 2015 Mar; 25 (3):162–73. (Chap. 10)
- Cox LA Jr. [How accurately and consistently do laboratories measure workplace concentrations of respirable crystalline silica?](#) *Regul Toxicol Pharmacol*. 2016 Nov;81:268–274. (Chap. 11)
- Cox T. Uncertain causation, regulation, and the courts. *Supreme Court Economic Review*. (In press.) (Chap. 14)
- Cox LA Jr., Cox ED. (2016) [Intergenerational Justice in Protective and Resilience Investments with Uncertain Future Preferences and Resources](#). Chapter 12 in P. Gardoni, C. Murphy, and A. Rowell (Eds). *Risk Analysis of Natural Hazards: Interdisciplinary Challenges and Integrated Solutions*. Springer. New York, New York. (Chap. 15)

We thank the publishers and coauthors of these works.

Discussions with Ron Josephson of the United States Environmental Protection Agency (EPA) in the context of reviewing research proposals on health effects of air pollution helped to inspire the idea of applying causal analysis methods to determine value of information in causal networks (Chap. 2). We thank Dennis Devlin and Bruce Copley of Exxon-Mobil and Will Ollison of the American Petroleum Institute for stimulating conversations and their unswerving commitment to discovering objective scientific truth from data to inform causally effective decision and policies. As we have developed and applied software to help automatically discover scientific truth about causality from data, we have found that this approach to pursuing more objective and reliable scientific inference is not always welcome. Advocates of expert judgment-based and modeling assumption-based approaches to causal inference in risk assessment have not always embraced the ideas that computer algorithms can now be far more accurate and objective than human experts in discovering true causal relations in data, and in identifying and rejecting false causal hypotheses; and that modeling judgments and expert interpretations of statistical patterns are not necessary or desirable for drawing valid causal inferences from data. We expect this analytics-centric perspective to continue to grow in popularity as causal discovery algorithms prove their value in a wide array of risk analysis applications. Meanwhile, we thank the visionaries who are pushing to make automated, objective, reproducible, algorithmic approaches to causal model discovery and validation a practical reality.

Finally, we thank Douglas Hubbard of Hubbard Decision Research for inviting lectures and discussions of the causal analytics framework in Chap. 2 at the American Statistical Association Symposium on Statistical Inference (*Scientific Method for the 21st Century: A World Beyond $p < 0.05$* in October of 2017) and Seth Guikema of the University of Michigan for inviting the 2017 [Wilbert Steffy Distinguished Lecture](#) on Causal Analytics for Risk Management: Making Advanced Analytics More Useful at the University of Michigan Department of Industrial Engineering and Operations Research in November of 2017. The opportunity to prepare and present these lectures and to participate in the very stimulating discussions that followed contributed to the final exposition in Chaps. 1 and 2.

Contents

Part I Concepts and Methods of Causal Analytics

1 Causal Analytics and Risk Analytics	3
Why Bother? Benefits of Causal Analytics and Risk Analytics	5
Who Should Read This Book? What Will You Learn?	
What Is Required?	6
What Topics Does This Book Cover?	7
Causality in Descriptive Analytics	9
Example: Did Customer Satisfaction Improve?	10
Example: Simpson’s Paradox	11
Example: Visualizing Air Pollution-Mortality Associations in a California Data Set	12
Example: What Just Happened? Deep Learning and Causal Descriptions	14
Example: Analytics Dashboards Display Cause-Specific Information	16
Causality in Predictive Analytics	17
Example: Predictive vs. Causal Inference—Seeing vs. Doing	18
Example: Non-identifiability in Predictive Analytics	19
Example: Anomaly Detection, Predictive Maintenance and Cause-Specific Failure Probabilities	21
Causality Models Used in Prescriptive Analytics	22
Normal-Form Decision Analysis	22
Markov Decision Processes	28
Improving MDPs: Semi-Markov Decision Processes and Discrete-Event Simulation (DES) Models	33
Performance of Prescriptive Models	33
Dynamic Optimization and Deterministic Optimal Control	35
Stochastic Optimal Control, Hidden Markov Models, and Partially Observable Markov Decision Models (POMDPs)	38
Bayesian Statistical Decision Theory	42

Simulation-Optimization	47
Causal Study Design and Analysis in Evaluation Analytics	55
Randomized Control Trials (RCTs)	55
Quasi-Experiments (QEs) and Intervention Time Series Analysis Are Widely Used to Evaluate Impacts Causes by Interventions	61
Counterfactual and Potential Outcome Framework: Guessing What Might Have Been	64
Change Point Analysis (CPA) and Sequential Detection Algorithms	66
A Causal Modeling Perspective on Evaluating Impacts of Interventions Using CPTs	71
Using Causal Models to Evaluate Total Effects vs. Direct Effects	73
Using MDPs and DES Causal Models to Evaluate Policies and Interventions	75
Causality in Learning Analytics	75
Causality in Collaborative Analytics	82
Causal Models in Game Theory: Normal and Extensive Forms, Characteristic Functions	83
Causal Models for Multi-agent Systems	85
Conclusions: Causal Modeling in Analytics	87
References	91
2 Causal Concepts, Principles, and Algorithms	97
Multiple Meanings of “Cause”	99
Probabilistic Causation and Bayesian Networks (BN)	102
Technical Background: Probability Concepts, Notation and Bayes’ Rule	102
Bayesian Network (BN) Formalism and Terminology	111
Using BN Software for Probability Predictions and Inferences	114
Practical Applications of Bayesian Networks	121
Non-causal Probabilities: Confounding and Selection Biases	122
Causal Probabilities and Causal Bayesian Networks	126
Dynamic Bayesian Networks (DBNs)	132
Causal Risk Models Equivalent to BNs	135
Fault Tree Analysis	135
Event Tree Analysis	139
Bow-Tie Diagrams for Risk Management of Complex Systems	142
Markov Chains and Hidden Markov Models	143
Probabilistic Boolean Networks	144
Time Series Forecasting Models and Predictive Causation	144
Structural Equation Models (SEMs), Structural Causation, and Path Analysis Models	149
Influence Diagrams	150

Decision Trees	153
Markov Decision Processes (MDPs) and Partially Observable MDPs (POMDPs)	153
Predictive Causality and Predictive Analytics Models	154
Classification and Regression Tree (CART) Models	154
The Random Forest Algorithm: Importance Plots and Partial Dependence Plots	161
Causal Concentration-Response Curves, Adjustment Sets, and Partial Dependence Plots for Total and Direct Effects in Causal Graphs	165
Predictive Analytics for Binary Outcomes: Classification and Pattern Recognition	178
Learning Causal BN Models from Data: Causal Discovery Algorithms	184
Comparison of Causal Discovery to Associational Causal Concepts and Methods: Updating the Bradford Hill Considerations	190
Strength of Association	192
Consistency of Association	195
Plausibility, Coherence, and Analogy of Association	197
Specificity, Temporality, Biological Gradient	200
Methods and Examples of Associational Causation: Regression and Relative Risks	201
Relative Risk and Probability of Causation in the Competing Risks Framework	209
Conclusions on Associational Causation	212
Comparison of Causal Discovery to Attributive Causal Methods	216
Example: Attributive Causation Is Not Manipulative Causation	217
Example: Nine Million Deaths per Year Worldwide Attributed to Air Pollution	218
Comparison of Causal Discovery to Counterfactual Causal Methods	219
Example: Attribution of Rainfall to Climate Change, and the Indeterminacy of Counterfactuals	222
Comparison of Causal Discovery to Structural and Mechanistic Causal Modeling	223
Example: Dynamic Causal Analysis of the Level of a Single Variable in a Compartment	224
Example: A CPT for a One-Compartment Model with Uncertain Inputs	228
Example: Causal Reasoning about Equilibria and Comparative Statics	229
Historical Milestones in Development of Computationally Useful Causal Concepts	229
Conclusions	233
References	240

Part II Descriptive Analytics in Public and Occupational Health

3 Descriptive Analytics for Public Health: Socioeconomic and Air Pollution Correlates of Adult Asthma, Heart Attack, and Stroke Risks	251
Introduction	251
Data Sources	252
Methods and Analytic Strategy	255
Results	259
Dependence of Health Effects on Age and Sex	259
Smoking Effects	261
Income Effects	263
Effects of Education and Ethnicity	264
Effects of Marital Status	268
Effects of Fine Particulate Matter (PM _{2.5}) and Ozone (O ₃) Air Pollution	268
Additional Interaction Analyses	271
Results of Logistic Regression Analysis	272
Results of Bayesian Network and Partial Correlation Analysis	273
Results of Regression Tree and Random Forest Analyses	275
Discussion	278
Study Limitations and Uncertainties	281
Conclusions	282
References	282
4 Descriptive Analytics for Occupational Health: Is Benzene Metabolism in Exposed Workers More Efficient at Very Low Concentrations?	285
Introduction	285
Background: Theories and Controversies in Benzene Dose-Response	287
Data	291
Methods	293
Results	297
Descriptive Statistics	297
Metabolites vs. Benzene Concentrations in Air	297
Inter-individual Variability and Declining DSM Ratios	301
Tying Up Some Loose Ends: Joint Frequency Distributions of Variables and Intra-individual Variability of DSM Ratios	303
Discussion and Conclusions	308
References	310

5 How Large Are Human Health Risks Caused by Antibiotics Used in Food Animals?	313
Methods of Quantitative Risk Assessment	314
Farm-to-Fork Risk Simulation Models	314
Dose-Response Models for Food-Borne Pathogens	318
Quantitative Risk Characterization for QMRA and Risk Management	320
Attribution-Based Risk Assessment and Controversies	321
Empirical Upper-Bounding	323
Case Study: Ampicillin-Resistant <i>E. faecium</i> (AREF) Bacteria	323
Summary of Results from Applying Empirical Upper-Bounding Risk Assessment to Other Antibiotic-Resistant Bacteria	328
Managing Uncertain Food Risks via Quality Principles: HACCP	329
Discussion and Conclusions	330
References	331
6 Quantitative Risk Assessment of Human Risks of Methicillin-Resistant <i>Staphylococcus aureus</i> (MRSA) from Swine Operations	333
Introduction: How Large Is the Human Health Risk from MRSA in Swine and Pork?	334
Potential Human MRSA Hazards Related to Pigs and Pork	334
Consumer Exposure to MRSA via Pork Meat Poses Little Risk	334
Direct Exposure to Pigs Can Increase Risk of Colonization with MRSA	335
Hospital Outbreaks of ST398 MRSA Are Extremely Rare	335
Community Outbreaks of ST398 Have Not Been Observed	336
ST398 MRSA Is Found in Retail Pork	336
MRSA Colonization from Food Handling Is Possible	336
Quantifying Pig-Associated MRSA Colonization Potential	337
Quantitative Estimation of Colonization Potential for Professional Food Handlers in the U.S. from Pork Meat	337
Quantitative Estimation of Colonization Potential from Consumer Food Handling	338
ST398 MRSA Colonizes Pig Farm Workers	340
Estimating the Number of U.S. Pig Farm Workers	340
Estimating the Proportion of Farms with MRSA	341
Probability Model for ST398 MRSA Colonization	342
Estimating the Annual Probability of MRSA Infection for Those Colonized	342
Probability Model for Infection Given Colonization	343
Estimating Secondary Cases	343
Hospital Cases	344
Hospital Estimation Model	344
Quantitative Risk Analysis Model	345

Results	346
Discussion and Conclusions	347
References	349
Part III Predictive and Causal Analytics	
7 Attributive Causal Modeling: Quantifying Human Health Risks Caused by Toxoplasmosis from Open System Production of Swine	355
Introduction	355
Background on Toxoplasmosis	356
Data and Methods	357
Distributions for <i>T. gondii</i> Prevalence in Pigs	358
Excess Risk Factor Distribution	362
Attribution to Pork	363
Pork Attributable Human Case Rates: Adult Hospitalizations and Death	364
Congenital Toxoplasmosis	365
QALYs Lost Assignment	366
Analysis	367
Results	368
Health Outcome Distributions	368
Excess Risk Factor and QALYs Lost Due to Reduced Confinement	369
Discussion and Conclusions	370
References	372
8 How Well Can High-Throughput Screening Tests Results Predict Whether Chemicals Cause Cancer in Mice and Rats?	375
Introduction	375
Case Study: Reassessing the Accuracy and Robustness of a Rodent Carcinogenicity Prediction System	379
Purpose, Scope, and Interpretation of the Original Study	379
Original Data, and Replication Process and Results	380
Original Methods, and Replication Process and Results	383
Discussion and Conclusions	393
References	395
9 Mechanistic Causality: Biological Mechanisms of Dose-Response Thresholds for Inflammation-Mediated Diseases Caused by Asbestos Fibers and Mineral Particles	397
Introduction	397
Biological Background: NLRP3 Inflammasome Responses to Mineral Particles	399

Thresholds in NLRP3 Priming: Receptor-Mediated Signal Transduction and Critical Mass of NLRP3 Protein Required for Activation	401
Thresholds for NLRP3 Assembly: Cooperativity in Oligomerization Kinetics	402
Thresholds in NLRP3 Activation: Lysosome Disruption and ROS	404
Thresholds in NLRP3 Signaling: Positive Feedback Loops and Bistability	404
From Cells to Tissues: Percolation Thresholds for Spread of Inflammation	408
Discussion and Conclusions	410
References	412
Part IV Evaluation Analytics	
10 Evaluation Analytics for Public Health: Has Reducing Air Pollution Reduced Death Rates in the United States?	417
Introduction: Using Data from Natural Experiments to Understand Causality	417
Data and Methods	423
Statistical Analysis Methods	424
Results	427
Descriptive Analytics	427
Results on Statistical Associations Between Pollutant Levels and Mortality Rates	430
Results on Correlations Between Changes in Variables over Time	432
Granger Causality Test and Control Results	434
Discussion and Conclusions: Caveats for Causal Interpretations of Regression Coefficients	436
Study Limitations	437
Comparisons to Conclusions from Other Studies	438
References	440
11 Evaluation Analytics for Occupational Health: How Well Do Laboratories Assess Workplace Concentrations of Respirable Crystalline Silica?	443
Introduction	443
Data and Methods	445
Results	449
Discussion	451
Conclusions	452
References	454

Part V Risk Management: Insights from Prescriptive, Learning, and Collaborative Analytics

12 Improving Individual, Group, and Organizational Decisions: Overcoming Learning Aversion in Evaluating and Managing Uncertain Risks	457
Introduction	458
Benefit-Cost Analysis	458
Example: A Simple BCA Justification for Banning Coal Burning	459
Example (Cont.): A BCA Justification for Banning Coal Burning May Be Regrettable	460
Aspirations and Benefits of BCA	463
Example: Majority Rule Without BCA Can Yield Predictably Regrettable Collective Choices	464
Decision-Making by <i>Homo economicus</i>	466
Example: Pareto-Inefficiency of BCA with Disagreements About Probabilities	467
Example: Impossibility of Pareto-Efficient Choices with Sequential Selection	469
How Real People Evaluate and Choose Among Alternatives	470
The Affect Heuristic Effects Risky Choice and BCA Evaluations via a Network of Decision Biases	471
Decision Biases Invalidate Straight-Forward Use of WTP Values	473
Example: Non-existence of WTP in a Social Context	475
Multiple Decision Biases Contribute to Learning Aversion	476
Example: Information Externalities and Learning Aversion in Clinical Trials	478
Example: Desirable Interventions with Uncertain Benefits Become Undesirable When They Are Scaled Up	478
Learning Aversion and Other Decision Biases Inflate WTP for Uncertain Benefits	479
Example: Overconfident Estimation of Health Benefits from Clean Air Regulations	481
Doing Better: Using Predictable Rational Regret to Improve BCA	482
Example: Rational vs. Irrational Regret	483
Conclusions	487
References	488
13 Improving Risk Management: From Lame Excuses to Principled Practice	493
Introduction	493
Why Do Catastrophes Happen? Bad Luck Is Rarely the Whole Answer	494
“It’s Not Our Fault”: Some Common Excuses for Bad Risk Management	495

Foundations for Better Risk Management	500
Understand the Causes of the Hazard, Then Its Potential Effects	500
Characterize Risk Magnitudes and Uncertainties	501
Identify Possible Risk Reduction Measures and Candidly Assess Their Costs, Benefits, and Uncertainties	502
Assess the Urgency of Immediate Action and the Value of Information	503
Anticipate, Monitor, and Prepare for Rare and Not-So-Rare Events	504
Deliberately Test and Learn	505
Learn from Near-Misses and Identify Accident Precursors	505
Establish and Maintain a Culture of Safety	505
Put the Right People in the Right Place with the Right Knowledge, Incentives and Resources	506
Clearly Define Leadership and Responsibilities	507
Share Knowledge and Experience Across Organizations	508
Conclusions	509
References	510
14 Improving Institutions of Risk Management: Uncertain Causality and Judicial Review of Regulations	513
Introduction: Principles of Law-and-Economics and Benefit-Cost Analysis of Regulations	514
Example: The Learned Hand Formula for Liability Due to Negligence	515
Example: The Cheapest Cost-Avoider Principle When Liability Is Uncertain	516
Uncertain Causation Encourages Ineffective and Potentially Harmful Regulations	518
Uncertain Causation Encourages Socially Reckless Regulation	518
Warnings from Behavioral Economics and Decision and Risk Psychology: The Tyranny of Misperceptions	519
Example: The Irish Coal-Burning Bans	521
Example: Estimated Benefits of Fine Particulate Matter (PM2.5) Regulation in the United States	523
Example: Food Safety Regulation Based on Assumed Causation	525
Lessons from the Examples	527
Better Causal Inferences and Benefits Estimates via More Active Judicial Review	529
Distinguishing Among Different Types of Causation	531
Example: Associations Do Not Necessarily Provide Valid Manipulative Causal Predictions	535
Can Regulatory Benefits Estimation Be Improved, and, If So, How?	536

Causation as Judgment: The Hill Considerations for Causality and Some Alternatives	539
Causation as Discoverable Empirical Fact: Causal Inference Algorithms and Competitions	542
Synthesis: Modernizing the Hill Considerations	545
Summary and Conclusions: Potential Roles of Judicial Review in Transforming Regulatory Causal Inference and Prediction	549
References	553
15 Intergenerational Justice in Protective and Resilience Investments with Uncertain Future Preferences and Resources	557
Introduction: How Much Care Does Each Generation Owe to Future Ones?	558
Simple Models of Intergenerational Justice: Sharing a Pie Over Time	561
Analytic Frameworks for Managing Cooperation Across Generations	562
Economic Growth Models	562
Behavioral Game Theory Framework for Strategic Interactions Among Generations	565
Axiomatic Solution Concepts from Cooperative Game Theory	567
Intergenerational Justice	568
Sustainability, Protective Principles, and Fundamental Trade-Offs	570
Investing in Building Resilient Communities and Societies: An Emerging Framework	571
Ethical Frameworks	572
Discussion: Principles for Applying the Frameworks to Improve Decisions and Policies	573
Conclusions	577
Epilog: A Vision for Causal Analytics in Risk Analysis	578
References	580
Index	583

Part I

Concepts and Methods of Causal Analytics

Chapter 1

Causal Analytics and Risk Analytics



Countless books and articles on data science and analytics discuss descriptive analytics, predictive analytics, and prescriptive analytics. An additional analytics area that is much less discussed links this world of analytics, with its statistical model-based descriptions and predictions, to the world of practical decisions in which actions have consequences that decision-makers, and perhaps other stakeholders, care about, and about which they are often uncertain. This is the area of *causal analytics*. How causal analytics relates to other analytics areas and how its methods can be used to predict what to expect next, explain past outcomes and observations, prescribe what to do next to improve future outcomes, and evaluate how well past or current policies accomplish their intended goals—for whom, and under what conditions—are the main topics of this book.

Causal analytics uses data, models, and algorithms to estimate how different actions change the probabilities of different possible future outcomes. By doing so, it provides the crucial information needed to solve both the prescriptive decision analysis problem of choosing actions to make desired outcomes more likely and undesired ones less likely, and also the evaluation challenge of determining what effects past actions and events have had. Useful recommendations from prescriptive analytics flow from understanding how actions affect outcome probabilities. Learning from experience the causal relationships between actions or policies and their consequences also makes it possible to evaluate and improve policies over time. Collectively, these activities of using data to quantify the causal relationship between actions and their outcome probabilities and then using this understanding to evaluate and improve decisions and policies contribute to what we shall call *risk analytics*: the application of algorithms to data to produce results that inform and improve risk management.

This book is largely about how to apply principles and methods of causal analytics to data to solve practical risk management decision problems and to inform and improve other steps in the risk analytics process. Table 1.1 outlines these steps, and they are discussed more fully in the rest of this chapter. Chapter 2 introduces

Table 1.1 Components of risk analytics

Risk Analytics Step	Typical Questions Addressed
Descriptive analytics	<ul style="list-style-type: none"> • What is the current situation? What's happening? • What has changed? What's new? • What should we focus on? What should we worry about?
Predictive analytics	<ul style="list-style-type: none"> • If we do not change what we are doing, what will (probably) happen next? When? How likely are the different possibilities? • Given observed (or assumed) values for some variables, what are the probabilities for values of other variables? How well can some variables be predicted from others? • How well can future outcomes be predicted now?
Causal analytics	<ul style="list-style-type: none"> • Diagnosis, explanation, and attribution: What explains the current situation? • What can we do about it? How would different actions change the probabilities of different future outcomes?
Prescriptive analytics	<ul style="list-style-type: none"> • What should we do next? What decisions and policies implemented now will most improve probabilities of future outcomes?
Evaluation analytics	<ul style="list-style-type: none"> • How well are our current decisions and policies working? • What effects have our decisions and policies actually caused? • How do different policies affect behaviors and outcomes for different people?
Learning analytics	<ul style="list-style-type: none"> • What decisions or policies might work better than our current ones? • How can we use data and experimentation to find out? • By how much do different items of information improve decisions? What is the value of information for different measurements?
Collaborative analytics	<ul style="list-style-type: none"> • How can we best work together to improve probabilities of future outcomes? • Who should share what information with whom, how and when? • What actions should each division of an organization or each member of a team take?

different concepts of causality and describes how they can be used to achieve such practical goals as quickly noticing important changes in a controlled industrial system or in an organization's performance or environment; explaining and predicting such changes and how they affect the performance of the system or organization; making more accurate predictions from limited data; and devising more effective interventions and policies to promote desired outcomes. Causal analytics provides algorithms for learning from data what works and what does not and for estimating how well different policies, treatments, or interventions perform in changing behaviors or outcomes for different people. These methods play a central role in the rest of risk analytics by providing information needed to address the questions in the right column of Table 1.1. Chapter 2 also discusses theoretical principles and existing algorithms and software for building causal models from data and knowledge and for using them to support the rest of the analytics steps of description, prediction, prescription, evaluation, learning, and collaboration in understanding and managing risks in uncertain systems.

The rest of the book illustrates practical applications of these methods, grouped roughly around the analytics steps in Table 1.1. The different chapters apply and extend principles, ideas and methods explained in this chapter and Chap. 2 to a variety of practical risk analysis problems and challenges. They emphasize public health risks, occupational health and safety, and possibilities for improving individual, organizational, and public policy decisions.

Why Bother? Benefits of Causal Analytics and Risk Analytics

Several large potential practical benefits from applying causal analytics provide ample motivation for mastering the technical methods needed to distinguish between association and causation and to estimate causal relationships among variables. The most important one is the ability to quantify how changes in the inputs to a system or situation change the probabilities of different outputs or results, which we refer to generically as *outcome probabilities*. This ability, in turn, allows the decision optimization question to be addressed of what to do to make preferred outcomes more likely. Crucially, causal analytics also provides relatively objective, data-driven methods for evaluating quantitatively how large the effects of policies, interventions, or actions on improving outcomes actually are. They enable quantitative assessment of what works, how well, for whom, and under what conditions.

In marketing science, medical research, and social science program evaluation studies, randomized control trials (RCTs), which randomly assign individuals to different “treatments” (actions, policies, or interventions), are often considered the gold standard for evaluating causal impacts of treatments on outcomes. Random assignment of individuals to treatments guarantees that any systematic differences in the responses to different treatments are not due to systematic differences between the individuals receiving different treatments. Causal analytics methods allow many of the benefits of RCTs to be achieved even when data are not obtained from RCTs. For example, they can be applied to data from observational studies in which there are no interventions or treatments. They can be applied to data from natural experiments or “quasi-experiments” in which random assignment has not been used. They also address several limitations of RCTs. While RCT results often do not generalize well beyond the specific populations studied, causal analytics methods provide constructive “transport formulas” for generalizing results inferred from one population to others with different population frequency distributions of risk factors; or, more generally, for applying causal relationships and laws discovered in one or more data sets to new data sets and situations (Bareinboim and Pearl 2013; Lee and Honavar 2013). Causal analytics can also help to explain why treatments work or fail, and for whom, rather than simply quantifying average treatment effects in a population. They provide powerful techniques for predicting how changes in causal drivers will change future outcomes (or their probabilities) and for deducing the values of unobserved variables that best explain observed data.

In all these ways, the methods we will be studying in this book improve ability to understand, explain, predict, and control the outputs of uncertain systems and situations by clarifying how different decisions about inputs affect probabilities of outcomes. However, to achieve these potential benefits, appropriate methods of causal analytics must be used correctly. This chapter and the next explain these methods; they also caution against unsound, incorrect, and assumption-laden methods of causal analysis. Two particularly important confusions to beware of are (a) Confusion between the effects *attributable* to a cause, such as numbers of illnesses or deaths attributed to a risk factor, and the effects *preventable* by removing or reducing that cause; and (b) Confusion between past effect levels that *would have been* observed had a risk factor been absent or smaller, as estimated by some model, and future effect levels that *will be* observed if the risk factor is removed or reduced. Chapter 2 discusses these and other important distinctions. They are often not drawn with precision in current policy analyses, risk assessments, and benefit-cost analyses. As a result, policy makers are too often presented with information that is claimed to show what should be expected to happen if different policies are implemented, when what is actually being shown is something quite different, such as historical associations between variables. A greater understanding and use of causal analytics can help to overcome such confusions. In turn, greater clarity about the correct causal interpretation of presented information can help to better inform policy makers about the outcome probabilities for different courses of action.

Who Should Read This Book? What Will You Learn? What Is Required?

This book is meant primarily for practitioners who want to apply methods of causal analytics correctly to achieve the benefits just discussed. Practitioners need to understand the inputs and outputs for different analytics algorithms, be able to interpret their outputs correctly, know how to apply software packages to data to produce results, and be aware of the strengths, limitations, and assumptions of the software packages and results. The details of how the algorithms compute outputs from inputs are less important for practitioners. We will therefore make extensive use of state-of-the-art analytics algorithms and highlight their key principles and while referring to the primary research literature and to software package documentation for technical and implementation details. This chapter and Chap. 2 seek to present the main ideas of key causal analytics and risk analytics, making them accessible for a broad audience of technically inclined policy-makers, analysts, and researchers who are not expected to be specialists in data science and analytics. The key ideas are independent of the software used to implement them, so we will explain the ideas and illustrate inputs and outputs without assuming or requiring familiarity any particular pieces of software. Thus, one audience for this book is fairly broad: we hope to make the main technical ideas of modern risk analytics and

causal analytics (Chaps. 1 and 2) and their practical applications to a variety of practical risk assessment and risk management problems (remaining Chapters) clear, interesting, and useful to those who must decide what to do, evaluate what has been done, or offer advice about what should be done next in order to cause desired results.

We anticipate that a subset of readers may want to personally master the algorithms discussed and start applying them to their own data. For those readers, Chap. 2 introduces several free software packages that can be used to carry out the calculations and analyses shown throughout the book. Many state-of-the-art algorithms for causal analytics and closely related machine-learning tasks are implemented as freely available packages in the R statistical programming environment. For Python developers, the scikit-learn machine learning package and other analytics packages provide valuable alternatives to some R packages. We mention such available software packages throughout the book where appropriate, but assume that many readers have limited interest in learning about their details here. Instead, we have created a cloud-based (in-browser) Causal Analytics Toolkit (CAT), introduced in Chap. 2, to let interested readers run algorithms on example data sets bundled with CAT, or on their own data sets in Excel, without having to know R or Python. The CAT software make it possible to perform the analyses we describe simply by selecting columns data table and then clicking on the name of the analysis to be performed; outputs are then produced.

Thus, we expect that most readers with a technical bent and interest in decision, risk, and policy analysis—or in causal analysis, either for its own sake or for other applications—will be able to use this book to understand, relatively quickly and easily, the main technical ideas at the forefront of current causal and risk analytics and how they can be applied in practice. Readers who want to do so can also this book, especially Chap. 2, to master technical skills including applying current state-of-the-art R packages (via the simplified Causal Analytics Toolkit (CAT) software) to quantify and visualize associations in data, analyze associations using parametric and non-parametric regression methods, detect and quantify potential causal relations in data, visualize them using causal networks, and quantify various types of important causal relations in real-world data sets. Finally, readers who care mainly about risk analysis applications can skip the rest of this chapter and Chap. 2 and proceed directly to the applications that begin in Chap. 3. We have sought to make the applied chapters relatively self-contained, recapitulating key ideas from Chaps. 1 and 2 where needed.

What Topics Does This Book Cover?

This chapter and the next describe the roles of causal analytics in the rest of risk analytics and provide technical background on current technical concepts and methods of causal analytics, respectively. The rest of this chapter walks through the risk analytics steps in Table 1.1: descriptive, predictive, prescriptive, evaluation,

learning, and collaborative analytics. It discusses how causal analytics is woven into each of them. A major goal is to show how causal analysis and modeling can clarify and inform the rest of the analytics process. Conversely, implicit and informal causal assumptions and causal interpretations of data can mislead analysts and users of analytics results. We provide caveats and examples of how this can occur and suggestions for avoiding pitfalls in data aggregation, statistical analysis, and causal interpretation of results. This chapter also surveys some of the most useful and exciting advances in each of area of risk analytics.

Chapter 2 delves into different technical concepts of causation and methods of causal analysis. These range from popular but problematic measures of associative, attributive, and counterfactual causation, which are based primarily on statistical associations and modeling assumptions, to more useful definitions and methods for assessing predictive, manipulative, structural, and mechanistic or explanatory causation. Chapter 2 emphasizes non-parametric methods and models, especially Bayesian networks (BNs) and other causal directed acyclic graph (DAG) models, and discusses the conditions under which these have valid causal, structural, and manipulative causal interpretations. It points out that many models used for inference and decision optimization in control engineering, decision analysis, risk analysis, and operations research can be brought into the framework of causal BN modeling and briefly discusses other models, such as continuous and discrete-event simulation models, that provide more detailed descriptions of causal processes than BNs.

The remainder of the book is devoted to applications that illustrate how causal analytics and risk analytics principles and methods can be applied in risk analyses, with the main applied focus being on human health risks. Chapters 3 through 6 illustrate how descriptive analytics can be used to address questions in public and occupational health risk analysis, such as: What factors are associated with risks of adult asthma, heart attacks, and strokes, and to what extent might such associations be causal? Do workers exposed to very low occupational concentrations of benzene have disproportionately high levels of risk due to relatively efficient production of metabolites at low exposure concentrations? How large are the risks to humans of antibiotic-resistant “super-bug” infections caused by use of antibiotics in farm animals? Chapters 7 and 8 illustrate attributive and predictive causal analytics by estimating human health risks caused by Toxoplasmosis from open system production of swine and by evaluating how well rodent carcinogenicity in multimillion dollar *in-vivo* 2-year assays can be predicted from much less expensive high throughput screening (HTS) data (the answer is somewhat disappointing), respectively. Chapter 9 is the sole chapter devoted entirely to mechanistic and explanatory causation. It examines health risks at the micro level of disease causation in individuals, focusing on one of the hottest topics in disease biology today: the role of the NLRP3 inflammasome in inflammation-mediated diseases. Mechanistic models of disease causation typically require more detailed applied mathematical and computational modeling than the directed acyclic graph (DAG) models emphasized in earlier chapters. Chapter 9 describes the types of mathematical models that might be useful in quantitative causal modeling of exposure-response relationships.

in which exposures activate the NLRP3 inflammasome and thereby cause increase risks of inflammation-mediated diseases such as lung cancer, mesothelioma, or heart attack.

Chapters 10 and 11 undertake retrospective evaluations of the results actually caused or achieved by two programs: effects on human mortality rates caused by historical reductions in air pollution in many different areas of the United States; and accuracy of efforts to use sampling to estimate workplace air concentrations of respirable crystalline silica (quartz sand and dust). The main purpose of these chapters is to demonstrate the value of retrospective evaluation of what works how well, under what conditions. They also show how to carry out such evaluations using a range of technical methods, from simple descriptive statistics to Granger causality tests. Chapters 12 through 15 shift focus from specific health risk analysis applications to broader questions of prescriptive analytics and individual, group, organizational, institutional, and societal learning and risk management. These chapters are more general and more theoretical than Chaps. 3–11, as they deal with topics such as improving individual and group decisions; better incorporating risk management disciplines into organizations; using judicial review, especially of causal reasoning, to improve the quality of regulatory science; and pursuing efficiency and justice in risk management decisions that span multiple generations.

To enable the most useful discussion of these topics, it is essential to have a solid grasp of the main aspects of causal analytics and risk analytics. To these, we now turn.

Causality in Descriptive Analytics

Descriptive analytics seeks to summarize, organize, and display data to answer questions about what has happened or is happening and to highlight features of the data that are relevant, interesting, and informative for deciding where to focus attention and what to do next. Questions addressed by descriptive analytics, such as “What has happened to the local unemployment rate recently?” or “How many people per year in this country are currently dying from food poisoning?” or “How has customer satisfaction changed since a year ago for this company?” often involve underlying causal assumptions. For example, an unemployed person may be defined as one who is actively seeking a job and able and willing to work but who does not have a job at the moment *because* the ongoing search has not yet succeeded. The word “because” indicates a causal attribution. Deciding whether someone should be counted as unemployed depends on the reason or cause for currently not having a job. An otherwise identical person lacking a job for a different stated reason—such as because he or she has given up looking after exhausting all promising leads in a market where employers are not hiring—would not be counted as “unemployed” by this definition, even if he or she is equally eager to find work and capable of working. Thus, describing how many people are currently unemployed requires assessing causes.

Similarly, suppose that food poisoning deaths occur disproportionately among people with severely compromised immune systems who would have died on the same days that they did, or perhaps very shortly thereafter, even in the absence of food poisoning. Then describing these deaths as being *due to* food poisoning might create an impressive-looking death toll for food poisoning, even if preventing food poisoning would not significantly reduce the death toll. Here, the phrase “due to” indicates a causal attribution, but the descriptive statistics on mortalities attributed to food poisoning do not necessarily reveal how changing food poisoning would change mortality rates. In the terminology introduced in Chap. 2, food poisoning in this example would be an attributive cause but not a manipulative cause of deaths. Descriptive analytics results that are not explicit about the kinds of causation being described can mislead their recipients about the probable effects of interventions.

How data are aggregated and summarized to answer even basic descriptive questions can reflect underlying causes more or less well, as illustrated by the following examples.

Example: Did Customer Satisfaction Improve?

Consider how to use data to describe how customer satisfaction has changed since a new customer relationship management (CRM) program was inaugurated a year ago. This question might arise for a retail chain trying to decide whether to roll out the same CRM program in other locations. For simplicity, suppose that each surveyed customer is classified as either “satisfied” (e.g., giving a satisfaction rating of 7 or more on a scale from 0 to 10) or “not satisfied” (giving a lower rating). Suppose also that the data are as follows:

- A year ago, 7000 of 10,000 randomly sampled customers were satisfied.
- Today, only 6600 of 10,000 randomly sampled customers were satisfied.

Based on this description of the data, it might seem plain that there is no evidence that the new CRM program has proved effective in increasing customer satisfaction. In fact, there has been a statistically significant decline in the proportion of satisfied customers (with a maximum likelihood estimate of $0.70 - 0.66 = 4\%$ and a 95% confidence interval of 2.7–5.3% for the decline, calculated via the on-line calculator for a difference of two proportions at http://vassarstats.net/prop2_ind.html).

The following different, more detailed, presentation of the same data tells a different story:

- A year ago, 55% of women customers and 85% of men customers were satisfied. There were equal numbers of both, so the aggregate fraction of satisfied customers was $0.5 \times 55\% + 0.5 \times 85\% = 70\%$.
- Today, 60% of women customers and 90% of men customers are satisfied. Moreover, after years of no growth, the number of women customers has quadrupled in the past year and the number of men customers has remained the

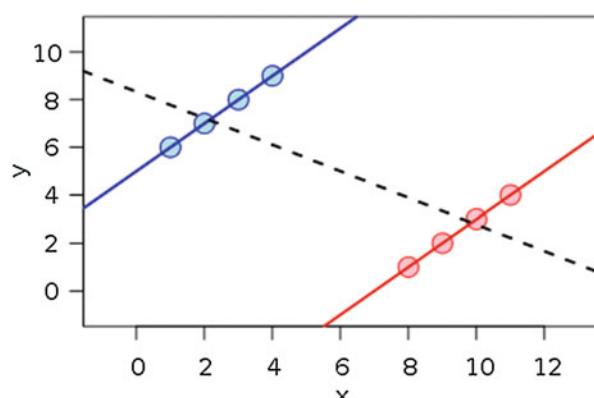
same. Thus, $4/5$ of the customers are now women. Hence, current aggregate satisfaction is $(4/5)*60\% + (1/5)*90\% = 66\%$.

This second presentation of the data reveals that the new CRM program was followed by an increase of 5% in the fraction of satisfied customers for both men and women and that the business expanded dramatically by attracting more female customers. If these changes were caused by the CRM program, then it might be well worth considering for broader adoption. The first presentation of the data made the program seem unsuccessful because it only reveals the decrease in overall percentage of satisfied customers without showing that this is caused by an influx of new female customers and that satisfaction increased within each sex-specific stratum of customers. The second presentation of the data provides more useful information to inform planners about the probable consequences of introducing the program. This example illustrates that how data should be aggregated and presented for use by decision-makers depends on what cause-and-effect relations are to be illuminated by the analysis. We will return later to how to use understanding of cause and effect to design descriptions of data that reveal causal impacts (e.g., that the CRM program increases the fraction of satisfied customers by 5% in each stratum) instead of obscuring them. One simple and important insight is that if an outcome such as customer satisfaction depends on multiple factors, such as gender and exposure to a CRM program, then estimating the effects of one factor may require adjusting for the effects of others.

Example: Simpson's Paradox

Figure 1.1 shows how adept use of data visualization can clarify otherwise puzzling or misleading patterns in a data set. It shows that the overall statistical association between two variables, x and y , such as annual advertising spent per customer and annual number of purchases per customer, might be negative (downward-sloping)

Fig. 1.1 Simpson's Paradox for quantitative data. A positive trend holds within each of two separate groups or clusters, although a negative trend appears when the data from the groups are combined.
Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox



dashed line) even if spending more on advertising per customer increases the average number of purchases per year for each individual (upward-sloping solid lines through the data points). In this display, it is clear that there are two clusters of individuals (e.g., men and women), and that the association between x and y is positive within each cluster but negative overall because the cluster with the higher x values has lower y values. In other words, increasing advertising per customer on the x axis increases the expected number of purchases per individual on the y axis for individuals in both clusters, even though the overall association between x and y is negative. This is an example of *Simpson's Paradox* in statistics, and the visualization makes clear how it arises.

Studies of Simpson's Paradox have been motivated by real-world examples in which it complicates interpretation of data. For example, a study of gender bias in university admissions might find that men applying to a graduate school are significantly more likely to be admitted than women, even if each department in the school is more likely to admit women than men. This can happen if women are more likely than men to apply to the departments with the lowest admissions rates. Or, a new drug for dissolving kidney stones might have a lower overall success rate than an older drug, even if it is more effective for both large kidney stones and small kidney stones, if it is more likely to be used on large kidney stones and these have a lower success rate than small kidney stones. In these and countless other examples, interpreting the implications of an overall statistical association requires understanding what other factors affect the outcome and then controlling for them appropriately to isolate the causal effect of the factor of interest, such as the effect of advertising on customer purchases, the effect of gender on admission decisions, or the effect of a drug or treatment on diseases. Causal analytics provides methods for determining which factors need to be controlled for (e.g., sex of customers, academic department applied to, size of kidney stone), and how, to isolate specific causal effects of interest. Chapter 2 introduced software packages (such DAGitty, see Fig. 2.21) that implement these methods.

Example: Visualizing Air Pollution-Mortality Associations in a California Data Set

Table 1.2 shows the first few records of a data set that is used repeatedly later to illustrate analytics methods and principles. The full data set can be downloaded as an Excel file from <http://cox-associates.com/CausalAnalytics/>. It is the file LA_data_example.xlsx. It is also bundled with the CAT software described in Chap. 2, appearing as the data set named “LA”. The rows contain daily measurements of fine particulate matter (PM2.5) concentrations, weather-related variables, and elderly mortality counts (number of deaths among people aged 75 years or older) for California’s South Coastal Air Quality Management District (SCAQMD), which

Table 1.2 Layout of data for PM2.5 concentration (“PM2.5”), weather, and elderly mortality (“mortality75”) variables for California’s South Coastal Air Quality Management District (SCAQMD)

Year	Month	Day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

contains Los Angeles. The full data set has 1461 rows of data, one for each day from January 1 of 2007 through December 31, 2010.

The variables (columns) in Table 1.2, and their data sources, are as follows:

- The calendar variables *year*, *month*, and *day* in the first three columns identify when the data were collected. Each row of data represents one day of observations. Rows are called “cases” in statistics and “records” in database terminology; we shall use these terms interchangeably. (They are also sometimes called “instances” in machine learning and pattern recognition, but we will usually use “cases” or “records.”)
- *AllCause75* is a count variable giving the number of deaths among people aged at least 75 dying on each day, as recorded by the California Department of Health. Columns in a data table typically represent “variables” in statistics or “fields” in database terminology; we shall usually refer to them as variables.
- *PM2.5* is the daily average ambient concentration of fine particulate matter (PM2.5) in micrograms per cubic meter of air, as recorded by the California Air Resources Board (CARB) at www.arb.ca.gov/aqmis2/aqdselect.php.
- The three meteorological variables *tmin* = minimum daily temperature, *tmax* = maximum daily temperature, and *MAXRH* = maximum relative humidity, are from publicly available data from Oak Ridge National Laboratory (ORNL) and the US Environmental Protection Agency (EPA): http://cdiac.ornl.gov/ftp/usshcn_daily/ www3.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdatal.htm.

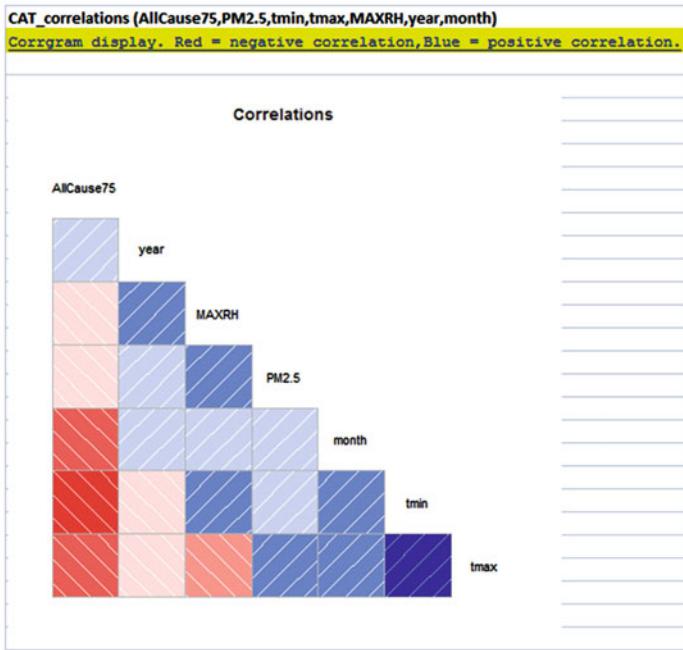
Lopiano et al. (2015) and the above data sources provide further details on these variables. For example, for the *AllCause75* variable, Lopiano et al. explain that elderly mortality counts consist of “The total number of deaths of individuals... 75+ years of age with group cause of death categorized as AllCauses... Note accidental

deaths were excluded from our analyses.” The definitions of the populations covered and the death categories used are taken from the cited sources. Clearly, average PM2.5 concentrations at monitor sites do not apply in detail to each individual, any more than the weather conditions describe each individual’s exposure to temperature and humidity. Rather, these aggregate variables provide data from which we can study whether days with lower recorded PM2.5 levels, or lower recorded minimum temperatures, relative humidity, and so forth, also have lower mortality, and, if so, whether various types of causal relationships, discussed in Chap. 2, hold between them.

Given such data, a key task for descriptive analytics is to summarize and visualize relationships among the variables to provide useful insights for improving decisions. A starting point is to examine associations among variables. Figure 1.2 presents two visualizations of the linear (Pearson’s product-moment) correlations between pairs of variables. (These, together with other tables and visualizations, were generated by clicking on the “Correlations” command in the CAT software described in Chap. 2.) The network visualization on the right displays correlations between variables using proximity of variables and thickness of links between them to indicate strength of correlation (green for positive, red for negative). In the corrgram on the left, positive correlations are indicated by boxes shaded with positively sloped hatch lines (blue in color displays). Negative correlations are indicated by boxes shaded with negatively sloped hatch lines, shaded red in color displays. Darker shading represents stronger correlations. Glancing at this visualization shows that the two strongest correlations are a strong positive correlation (dark blue) between t_{min} and t_{max} , the daily minimum and maximum temperatures; and a moderate negative correlation between t_{min} and $AllCause75$, suggesting that fewer elderly people die on warmer days. Whether and to what extent this negative correlation between daily temperature and elderly mortality might be causal remains to be explored. Chapter 2 discusses this example further.

Example: What Just Happened? Deep Learning and Causal Descriptions

A useful insight from machine learning and artificial intelligence is that raw data, e.g., from sensor data or event logs, can usually be described more abstractly and usefully in terms of descriptive categories, called *features*, derived from the data. Doing so typically improves the simplicity, brevity, and noise-robustness of the resulting descriptions: abstraction entails little loss, and often substantial gain, in descriptive power. Descriptively useful features can be derived automatically from lower-level data by data compression and information theory algorithms (of which autoencoders and deep learning algorithms are among the most popular) that map the detailed data values into a smaller number of categories with little loss of information (Kale et al. 2015). (Readers who want to apply these algorithms can do so using the



Network generated by qgraph for Pearson.Correlations
In qgraph, heavier lines and shorter distances show stronger correlations.

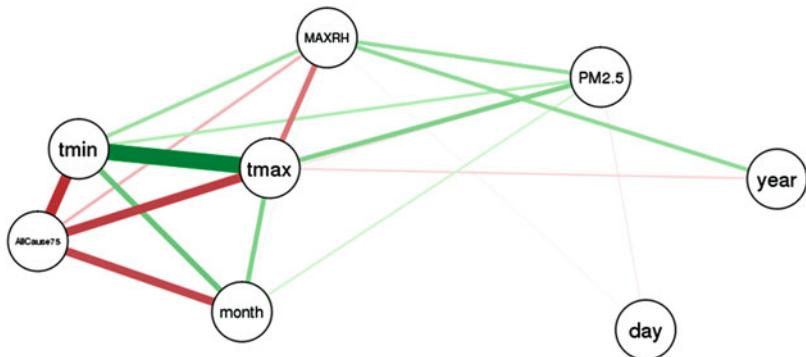


Fig. 1.2 A corrgram (left side) and network visualization (right side) displaying correlations between pairs of variables in Table 1.2

free H2O package in R or Python; see <https://github.com/h2oai/h2o-tutorials/tree/master/tutorials/deeplearning>.) Describing the data in terms of these more abstract features typically improves performance on a variety of descriptive, predictive, and prescriptive tasks. For example, describing what a nuclear power plant or an airplane is doing in terms of the stream of detailed events and measurements recorded in

sensor log data is apt to be far less useful for both descriptive and predictive purposes than describing it using higher-level abstract terms such as “The core is overheating due to loss of coolant; add coolant to the reactor vessel now to prevent a core melt” or “The plane’s angle of attack exceeds the critical angle of attack at this air speed; if this continues, a stall is imminent.” The natural language of descriptions, predictions, warnings, and recommendations implicitly reflects causal relationships, as in “Your car’s engine is overheating; add coolant to the radiator now” or “You are descending too fast; pull up!” The brevity and high information value of such communications indicate that they are using terms at the right level of abstraction to communicate effectively: they express using only a few variables and rough values or comparative terms, such as “too steep” or “too hot” or “too fast” the essential information needed to understand what is going on and what to do about it to avoid undesirable outcomes. Individual terms such as “overheating” convey a wealth of causal information. Such examples suggest a close relationship between causality, information, and effective (brief and informative) communication. Chapter 2 develops these connections further by using information-based algorithms to develop parsimonious descriptions and predictive models.

Example: Analytics Dashboards Display Cause-Specific Information

It is common for contemporary analytics dashboards to combine descriptive, predictive, and causal information by comparing observed values for cause-specific outcomes to their desired, predicted, and past values. Such visualizations make unexpected deviations and recent changes visually obvious. For example, Fig. 1.3 shows a clinical dashboard that, on its left side, displays actual rates (thick red pointers) and expected rates (thin grey pointers) of patients entering a disease register for various groups of cause-specific reasons—coronary heart disease, cancer, chronic obstructive pulmonary disease (COPD), and palliative care. The COPD rate is about double its expected value, at close to 7 patients per 1000. The right side shows stacked bar charts of the numbers of patients per 1000 entering the practice’s register for each of these groups of reasons in each of three successive time intervals. The cause-specific rates have remained fairly stable over time.

Many dashboards allow the user to drill down on high-level aggregate descriptive categories of causes or effects, such as “cancer,” to view results by more specific sub-types.

Such analytics dashboards are now widely used in business intelligence, sales and marketing, financial services, energy companies, engineering systems and operations management, telecommunications, healthcare, and many other industries and applications. They provide constructive visual answers to the key questions of descriptive analytics: “What’s happening?”, “What’s changed?”, “What’s new?” and “What should we worry about or focus on?” for different groups of causes and

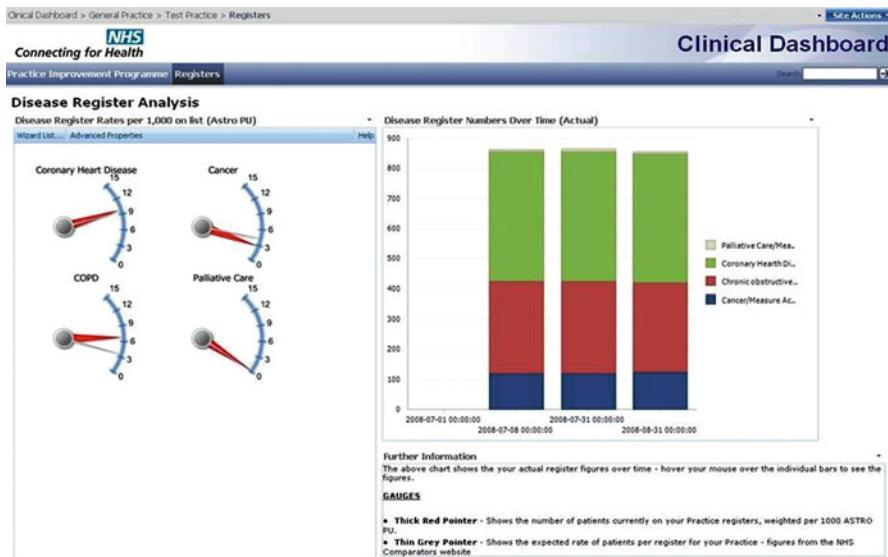


Fig. 1.3 A clinical dashboard displaying descriptive analytics results for different clusters of causes. Source: National Health Service, U.K. <https://developer.nhs.uk/library/systems/clinical-dashboards/clinical-dashboard-user-interface-design-guide/>

effects. Readers who want to build dashboards for their own data and applications can do so with commercial products such as Tableau (www.tableau.com/learn/training) or with free dashboard development environments such as the *flexdashboard* R package (<http://rmarkdown.rstudio.com/flexdashboard/>).

Causality in Predictive Analytics

Two main forms of predictive analytics are *forecasting* future values from past ones and *inferring* values of unobserved variables from values of observed ones. When time series data are arrayed as in Table 1.2, with time increasing down the rows, then forecasting consists of using data values in earlier rows to predict values in later ones. Predictive inference consists of using the data in some columns to predict the values in other columns. The two can be combined: we might use the past several days (rows) of data on air pollution and weather variables to predict the next several days of data on these same variables and also elderly mortality, for example.

Methods of causal analytics overlap with methods for predictive analytics, statistical inference and forecasting, artificial intelligence, and machine learning. Chapter 2 discusses these overlapping methods. Software for predictive analytics is widely available; several dozen software products for predictive analytics, and a handful for prescriptive analytics, are briefly described at the web page “Top

52 predictive analytics and prescriptive analytics software,” www.predictiveanalyticstoday.com/top-predictive-analytics-software/. However, causal analytics is distinct from predictive analytics in that it addresses questions such as “How will the probabilities of different outcomes change if I take different actions?” rather than only questions such as “How do the probabilities of different outcomes change if I observe different pieces of evidence?” Statistical inference is largely concerned with observations and valid probabilistic inferences that can be drawn from them. By contrast, causal analytics is largely concerned with actions and their probable consequences, meaning changes in the probabilities of other events or outcomes brought about by actions. This distinction between *seeing* and *doing* is fundamental (Pearl 2009). Techniques for predicting how actions change outcome probabilities differ from techniques for inferring how observations change outcome probabilities.

Example: Predictive vs. Causal Inference—Seeing vs. Doing

Table 1.3 shows a small hypothetical example data set consisting of values of variables for each of three communities, A, B, and C. For each community (row), the table shows values for each of the following variables (columns): average exposure concentration, C , for a noxious agent such as an air pollutant, in units such as parts per million; income levels, I , in units such as dollars per capita-year; and rates of some adverse health effect, R , such as mortality or morbidity, in units of cases per person-per year.

For purposes of *descriptive* analytics, the dependencies among these variables over the range of observed values are described equally well by any of the following three models (among others):

Model 1: $R = 2C$ (and $I = 140 - 10C$)

Model 2: $R = 35 - 0.5C - 0.25I$

Model 3: $R = 28 - 0.2*I$

For purposes of *predictive* analytics, these three models all predict the same value of R for any pair of C and I values that satisfy the same descriptive relationship between I and C as the data in Table 1.2, $I = 140 - 10C$. For example, for a community with $C = 6$ and $I = 80$, all three models predict a value of 12 for R .

Table 1.3 A hypothetical example data set

Community	Exposure concentration, C	Income, I	Response rate, R
A	4	100	8
B	8	60	16
C	12	20	24

A statistical inference question for these data is: What response rate should be predicted for a community with exposure concentration $C = 10$? A causal question is: How would changing exposure concentrations to 10 affect response rates R in communities A, B, and C?

But for other pairs of C and I values, the models make very different predictions. For example, if $C = 0$ and $I = 120$, then model 1 would predict a value of 0 for R ; model 2 would predict a value of 5; and model 3 would predict a value of 4.

For purposes of *causal* analytics, it is necessary to specify how changing some variables such as C or I would change others such as R . If it is assumed that changing C or I on the right-hand side of one of the above equations would cause R to adjust to restore equality, then model 1 would predict that each unit of decrease in C would cause 2 units of decrease in R . By contrast, model 2 implies that each unit of decrease in C would *increase* R by 0.5 units. Model 3 implies that changing C would have no effect on R . Thus, the causal impact of changing C on changing R is under-determined by the data. It depends on which causal model is correct.

In practice, of course, it is unlikely that the numbers in a real data set would be described exactly by linear relationships, as in Table 1.2, or that the different columns would be exactly linearly related. But it is common in real applications for many different descriptive models to fit the data approximately equally well and yet to make importantly different predictions. A key challenge for causal analytics is discovering from data which of many equally good descriptive models, if any, best predicts the effects of making changes in some variables—those that a decision-maker can control, for example—on other variables of interest or concern, such as health or financial or behavioral outcomes. Whether data suffice to identify a unique causal model, to make unique predictions, or to uniquely estimate the causal impact of changing one variable on the average values of other variables, are often referred to as questions of *identifiability* in the technical literatures of machine learning, statistics, causal analysis, and econometrics. We shall see later how modern causal analytics algorithms determine which effects of interventions can be identified from observed data and how to quantify the effects that can be identified.

Example: Non-identifiability in Predictive Analytics

Table 1.4 provides another small hypothetical data set to illustrate the identifiability challenge in a different way. For simplicity, all variables in this example are binary (0–1) variables.

Table 1.4 A machine learning challenge: what outcome should be predicted for case 7 based on the data in cases 1–6?

Case	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Outcome
1	1	1	1	1	1
2	0	0	0	0	0
3	0	1	1	0	1
4	1	1	0	0	0
5	0	0	0	0	0
6	1	0	1	1	1
7	1	1	0	1	?

Suppose that cases 1–6 constitute a “training set”, with 4 predictors and one outcome column (the right-most) column to be predicted from them. The challenge for predictive analytics or modeling in this example is to predict the outcome for case 7 (the value, either 0 or 1, in the “?” cell in the lower right of the table). For example, predictors 1–4 might represent various features (1 = present, 0 = absent) of a chemical, or perhaps results of various quick and inexpensive assays for the chemical (1 = positive, 0 = negative). The outcome might indicate whether the chemical would be classified as a rodent carcinogen in relatively expensive 2-year live-animal experiments. Chapter 2 reviews a variety of machine-learning algorithms for inducing predictive rules or models from such training data. But identifiability places hard limits on what can be learned and on the accuracy of predictive models learned from data. No algorithm can provide trustworthy predictions for the outcome in case 7 based on the training data in cases 1–6, since many different models fit the training data equally well but make opposite predictions. For example, the following two models each describe the training data in rows 1–6 perfectly, yet they make opposite prediction for case 7:

- Model 1: Outcome = 1 if the sum of predictors 2, 3, and 4 exceeds 1, else 0
- Model 2: Outcome = value of Predictor 3.

Likewise, these two models would make opposite predictions for a chemical with predictor values of (0, 0, 1, 0). If these models are interpreted causally, with changes in predictors on the right side causing the dependent variable (*Outcome*) on the left side to change to make the equality hold, then Model 1 would imply that setting the values of any two of the values for predictors 2, 3, and 4 equal to 1 would suffice to achieve an *Outcome* value of 1, but Model 2 would imply that this can be achieved only by setting predictor 3 equal to 1. Additional models or prediction rules such as

- Model 3: Outcome is the greater of the values of predictors 1 and 2 except when both equal 1, in which case the outcome is the greater of the values of predictors 3 and 4
- Model 4: Outcome is the greater of the values of predictors 1 and 2 except when both equal 1, in which case the outcome is the lesser of the values of predictors 3 and 4

also describe the training data, but make opposite predictions for case. Thus, it is impossible to confidently identify a single correct model structure from the training data in this case (the data-generating process is *non-identifiable* from the training data), and no predictive analytics or machine learning algorithm can determine from these data a unique model (or set of prediction rules) for correctly predicting the outcome for new cases or situations or for determining how manipulating the values of the predictors, if some or all of them can be controlled by a decision-maker, will affect the value of the outcome variable. Chapter 2 discusses conditions under which unique causal or predictive models can be identified from available data and what to do when this is impossible.

Example: Anomaly Detection, Predictive Maintenance and Cause-Specific Failure Probabilities

When a complex system fails, it is often possible in retrospect to identify precursors and warning signs that might have helped the system's operators to realize that failure was imminent. This has inspired the development of anomaly detection and predictive maintenance algorithms to seek patterns in data that can help to predict costly failures before they happen, so that maintenance can be taken to prevent them. One very useful principle is to train an autoencoder to predict a system's outputs from its inputs during normal operation. This trained autoencoder then serves as a model for normal, problem-free operation. An anomaly is detected when the observed outputs stop matching the autoencoder's predicted outputs. (For implementation details, see the free TensorFlow or H2O package documentation on autoencoders, e.g., at <http://amunategui.github.io/anomaly-detection-h2o/>.) Using discrepancies between observed and expected normal behaviors provides a powerful way to detect fraud in financial systems, cyberattacks, and other perturbations in normal operations due to intelligent adversaries, as well as to changes in the performance of system components. Anomaly detection algorithms provide a way to automatically notice the early warning signs of altered input-output behaviors that can show that a complex engineering system—or, for that matter, a human patient—is losing normal homeostatic control and may be headed for eventual system failure.

Several companies, including IBM and Microsoft, offer commercial predictive maintenance software products that go beyond merely detecting anomalies. They apply deep learning algorithms—especially, Long Short Term Memory (LSTM) algorithms for learning from time series with long time lags between initiating events and the consequences that they ultimately cause—to predict specific causes or modes of system failure and to identify components or subsystems that are at risk. They also quantify the probability distributions of remaining times until failure is predicted to occur from various causes (Liao and Ahn 2016). This information can then be displayed via a predictive analytics dashboard that highlights which failures are predicted to occur next, quantifies the probability distributions for remaining time until they occur, and recommends maintenance actions to reduce failure risks and increase the remaining time until failure occurs. In industrial applications, predictive maintenance has significantly reduced both maintenance costs (by avoiding unnecessary maintenance) and costs of failures (by targeting maintenance to prevent failures or reduce failure rates). Similar algorithms have started to be applied in health care recently, for example, to predict heart failure diagnosis from patients' electronic health records (EHRs) (Choi et al. 2017).

Identifying from data the symptoms of fault conditions or potential causes that can eventually lead to systems failure is one key challenge for predictive analytics. Conversely, identifying the potential long-term consequences of currently observed aberrations in subsystem or component performance is another. Practical algorithms and software for meeting these challenges are now advancing rapidly under the impetus of new ideas and methods for machine learning. In doing so, they are

clarifying methods for learning about delayed relationships between causes and their effects from time series data. Chapter 2 will discuss further the important principle that causes help to predict their effects, and how this principle can be used to draw inferences about causation between variables from data recording the values of different variables over time.

Causality Models Used in Prescriptive Analytics

Prescriptive analytics addresses the question of how to use data to decide what to do next. It uses a combination of data, causal modeling or assumptions, and optimization to decide on a best course of action. To do so, it is common practice to model how probability distributions of outcomes would change if different actions were taken. In very simple decision analysis models, outcome probabilities can simply be tabulated for different acts. The “best” act, as defined by certain axioms of rational choice, is then one that maximizes expected utility. This rule is explained and justified in detail in decision analysis.

Normal-Form Decision Analysis

In mathematical notation, the expected utility (EU) of an act a is defined by the following sum:

$$EU(a) = \sum_c u(c) * P(c|a) \quad (1.1)$$

Here, $EU(a)$ is the expected utility of act a , c is a consequence or outcome, $u(c)$ is the utility of consequence c , and $P(c | a)$ is the conditional probability of consequence c if the decision-maker (d.m.) chooses act a . (If needed, the beginning of Chap. 2 provides a quick review of probability and conditional probability concepts and notation; see Eqs. (2.1) and (2.2). The remainder of this section assumes familiarity with both.) $P(c | a)$ is a *causal model* of the probabilistic relationship between exogenous acts and their consequences. Interpretively, Eq. (1.1) says that the expected utility of an act is the mean or average value of the utilities of the consequences that it might cause, weighted by their respective probabilities.

Normative decision analysis describes choosing among alternatives as choosing among different sets of outcome probabilities. Prescriptively, the choice is to be made to maximize expected utility or to minimize expected loss, respectively. In each case, *a decision is represented by the outcome probabilities that it causes*.

Table 1.5 A simple example of a normal form decision table with two acts and three states

	State 1	State 2	State 3
act 1	3	1	4
act 2	1	5	9
	$P(\text{state 1}) = 0.2$	$P(\text{state 2}) = 0.3$	$P(\text{state 3}) = 0.5$

Example: Identifying the Best Act in a Decision Table

Suppose that a decision-maker (d.m.) must choose between two acts, acts 1 and 2, represented by rows in Table 1.5. The consequence of each act depends on which of three possible states, 1, 2, or 3, occurs. The probabilities of states 1, 2, and 3 are 0.2, 0.3, and 0.5, respectively, as shown in the bottom row of Table 1.5. The cells of the table show the rewards, payoffs or expected utilities, expressed on a scale with numbers between 0 to 10, for choosing each act if each state. (The two endpoints of a von Neumann Morgenstern utility function can be chosen arbitrarily, much as the numbers corresponding to the boiling and freezing points of water can be chosen arbitrarily on a temperature scale. Both temperature and utility are measured on interval scales, and such a scale is uniquely determined by specifying the values of two points on it, such as by making 0 the value of the least-preferred outcome and 10 the value of the most-preferred outcome. Luce and Raiffa (1957) provide an excellent full exposition of utility theory and normal form decision analysis.)

In Table 1.5, the utility of the consequence if the d.m. chooses act 2 and state 2 occurs is 5; the utility from act 2 if state 3 occurs is 9; and the utility from act 2 if state 1 occurs is 1. If a particular choice of act combined with a particular state does not determine a unique outcome as the consequence, but only a conditional probability distribution of outcomes, then the numbers in the cells in Table 1.5 should be interpreted as expected utilities (i.e., expected values of the utilities) of the random outcomes for the different act-state pairs. This formulation of a decision problem, in which the d.m. chooses an act (i.e., row of the table) from a set of feasible alternatives, “Nature” or “Chance” chooses a state (i.e., column of the table) at random according to known probabilities from a set of possible states; and the intersection of the selected act and the selected state determine an outcome (or, more generally, the conditional probabilities of different outcomes or consequences, each having a known utility) is called the *normal form* of decision analysis (Luce and Raiffa 1957).

Problem: Given the decision problem data, in Table 1.5, which act, act 1 or act 2, should the d.m. choose? Assume that the goal of decision-making is to maximize expected utility.

Solution: The expected utility of act 1 is chosen is $3*0.2 + 1*0.3 + 4*0.5 = 2.9$. The expected utility if act 2 is chosen is $1*0.2 + 5*0.3 + 9*0.5 = 6.2$. Since $6.2 > 2.9$, the d.m. should choose act 2.

Table 1.5 represents the probabilistic causal relationship between acts and the outcomes that they cause in the form of a *decision table*. In such a table, *acts* are represented by rows, *states* (meaning everything other than the act that helps

determine the outcome) are represented by columns, and *state probabilities* (shown in the bottom-most row) are assumed to be known and to be independent of the act. The numbers in the cells of the table show the (von Neumann-Morgenstern) *utilities* to the decision-maker (d.m.) of different outcomes. Each outcome is determined by an (*act, state*) pairs. These ingredients—acts, states, consequences, state probabilities, and utilities—are all of the elements of a normal form decision analysis. A decision table is a simple causal model: in Table 1.5, choosing act 1 causes outcome probabilities of 0.2 for an outcome with a utility of 3, probability 0.3 for an outcome with utility 1, and probability 0.5 for an outcome with utility 4. This idea of representing acts by the probabilities of the outcomes that they cause (or, as in Table 1.5, the probabilities of the utilities of those outcomes), is often formalized in technical work by representing decision problems as choices among cumulative probability distributions (CDFs), probability density functions (PDFs), or probability measures, which are then used to calculate expected utilities or expected losses. In each case, however, all relevant differences among decision alternatives are assumed to be captured by differences in the outcome probabilities that they cause.

Causal models can be more elaborate, of course. One generalization is to let the conditional probabilities of the states depend on which act the d.m. selects. The choice of act can then affect outcome probabilities both directly, via their dependence on the act, and also indirectly via their dependence on the state, which in turn depends probabilistically on the act. In notation, the probability of consequence c if act a is selected is given by the following sum, expressing the law of total probability:

$$P(c|a) = \sum_s P(c|a, s)P(s|a) \quad (1.2)$$

where the sum is over all possible states, s . In words, the probability of a particular consequence of an act is the sum of the probabilities that it occurs in conjunction with each possible state, weighted by the probability of that state given the selected act. This type of causal structure arises, for example, if the decision variable a is the price per unit that a retailer charges for a good; the state s is the number of units of the good purchased at that price; and the consequence of selling s units at price a is the revenue, a^*s . (In this case, $P(c|a, s) = 1$ for $c = a^*s$ and 0 otherwise.)

The following examples illustrate how the main ideas of normal-form decision analysis can be applied even in cases where there are too many acts or states to be compactly summarized in a small decision table.

Example: Optimizing Research Intensity

Setting: Suppose that a start-up company has a 1-year window in which to solve an applied research and development (R&D) problem. Successfully solving the problem within a year is worth \$1 M; otherwise, the company earns \$0 from its effort on this problem. The company can assign employees to work on the problem. The causal relationship between number of employees assigned and probability of

success is that each employee assigned has a 10% probability of solving the problem, independently of anyone else. (Allowing for teamwork and interaction among employees might be more realistic, but would not affect the points to be illustrated.) Each employee assigned to this effort costs \$0.05 M.

Problem: How many employees should the company assign to this R&D effort to maximize expected profit?

Solution: The R&D problem will be solved successfully unless everyone assigned to it fails to solve it. If N employees are assigned to work on it, then the probability that they succeed in solving the problem is one minus the probability that they all fail to solve it: $P(\text{success} \mid N \text{ employees}) = 1 - 0.9^N$, since each independently has a 0.9 probability of failure and therefore all N together have a 0.9^N probability of failure. The cost of assigning N employees, expressed in units of millions of dollars, is $0.05N$. The expected benefit is $(1 - 0.9^N) * 1$ million dollars. The expected profit from assigning N employees to this effort is therefore $(1 - 0.9^N) - 0.05N$. Clearly, if N is too large (e.g., greater than 20) then the expected profit will be negative, and if $N = 0$ then the expected profit is 0. Searching over this range e.g., using the online “Wolfram Alpha extrema calculator,” or simply evaluating the profit function for each integer in the range from 0 to 20, reveals that the most profitable number of employees to assign to the R&D effort is 7. For readers familiar with R, the following R script returns the answer:

```
profit <- c(1:20)
for (N in 1:20) {
  profit [N] <- 1 - .9^N - 0.05*N}
print(which.max(profit))
[1] 7
```

In this example, the outcome, *profit*, depends on the decision variable, N = number of employees assigned, both directly through the effect of N on cost, and indirectly through the effect of N on the probability that the effort will succeed. The general formula in Eq. (1.2) is instantiated in this case by the specific assignments $P(c \mid a, s) = 1$ for $c = s - 0.05a$ and $P(c \mid a, s) = 0$ otherwise and $P(s = 1 \mid a) = 1 - 0.9^a$, where we define the state variable s to have value 1 if the R&D problem is solved successfully within a year and $s = 0$ otherwise and we define the decision variable as $a = N$ = number of employees assigned to the effort.

Example: Optimal Stopping in a Risky Production Process

Setting: Suppose that a hazardous production process or facility, such as a chemical plant, an oil rig, or an old mine, produces a profit of \$10 M per year while it is operating. If it fails due to a catastrophic accident during operation, this costs \$50 M and destroys the process. The random lifetime until such a catastrophic accident

occurs is uniformly distributed between 0 and 60 years with a mean of 30 years. The process can be voluntarily closed down at any time before failure occurs.

Problem: When should the production process be voluntarily closed (if it has not yet failed) to maximize expected profit? For simplicity, ignore interest rates, acquisition and replacement costs, and discounting: assume that the goal is to maximize expected profit, where the profit is given by $10A$ if the process is voluntarily closed at age A and is given by $10*T - 50$ if the process fails at age T before it reaches age A . More detailed and realistic objective functions can be devised, but this simple one suffices for purposes of illustration.

Solution: One way to solve this problem is to consider different decision rules, evaluate the expected reward from each one, and use optimization to find the decision rule that maximizes the specified objective function. Suppose that the decision rule adopted is to voluntarily shut down the process when and if it reaches age A years. The probability that it survives until age A without an accident is $1 - A/60$, and the reward if it does so is specified to be $10A$. On the other hand, if an accident terminates the process at some time before age A , which occurs with probability $A/60$, then the expected net reward is $10*(A/2) - 50$. This is because the expected age at failure, given that failure occurs before age A , is just $A/2$. (Conditioning on the event of failure by age A replaces the original uniform distribution between 0 and 60 years for the failure time with a new uniform distribution between 0 and A years, having conditional mean $E(T | T < A) = A/2$.) The expected value of the process with the decision rule determined by A is therefore $(A/60)*(10*(A/2) - 50) + (1 - (A/60))*10*A$. The value of A that maximizes this average reward per unit time can be found using free online solvers such as the WolframAlpha Max/Min Finder widget, or can be searched for (here, to the nearest year) in R as follows: $A \leftarrow 1:60$; $J \leftarrow (A/60)*(10*(A/2) - 50) + (1 - (A/60))*10*A$; $\text{print}(\text{which.max}(J))$. The solution is that the process should be closed when it reaches age 55 years.

A different way to solve this problem is to apply the following concepts from reliability theory and economics. Intuitively, the process should be operated as long as the expected marginal benefit from continuing for a little longer exceeds the expected marginal cost. The expected marginal benefit from additional product produced by continuing for an additional time increment of length dt is $10dt$. The expected marginal cost is $50 h(t)dt$, where $h(t)$ is the age-specific hazard function for failure at time t . From reliability theory, a formula for $h(t)$ is $h(t) = f(t)/(1-F(t))$ where $F(t) = P(T \leq t)$ is the cumulative probability distribution function (CDF) for random lifetime T , i.e., for failure by time t ; $1 - F(t) = P(T > t)$ is the survivor function, i.e., the probability that the process survives until time or age t without failing; and $f(t) = F'(t)$ is the probability density function (PDF) for the failure time. For a uniformly distributed lifetime, the hazard function $h(t)$ increases with time. The process should be operated until the expected marginal benefit from continuing equals the expected marginal cost, $10dt = 50 h(t)dt$, i.e., until $h(t) = 0.2$. Free online calculators for various hazard functions are available, such as the one at http://reliabilityanalyticstoolkit.appspot.com/normal_distribution for normally distributed lifetimes, but for a uniform distribution between 0 and 60 years, $h(t)$ can readily be

calculated by hand: $h(t) = f(t)/(1 - F(t)) = (1/60)/(1 - t/60) = 1/(60 - t)$. Equating this to 0.2 and solving for t yields $60 - t = 5$, or $t = 55$ years. Hence, the process should be voluntarily closed at age 55 years if it has not failed by then. The solution in this case has the attractive and intuitive form of an *instantaneous look-ahead rule* (or, if decisions are made at discrete intervals, a one-step look-ahead rule) that calls for continuing an activity until the marginal costs of doing so are no longer less than the marginal benefits, and then stopping. This works because the hazard function is increasing: if it is not worth continuing at time t , it will be even less worthwhile to continue thereafter.

In this example, the act and state are both continuous variables: when to stop the process and when the process will spontaneously fail, respectively. The conceptual framework of normal form decision analysis still applies, but finding the best act in the way just illustrated requires using specialized concepts such as the age-specific hazard function, $h(t)$. The problem could also be solved by brute force by trying out different decision rules, i.e., ages at which to voluntarily close the process, and using simulation to estimate the average profit from different choices and to identify the best choice. A more efficient version of this approach uses optimization algorithms to home in more quickly on the best decision: this is a key idea of *simulation-optimization*, discussed further later in this section.

Example: Harvesting Timber

Suppose that a commercial timber stand has a market value that grows linearly with time: if it is harvested at age t years, its value will be $0.1t$ million dollars. There is a 5% probability of a fire that destroys the current stand and resets its value to zero. At what age should the timber be harvested? If we again focus just on the trade-off between the risk and benefit of waiting to harvest, ignoring interest rates, discounting, costs, age structure of the tree population, price changes and uncertainties, and other important realistic details, then the simplified problem consists of finding the age for which the expected marginal benefit of waiting to harvest equals the expected marginal cost. Waiting an additional time increment of length dt brings an expected marginal benefit of $0.1dt$ million dollars and an expected marginal cost of $(0.05dt)*0.1*t$, which is the probability of loss due to fire (approximately $0.05dt$) times the size of the loss if a fire occurs, which is the accumulated value of the timber so far, $0.1t$. Equating these and solving for t yields $0.1 = 0.05*0.1*t$, or $t = 20$ years. Thus, the optimal decision rule for the simplified problem is to grow the stand for 20 years and then harvest it if it has not yet burned down. Real-world harvesting decision rules and calculations are more complex because they must consider the various important factors that we have omitted, especially the opportunity cost of foregoing interest on sales while harvesting is postponed. However, many such *renewal reward processes*, in which the process resets to its initial condition (i.e., “renews” itself) when certain random events occur (such as a forest fire or harvesting when a pre-specified stopping time is reached) can be solved by formulating an objective function expressing the average profit per unit time per cycle, i.e., per

interval between renewals of the process, and then choosing the decision variables to maximize this objective function.

Markov Decision Processes

An important generalization of one-shot decision-making allows the d.m. to make a *sequence* of decisions. Each decision produces an outcome. Outcome probabilities in each period may depend on the state at the start of that period; on the decision (or act) selected; and possibly on the next state, i.e., the state to which the system transitions by the start of the next period. Acts may also affect the probabilities for the next state. A *Markov decision process* (MDP) models such a sequence of decisions and states over time, where the decisions taken in one period affect the probabilities of the states for the next period as well as the rewards received in the current period. More specifically, the d.m. chooses an act in each period. This act, together with the state of the system at the start of the period, determines, or causes, conditional probabilities of different outcomes. The outcome has two components, both an *immediate reward* to the d.m., and also a *next state* of the system. In other words, the d.m.'s choice causally affects not only the probability distribution of immediate rewards, which in general depends on both the chosen act and the current state; but also the probabilities of making transitions from the current state to each possible next state. For example, if the state is the amount of inventory at the start of the period and the decision variable is the number of items purchased and added to inventory in this period, then the next state will be the current state plus the number of new items added to the inventory (the decision variable) minus the number of items sold, and hence withdrawn from the inventory, in this period. The number of items sold reflects current demand for the item, usually modeled as a random variable. The reward for this period is then the net profit from these purchases and sales. The probability distribution for the next state is determined by the decision of how many items to purchase and by the probability distribution for the number of items demanded in this period.

MDPs with known rewards and state transition probabilities can be solved via well-developed optimization algorithms to maximize objective functions such as the expected net present value of rewards or the average reward per period. To “solve” an MDP means to specify what act to take in each state to maximize the objective function. More generally, a *decision rule* or *policy* is a function that maps available information to decisions: it specifies what to do whenever a decision must be made, given the information available then. An *optimal decision rule* or *optimal policy* is one that optimizes the expected value of an objective function, e.g., maximizing expected utility or minimizing expected loss. In an MDP, a decision rule maps the current state, such as the amount of inventory on hand at the start of a day, to a feasible act to take from that state, such as how much additional inventory to purchase that day.

A quick recap of standard theory for MDPs follows. Readers who do not thirst for such details can safely skip them. If an optimal policy is followed, then the expected discounted value of the MDP starting from current state s , denoted by $V(s)$, can be expressed as follows:

$$V(s) = \max_a \left[R(a, s) + \gamma \sum_{s'} P(s'|a, s) V(s') \right] \quad (1.3)$$

Here,

- s denotes the current state
- s' denotes the next state
- a = current act
- $V(s)$ = expected discounted value of the stream of rewards generated by the process starting from state s , assuming that an optimal policy is followed henceforth.
- $R(a, s)$ = immediate reward (or, if it is uncertain, the expected value of the immediate reward) from taking act a in state s . In some models, this *reward function* is generalized to depend not only on the current state and act, but also on the next state. In this case, if the reward function is specified by a table, then instead of showing the value of the reward for each pair of values (a, s) , it is necessary to show its value for each triple (a, s, s') where s' denotes the next state.
- $P(s'|a, s)$ = conditional probability that the next state is s' , given that the current state is s and that the act taken now is a . This is also called the *transition probability* for a transition to next state s' , given that act a is chosen from current state s . It provides a simple causal model for the relationship between the current choice of act and the probability distribution for the next state. It could be diagrammed as $a \rightarrow s' \leftarrow s$, signifying that this probability distribution depends on, or is determined by, the current state and on the current act. When the sets of states and acts are finite, the array of numbers $P(s'|a, s)$ as a , s , and s' range over all possible acts starting from the current state, all states, and all possible next states starting from the current state, respectively, constitute a *conditional probability table* (CPT) for the value of the next state given the current state and act. In continuous time, instead of transition probabilities, the causal dynamics of Markov processes are described by *transition intensities* giving the expected rate of transitions (in units of transitions per unit time) from the current state to each possible next state. The probability of each next state is then the ratio of the transition intensity into it from the current state to the sum of these transition intensities over all possible next states.
- $\sum_{s'} P(s'|a, s) V(s')$ = expected value of the future reward from the process, starting one period from now. This is its expected value starting from each possible next state weighted by the transition probability that it will indeed be the next state, given the current state and act.
- γ = one-period discount factor, i.e., the amount that a reward postponed by one period is worth now.

- $R(a, s) + \gamma \sum_{s'} P(s'|a, s)V(s')$ = expected value starting from the current state, assuming that optimal decisions are taken from each state henceforth. It is the sum of the immediate rewards and the expected discounted value of future rewards.
- By definition, the best act to take now is the one that maximizes expected value starting from the current state. The “ \max_a ” on the right side of the equation signifies that this is the act selected.

Readers familiar with dynamic programming optimization methods will recognize this equation as a version of the Bellman equation. In words, it states that the optimized value of the process (meaning the expected discounted value of the stream of rewards that it generates over time if optimal decisions are made) is the sum of the immediate reward that it generates in this period plus its optimized value thereafter. The *optimal value function* defined by this equation, once it is known, yields a straightforward way to decide what to do in any state: choose the act that maximizes the optimized value, $R(a, s) + \gamma \sum_{s'} P(s'|a, s)V(s')$. If the set of acts is small and finite (or, more generally, easily searched), so that this expression can easily be evaluated and the optimal act found, then this provides an easily computed optimal decision rule.

In practice, of course, the optimal value function is usually initially unknown. However, if the reward function $R(a, s)$ (or, in some formulations, $R(a, s, s')$), the transition probabilities $P(s' | a, s)$, and the discount factor γ are all known, then the optimal value function can be calculated from via the following *value iteration algorithm*:

1. *Initialize the algorithm* by assigning a numerical value to each state. This assignment may be thought of as an initial guess at the optimal value function. Let's call it $Q(s)$. If there are only a few states, then $Q(s)$ can be shown as a table with a row for each state and with two columns, the first listing each state and the second showing the value for each state. If there is no other information, then the values in this initial guess at the optimized value function may be assigned arbitrarily, or at random. If there is some information about which states have the highest values (as in many chess-playing heuristics, for example), then using it to guess at the optimized value function may speed convergence of the value iteration algorithm.
2. *Iteratively improve the estimated value function.* This is done as follows. Denote the current iteration (or guess or estimate) for the optimal value function by $Q(s)$. This function gives a numerical value, the estimated expected discounted value of current and future rewards, for each possible state, s . Similarly, denote by $Q(s, a)$ the estimated expected value of taking act a in state s and then acting optimally in future, where this estimate is formed using the current iteration $Q(s)$. No confusion should arise from the use of the same function name, Q , for both of these closely related functions: $Q(s)$ denotes the estimated optimal value of the MDP starting from state s if optimal decisions are made, and $Q(s, a)$ denotes the estimated value of the MDP starting from state s if act a (which may not be optimal) is taken now and then all future decisions are made optimally. In symbols,

$$Q(s, a) = \sum_{s'} (P(s'|a, s)(R(a, s, s') + \gamma Q(s')). \quad (1.4)$$

Here, $Q(s')$ on the right side refers to the current guess at the optimal value starting from next state s' . $Q(s)$ is derived from $Q(s, a)$ by “optimizing out” the current act, a , i.e., by choosing it to maximize $Q(s, a)$; in symbols, $Q(s) = \max_a Q(s, a)$.

Equation (1.4) shows that the estimated expected value of taking act a in state s is the sum over all possible next states of the conditional probabilities $P(s'|a, s)$ of the next state (which is probabilistically caused by the choice of act and by the current state), multiplied by the sum of the immediate reward $R(a, s, s')$ and the discounted value starting from the next state reached, $\gamma V(s')$. (For generality, we address the case where the immediate reward depends not only on the current state and act, but also on the next state. If it depends only on the current state and act, then the equation would simply to $Q(s, a) = R(a, s) + \gamma \sum_{s'} (P(s'|a, s)Q(s'))$.) With this estimate of the expected value of the process starting from state s if act a is selected, it is clear that estimated expected value is maximized simply by choosing an act that maximizes the function $Q(a, s)$. Since the current state s is known and $P(s'|a, s)$, $R(a, s, s')$, γ , and the current estimate of the optimal value starting from the next state, $Q(s')$, are all known quantities, $Q(a, s)$ can be evaluated for each a and its maximum achievable value determined. This estimated maximum achievable value starting from state s , denoted by $\max_a Q(a, s)$ in mathematical notation, now becomes the new estimate of $Q(s)$, the optimal value starting from state s . In other words, we can improve the old estimate of $Q(s)$ by replacing it with an updated estimate, as follows:

$$Q(s) \leftarrow \max_a Q(a, s). \quad (1.5)$$

The update arrow “ \leftarrow ” means that the old value on the left is to be replaced by the new value on the right. Once the function $Q(s)$ has been updated for all states via Eq. (1.5), this new estimated function can be used to derive updated values for $Q(a, s)$ via Eq. (1.4). These then lead to further updated values of $Q(s)$ via Eq. (1.5), and the iteration between Eqs. (1.4) and (1.5), updating $Q(a, s)$ and $Q(s)$ respectively, continues until the values of $Q(s)$ and $Q(a, s)$ stop changing appreciably, i.e., to within the desired level of accuracy, between successive iterations. That such convergence will occur, and that the resulting $Q(s)$ derived at the end of this process is indeed the true optimal value function $V(s)$ (to within the desired level of accuracy used in deciding when to terminate the iteration), follow from mathematical analysis of the value iteration algorithm, which establishes that the iteration of Eqs. (1.4) and (1.5) converges to a unique fixed point at $V(s)$ (Kaelbling et al. 1996).

The value iteration algorithm just described is one of three main classical approaches, each with many variations, for solving MDPs. Linear programming and policy iteration are the other two. *Policy iteration* works by iteratively improving an initial policy, which may be random. Recall that a *policy* for an MDP is an assignment of acts to states, specifying which act to take in each state. An optimal policy maximizes an objective function such as expected discounted reward or average reward per period. Given any non-optimal policy (including one formed by randomly selecting which act to take in each state), an improved policy is created

by assigning to each state the act that maximizes the objective function when all future decisions are made according to the current policy. Iterating such policy improvement steps leads to an optimal policy; see Kaelbling et al. (1996) and references therein for details. Linear programming provides an alternative way to optimize the objective function by assigning acts to states subject to the linear constraints implied by Eq. (1.4).

For practical work, free MDP solvers such as the MDPtoolbox package in R provide efficient numerical optimization algorithms that return optimal policies as outputs, given as inputs a finite set of states, a finite set of acts, act-dependent transition probability matrices $P(s'|a,s)$, a reward matrix specifying the immediate reward from taking each act in each state (with the option of having the reward also depend on the next state), and a discount factor, γ . The transition probability matrix for a given policy can be estimated from appropriate data on observed acts and transitions, if such data are available, using statistical software such as the markovchain package in R. The act-specific state transition probabilities and rewards together constitute a causal model linking decisions to the probabilities of consequences, namely, rewards and state transitions. MDP solvers such as MDPtoolbox use this causal model to prescribe what is best to do in each state. They apply value iteration, policy iteration, or other algorithms to solve the MDP and determine the optimal policy.

Once the best act for each state has been specified, the MDP becomes just an ordinary Markov process, with the probabilities for each possible next state fully determined by the current state (and by the optimal act prescribed for that state; since this act in turn is determined by the state, the current state fully determines the probabilities for the next states). The acts are then said to have been optimized out, leaving a purely random (Markov) process that is determined by the initial state (or probability distribution for the initial state, if it is uncertain) and by the optimal policy. The probabilities of sequences of future states, i.e., of the future state trajectories, of this process can now be solved for or simulated using standard algorithms for Markov processes. For example, the markovchain package in R provides straightforward commands for carrying out calculations of future state probabilities and for probabilistic simulation of future state trajectories using specified state transition probabilities. For systems that can be modeled well by MDPs, such software enables predictions of the probabilities of future state trajectories and related quantities, such as the probability distribution of the time at which the optimized MDP will first enter or leave a specified target subset of states, or the probability distributions of the reward or of the times spent in different states over any specified time interval. Thus, excellent predictive analytics are possible for such systems, provided that enough data are available to determine the optimized transition probabilities.

MDPs have been used in a variety of practical applications, ranging from optimizing screening for cervical cancer (Akhavan-Tabatabaei et al. 2017) to designing treatment pathways to manage progression of diseases in individual patients by trying to optimize assignment of treatments to disease states (Bala and Mauskopf 2006). However, the main practical value of MDP methodology is that it serves as a fruitful point of departure for more flexible and realistic decision optimization models.

Improving MDPs: Semi-Markov Decision Processes and Discrete-Event Simulation (DES) Models

In practice, MDPs are often too simple to describe complex systems or diseases because the fundamental *Markov assumption* does not hold. This is the assumption that the probabilities for the next states depend only on the current state and act, and hence are conditionally independent of past states given the current state (the “memoryless” property of Markov processes). This property also implies that the transition probability (or, in continuous time, the transition intensity, i.e., expected transitions per unit time) into each possible next state does not depend on how long the process has been in the current state. This in turn implies that the dwell time in each state has a geometric distribution for discrete-time systems, or an exponential distribution for continuous-time systems. These are restrictive assumptions in trying to model the real world. They can be relaxed by allowing transition probabilities or intensities from the current state to each possible next state to depend on the time since the system entered the current state; the results is called a semi-Markov process. Still more generally, transition probabilities or intensities for occurrences of events can depend on the history of events (including acts by the d.m. or controller) that have occurred so far and when they occurred. This very flexible modeling framework is called discrete-event simulation (DES). DES includes simulation of MDPs and semi-Markov decision processes as relatively simple special cases. It can be carried out using free packages such as SimPy in Python (<https://simpy.readthedocs.io/en/latest/>) or Simmer in R. Considerable ingenuity has been devoted to computational efficiency in such packages, making it practicable to simulate the probabilistic time courses (also called sample paths, trajectories, or histories) of large and complex systems if they are understood well enough so that event occurrence intensities at any moment can be specified as functions of the history to date. Different events and processes and their interactions can be modeled without having to explicitly enumerate all possible states. For example, if a system has 100 components, each of which can be working or failed at any moment, then there is no need to consider transitions among all 2^{100} (which exceeds 10^{30}) logically possible states. Instead, the failures and repairs of each component can be simulated separately, if they are all independent, or the transition intensities between the working and failed states of each component can be specified to depend only on the states of those other components that affect them. The sparsity of causal connections, or probabilistic dependencies, among the components in many systems allows a great reduction in the computational complexity of DES models.

Performance of Prescriptive Models

Given a large data set recording the times of observed transitions (or other events) in a large number of human patients or other observational units (e.g., computers and

other components in communications networks whose reliability and performance are being assessed), together with the histories of covariates thought to affect the event occurrence rates, a risk analyst typically has a choice of fitting a Markov model, a semi-Markov model, or a more flexible DES model to the data to describe causal relationships among the measured variables and event occurrence rates. The technical challenges of estimating such models from observed event history and covariate data are met by software such as the markovchain and SemiMarkov or depmixS4 packages in R (Król and Saint-Pierre 2015). However, using any model to predict the probabilities of future states or state trajectories and to prescribe current decisions based on the model’s assumptions and on available data invites the following crucial questions of model validation, evaluation, and selection:

- *Model validation and evaluation:* How well do the predictions and recommendations from a model actually work, as judged in hindsight and as compared to the predictions and recommendations from other models?
- *Model selection and combination:* Which of the multiple models that can be fit to the data should be used in making predictions and recommendations? How should the predictions and recommendations from multiple models be combined?

For example, how much difference does the extra effort needed to develop a semi-Markov model with state transition intensities that depend on elapsed time in a state make to the quality of predictions and recommendations, compared to fitting a simpler Markov model to the same data?

Chapter 2 introduces modern algorithms (such as Random Forest) that include both model validation and evaluation steps and automatic selection and combination of results from multiple models—an approach often referred to as *ensemble modeling* that has proved extremely valuable in improving model-based predictions. Meanwhile, practical experience in different domains also offers some guidance on the value of using more sophisticated vs. less sophisticated models. For example, Cao et al. (2016) find that a semi-Markov model fit to data on disease management and outcomes for heart failure patients substantially out-performed a Markov model with the same states, yielding not only much better fits to the data (i.e., better descriptive validity), but also about a nearly 50% reduction in mean total cost and about a two month increase in mean survival times for patients. On the other hand, Simpson et al. (2009) compared Markov and DES modeling for predicting long-term (5-year) clinical and economic outcomes for HIV patients. They found that the predictions from both models matched clinical results well for a 1-year forecast horizon, but that the DES model had slightly better 5-year predictive validity than the Markov model and provided more detailed predictions for outcomes. Both models agreed in their recommendations for which of several alternative treatment regimens being evaluated generated the lowest costs and best results for disease management. Bennett and Hauser (2013) estimate that applying MDPs and related models, including methods for “partially observable” health states discussed later in this section, to electronic health records can have very substantial benefits in healthcare, increasing measures of health improvement for patients by about 50% while cutting costs approximately in half. As advances in analytics software and

access to plentiful data make it easier to fit more flexible and detailed models to available data, it will become more desirable to use semi-Markov and DES causal models instead of simpler Markov models to achieve the gains in descriptive and predictive validity and the finer detail in predictions that the more sophisticated models offer.

Dynamic Optimization and Deterministic Optimal Control

Markov decision processes (MDPs) have the simplifying feature that the d.m. is assumed to know the current state. The state represents all the conditions other than the d.m.'s act that affect probabilities of outcomes, i.e., rewards and next states. In many practical applications, however, the d.m.'s knowledge is far less complete than this: all the d.m. knows is the values of some observed variables that depend probabilistically on the state, but that do not necessarily fully reveal it. In this case, the concept of a policy or decision rule must be generalized. It is now a rule specifying what act to take as a function of the information available to the d.m. at the time—that is, the values of observed variables—rather than as a function of the true underlying state. An optimal policy or decision rule maximizes the expected value of some objective function conditioned on available information and assuming that future decisions will also be made optimally with respect to the future information that will be available when they are made. Formalizing these ideas has led to useful techniques for solving both stochastic (random) and deterministic dynamic programming and optimal control problems. These techniques have been extensively developed by mathematicians, control engineers, and operations researchers since the 1960s. Practical solution techniques today are largely being driven by advances in machine learning, especially, variations on reinforcement learning, discussed later.

Some important special cases of the general problem of devising high-performing decision rules are as follows. A *discrete-time deterministic optimal control problem* has the form shown in Eqs. (1.6a) through (1.6c).

$$x(t+1) = f(x(t), u(t)) \quad (1.6a)$$

$$y(t) = g(x(t), u(t)) \quad (1.6b)$$

$$u(t) = h(y(t-1)) \quad (1.6c)$$

Following conventional notation, the *state* of the dynamic system in time period t is denoted by $x(t)$ and the controllable input to the system at time t is denoted by $u(t)$; both of these can be vectors if the state and controls have multiple components. These are the same as the state s and the act a of normal form decision analysis, but now renamed x and u in deference to control system tradition and equipped with the time index (t), often shown as a subscript, to explicitly index time periods.

The dynamic evolution of the system is described by the difference Eq. (1.6a), where f is the deterministic law mapping the current state and input to the next state. This is sometimes called the “plant dynamics equation” or “plant model” in control systems engineering for power plants and other industrial facilities. If it changes over time, then Eq. (1.6a) can be generalized to $x(t + 1) = f(x(t), u(t), t)$, or, in subscript notation, $x_{t+1} = f(x_t, u_t)$. Equation (1.6b) shows that the observable output for period t , denoted by $y(t)$, is fully determined by the state and the input decisions (which may include decisions to sample, observe, or experiment) that are taken then, $x(t)$ and $u(t)$, respectively. The output function $g()$ describes this dependency. The observed output (or outputs, if $y(t)$ is a vector) can be used to inform decision-making. Eq. (1.6c) shows the form of a *feedback control policy* in which the input in each period is determined by the observed output from the previous period; more generally, it may depend on observed outputs from many previous periods, i.e., on all the information or history observed so far. Designing the feedback function $h()$ to optimize an objective function or cost function, which may depend on both the control decisions $u(t)$ and the state and output trajectories (e.g., on whether the system ends up in a desired target set of states, and how quickly it gets there) is the central challenge for discrete-time optimal control.

Equation (1.6a) shows that the state at the start of the next period depends only on the state at the start of this period and the control $u(t)$ applied in this period. It constitutes a deterministic dynamic model of the system or situation being controlled, and summarizes the causal relationship between each state and input and the next state. For example, if $x(t)$ represents the number of fish in a commercial fishery, the number of items in an inventory, or the number of customers who have not yet received a new offer at the start of the current period (e.g., the current year, day, or minute, respectively); and if $u(t)$ represents the number of fish caught, the net number of items removed from the inventory, or the number of customers who receive an offer during the current period, respectively, then the system dynamics might be modeled by the first-order linear difference equation $x(t + 1) = (1 + r)x(t) - u(t)$, where the growth rate $r = 0$ for the inventory and offer examples but positive for the fishery example. More interesting dynamics result from nonlinearities. Redefining $x(t)$ as the current fraction of the fishery’s maximum carrying capacity and setting $u(t) = 0$ for the quantity lost to fishing in each period, the logistic difference equation $x(t + 1) = 4x(t)*(1 - x(t))$ displays chaotic behavior starting from most initial states, although of course 0 is a trapping state that is reached from initial values for x of 0, 1, or 0.5. Such difference equations can easily be solved, starting from any initial state, by substituting consecutive states into Eq. (1.6a), generating each from its predecessor. For example, the following R script plots the first 60 periods of the chaotic trajectory of $x(t)$, starting from an initial value in period 1 of $x[1] = 0.2$, for the logistic difference equation $x(t + 1) = 4x(t)*(1 - x(t))$:

```
N <- 60; x <- 1:N; x[1] <- 0.2; time <- 1:N
for (t in 2:N) {x[t] <- 4*x[t-1]*(1-x[t-1])}
plot(time, x); lines(x)
```

In continuous time, the causal model (1.6a) is replaced by the following ordinary differential equation (ODE) model:

$$dx/dt = f(x(t), u(t)) \quad (1.7)$$

(If the dynamics are time-dependent, then this ODE is generalized to $dx/dt = f(x(t), u(t), t)$.) The trajectory of x values starting from any initial state and given any input history (or a closed-loop feedback control law for determining $u(t)$ from observations so far) is derived by integrating this ODE, and freely available, high-quality numerical integration software provides the necessary tools for practitioners to do so. Continuous-time deterministic simulation modeling and solver software packages, languages, and environments, such as the *deSolve* package in R, return numerical solutions to systems of ODEs such as Eq. (1.7) (and also numerical solutions to partial differential equations, delay differential-delay, and differential algebraic equations, in case the relevant causal model has one of these forms).

Solving the optimal control problem to determine an optimal policy for selecting $u(t)$ given the information available at time t , which typically consists of the history of inputs and observed outputs until then, requires specialized techniques. Classical solution algorithms draw on methods of dynamic programming, the calculus of variations, Pontryagin's maximum principle for optimal control, the Hamilton-Jacobi-Bellman equation, and related methods of dynamic optimization; these are well synthesized and explained by Kamien and Schwartz (2012). Sophisticated algorithms based on these methods are increasingly being made available to non-specialist practitioners via relatively easy-to-use solvers for optimal control problems, such as the open-source BOCOP (<http://bocop.org/>) package and the DIDO commercial solver (www.elissarglobal.com/). However, the vast and sophisticated older literature on deterministic and stochastic optimal control is largely being overtaken by developments and software emerging from machine learning, especially reinforcement learning algorithms discussed later in this chapter.

Example: Optimal Harvesting

The study of dynamic optimization and optimal control of dynamic systems often yields intuitively appealing, economically interpretable, qualitative insights into the nature of optimal control policies in many interesting real-world applications. Indeed, in a famous essay, Dorfman (1969) showed that the equations of the Pontryagin maximum principle used in optimal control can be derived and interpreted by purely economic reasoning, e.g., by making moment-by-moment decisions to equate the instantaneous marginal gains from consumption or use of a stock of capital or resource to the marginal value of its contribution to the growth of the stock. For optimal control trajectories, the stock of capital is such that it depreciates at the same rate that it contributes to the net present value of present and future consumption. Dynamic optimization models have been used to characterize optimal harvesting and extraction policies over time for both renewable

resources such as forests (Amacher et al. 2009) and fisheries and for non-renewable resources such as mineral deposits. Optimal policies balance the value of present vs. future consumption while taking into account time preferences (e.g., discount factors) that typically give extra weight to earlier relative to later consumption.

Starting from an initial age-structured population of trees or fish in a managed forest or fishery, respectively, an optimal control law specifies how much of the current population should be harvested, and which individuals should and should not be harvested if they differ in maturity or other attributes, in each period to maximize the value of the stream of consumption of the harvested resource over time. Typical questions analyzed in deterministic optimal control deal with the existence and uniqueness of optimal controls (i.e., harvesting trajectories); whether they are periodic; the amplitudes and length of periodic cycles in harvesting if they exist; how to allocate harvesting effort across areas if the resource is spatially distributed; whether the optimal control laws have simple forms (e.g., harvest all trees over a certain age, and no others); and how the answers to these questions depend on problem parameters such as the discount rate, age-specific growth rates, the prices or values of the harvested resource, and the initial size and composition of the population (Amacher et al. 2009). By contrast, stochastic control models, discussed next, must also consider how probabilities of catastrophic loss (e.g., forest fires), extinction of the population, and random variations in growth rates and prices affect optimal harvesting.

Stochastic Optimal Control, Hidden Markov Models, and Partially Observable Markov Decision Models (POMDPs)

Key ideas of deterministic optimal control can be generalized to probabilistic systems. One way to do so is to replace the causal model described by the deterministic ODE Eq. (1.7) with a probabilistic generalization, corresponding to the discrete-time random difference equation

$$x(t+1) = f(x(t), u(t), w(t), t) \quad (1.7a)$$

where $w(t)$ is a random variable (called a disturbance term) representing uncontrolled inputs from the environment. Constructing a continuous-time analog requires some mathematical care and yields a stochastic differential equation (SDE) that can be solved via stochastic (Ito-Stratonovich) integration (Bertsekas and Shreve 1996). If the state dynamics are described by time-invariant causal laws, as in many physics and engineering systems, then (1.7a) simplifies because the causal laws encoded in the ODE do not depend on time:

$$x(t+1) = f(x(t), u(t), w(t)) \quad (1.7b)$$

Modeling stochastically evolving systems via solution of SDEs has yielded important results and applications in mathematical finance, including the Black-Scholes formula for pricing options and derivatives, as well as in the study of diffusion

processes and more general stochastic processes in physics and engineering. A well-developed field of *stochastic optimal control* deals with both discrete-time and continuous-time optimal control of systems in which the state dynamics include randomness, evolving according to a controlled stochastic process. In many such problems, the observed outputs have conditional probability distributions that depend on the underlying state, but the underlying state is not observed directly. Stochastic dynamic programming provides a fruitful framework for the study and solution of such problems (Bertsekas and Shreve 1996). We defer further discussion of solution methods for stochastic dynamic programming problems to the discussion of causality in learning.

It is also possible to generalize the comparatively simple MDP model with probabilistic causal dynamics given by the transition probabilities $P(s' | a, s)$, which could also be written in discrete time control system notation as

$$P(x_{t+1}|x_t, u_t), \quad (1.7c)$$

Where $x_{t+1} = s'$ is the next state; $x_t = s$ is the current state; $u_t = a$ is the current act; and the transition probability $P(x_{t+1} | x_t, u_t)$ is the conditional probability that the next state will be x_{t+1} , given that the current state is x_t and that the current act is u_t . The key idea for the generalization is that the decision-maker or controller now only sees a signal that depends probabilistically on the underlying state, rather than observing the true state directly. The observation Eq. (1.6b) can be generalized to describe such probabilistic observations by specifying an *observation function*, *experiment*, or *information channel* (all three terms are used) as follows:

$$P(y|a, s) = \text{probability of seeing signal } y \text{ if state is } s \text{ and act } a \text{ is taken} \quad (1.8)$$

Equivalently, this idea can be expressed using systems theory notation by generalizing the observation Eq. (1.6b) to include a *measurement error* or *observation noise* term, $v(t)$:

$$y(t) = g(x(t), u(t), v(t), t) \quad (1.8a)$$

where $v(t)$ is a random variable. (The time argument t can be dropped unless the observation structure itself changes over time.) In discrete time control system notation, Eq. (1.8) can also be shown as

$$P(y_t|x_t, u_t) \quad (1.8b)$$

Here, x_t corresponds to the state s and u_t corresponds to the act a for period t . This formulation assumes that the state holds at the start of a period; an act is taken at that time; and then an observation results that depends on the act and the state for that period. Slightly different formulations use different timing conventions, such as having the current state and act determine the probabilities of observations at the start of the next period. No matter which timing convention is used, the essential point is that the true state is now an unobserved variable (also called a *hidden* or

latent variable) that affects the probabilities of observations. Therefore, decisions must be made based only on observations rather than on perfect knowledge of the state. In many engineering applications, the observation equation is assumed to have the special form

$$y(t) = g(x(t), u(t), t) + v(t) \quad (1.8c)$$

where $v(t)$ is assumed to be a Gaussian (i.e., normally distributed) additive noise term.

A causal model consisting of a Markov chain with information about the state obtained only via an observation function such as Eq. (1.8) is called a *Hidden Markov Model* (HMM). HMMs are widely used to model unobserved causal processes in applications as diverse as handwriting recognition (observed y = handwritten words, unobserved x = the motions executed to produce them) and cancer progression modeling, where clinical measurements and symptoms provide probabilistic signals about the progression of the underlying disease (Taghipour et al. 2017). Software for estimating HMMs and Hidden Semi-Markov models (i.e., models in which transition intensities from the current state to possible next states change with elapsed time in the present state) from sequences of data on observed transition times includes the R packages `msm` for HMMs and `mhsmm` for HMMs and Hidden Semi-Markov models, respectively.

A *partially observable Markov decision process* (POMDP) is an MDP with hidden states. Inputs must be selected based only on observations—that is, on the data obtained via an observation function such as (1.8b)—to optimize an objective function such as expected discounted value of rewards or the time to first reach or first exit from a target subset of states. In formal discussions, it is usual to represent a POMDP by a tuple such as (S, A, T, R, Y, g, b) , where S denotes the set of states, A the set of possible acts, T the transition probability matrix with elements $P(s' | a, s)$, R the reward matrix with elements $R(a, s)$ or $R(a, s, s')$, Y the set of possible signals, g the observation function specifying $P(y | a, s)$, and b a probability distribution for the initial state of the POMDP. The probability distribution b may also be thought of as a Bayesian prior probability distribution or belief about the initial state of the system. After each act and each observation, this belief is updated (via conditioning, e.g., using Bayes’ Rule, which is discussed in Chap. 2) to yield a new belief, or probability distribution for the states, given the history of acts and observations so far. Mathematically, the transitions among these successive beliefs, each of which can be represented by a vector of state probabilities, can be viewed as taking place within a larger MDP whose states are the probability distributions for the states of the underlying POMDP. From this perspective, it is natural to consider applying MDP solution techniques such as value iteration to determine the best sequence of actions, given the history of observations. However, in many practical applications, the number of states in the larger MDP is so large that approximation techniques must be used. Excellent results with performance guarantees for the difference between achieved and theoretically optimal results can be obtained by “point-based” sampling approaches that update beliefs only on a sample of points (i.e., probability

vectors) reachable from the initial belief vector via a sequence of actions (Shan et al. 2013). POMDP solvers are now widely available, including pomdp-solve (written in C), MDPSOLVE (written in MATLAB), ZMDP (written in C++), and others (<https://github.com/JuliaPOMDP>). These implement both exact solution methods for small POMDP problems and sophisticated numerical solvers for larger-scale POMDPs.

Example: Administering Antibiotics to Avoid Septic Shock

Tsoukalas et al. (2015) consider the dynamic decision problem of determining which antibiotics to administer to a patient when, and at what doses, to control sepsis and avoid potentially lethal responses to bacterial infections. Figure 1.4 shows a clinical decision support dashboard developed for this purpose. In the upper left-hand corner are the subjectively assessed rewards for each unit of time that a patient spends in each health state; these may be thought of as expected utilities. The health states considered range from *Healthy* at the top of the list down through increasingly deteriorated states ending in *Septic Shock* and then *Death*. The goal of decision-making in this case is to postpone or reverse transitions from better to worse states by

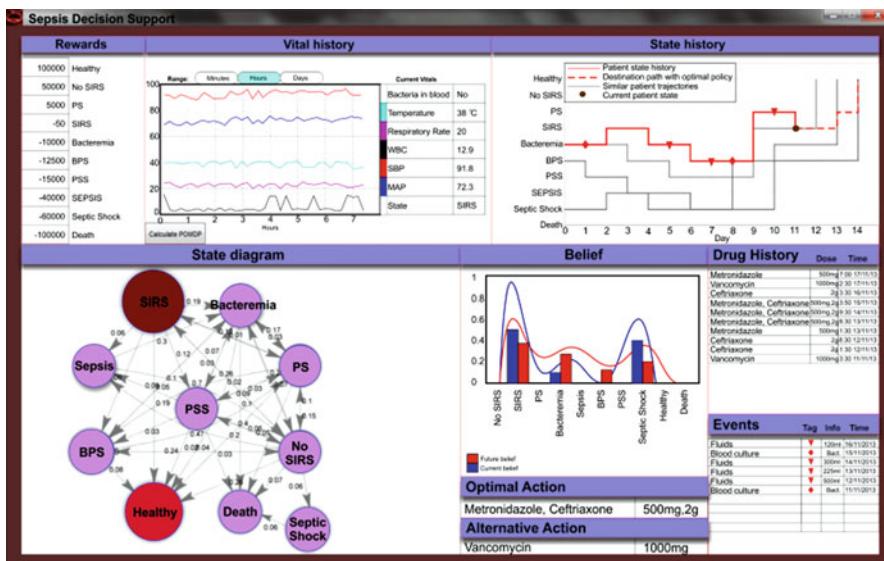


Fig. 1.4 Dashboard for sepsis risk management control problem. Source: Tsoukalas et al. (2015) www.ncbi.nlm.nih.gov/pmc/articles/PMC4376114/. Copyright © Athanasios Tsoukalas, Timothy Albertson, Ilias Tagkopoulos. Originally published in JMIR Medical Informatics (<http://medinform.jmir.org/>, 24.02.2015. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>)

administering antibiotics at appropriate times, doses, and formulations to maximize the expected reward from the time spent in various health states.

To the right of the rewards in Fig. 1.4 are “Vital history” summaries of the observed data, consisting of time series histories and current values for observed variables such as temperature and white blood cell counts (WBC); these are the components of the y vector of observed data values produced by the observation function of the POMDP. In the upper right are patient state trajectories and anticipated target “destination trajectories” that can probably be achieved via optimal policies for administering antibiotics. At the lower right are recommended optimal next actions and alternative nearly-optimal next actions based on inferences about the probabilities of different health states for the patient (The “Belief” profiles shown for current and next states) and records of the acts taken recently “Drug History” and other events. The lower left shows estimated transition probabilities among states. The authors conclude that this POMDP-based tool greatly increased the proportion of patients whose next transitions were to better health states and that mortality and length-of-stay were also predictable with accuracies of about 70–80% using the POMDP model.

Bayesian Statistical Decision Theory

The basic causal structure in which outcomes of decisions depend on (i.e., have probabilities that are affected by) an underlying state variable that cannot be observed directly, but whose value affects the probability distribution of observable signals or data, arises very frequently in applied statistical decision theory. This is the normative theory of how one should make decisions under uncertainty when informed by statistical evidence from observed data and by a causal model, theory, or beliefs about how the probabilities of observations and outcomes depend on the unobserved state (Luce and Raiffa 1957). The abstract structure of many such problems can be diagrammed in DAG form as follows:

$$a \rightarrow c \leftarrow s \rightarrow y, \quad (1.9)$$

indicating that the consequence or outcome c depends on the d.m.’s choice of an act, a , and on the state variable, s , and that the observable data or signal y also depends on state s . If the data y result from a costly sampling effort or experiment designed by the d.m., then there could also be an arrow from a to y to indicate the statistical dependence of the data on these design decisions, such as what sample sizes to use and what variables to collect data on. The quantitative elements of the statistical decision problem are as follows:

- Conditional probability tables (CPTs) for the outcomes and observations, $P(c | a, s)$ and $P(y | a, s)$, respectively. If the data observed does not depend on the act, then the CPT for y simplifies to $P(y | a)$.
- A utility function for consequences, $u(c)$

- A marginal probability distribution for the state, $P(s)$.

These elements complete the specification of the Bayesian statistical decision theory problem shown in the DAG. The marginal distribution for the state, $P(s)$, is also called the *prior distribution* for the state, as it summarizes knowledge or beliefs about the state, expressed via probabilities for the possible values of the state variable, prior to conditioning on the data, y . The standard solution approach to such a statistical decision theory problem, as detailed at the beginning of Chap. 2, is to use Bayes' Rule to infer updated probabilities for the state after conditioning on the observed data. This updated distribution of state probabilities, denoted by $P(s|y)$, is called the *posterior distribution* for the state, because it is formed after observing the data. Together with the utility function $u(c)$ and the conditional probabilities of consequences specified by the causal model CPT, $P(c|a, s)$, the posterior distribution allows the expected utility of each act to be evaluated after taking the data into account, as follows:

$$EU(a|y) = \sum_0 u(c) * P(c|a, y) \quad (1.10a)$$

The conditional probability of each consequence, c , given act a and observed data y , in turn, is found from the causal model $P(c|a, s)$ via an application of the law of total probability:

$$P(c|a, y) = \sum_s P(c|a, s)P(s|y) \quad (1.10b)$$

Finally, the posterior probability distribution for s is given by Bayes' Rule applied to the observation model $P(s|y)$ and the prior probabilities for states, $P(s)$:

$$P(s|y) = P(y|s)P(s)/\sum_{s'} P(y|s')P(s') \quad (1.11)$$

where the sum in the denominator is taken over all possible states. (As usual, if the state is a continuous variable, then sums are replaced by integrals and discrete state probabilities are replaced by probability density functions.)

If the d.m.'s choice of act affects the data collected, so that an arrow from a to y is added to DAG model (1.9) then appropriate act-specific conditional probabilities must be used:

$$P(s|y) = P(y|a, s)P(s)/\sum_{s'} P(y|a, s')P(s') \quad (1.11a)$$

Likewise, if the state probabilities also depend on the act, then the calculation would be:

$$P(s|y, a) = P(y|a, s)P(s|a)/\sum_{s'} P(y|a, s')P(s') \quad (1.11b)$$

None of these variations changes the basic idea, which is to choose an act that will maximize expected utility after state probabilities are updated based on observations using Bayes' Rule and appropriate conditional probabilities.

When the sets of possible acts, states, and observations are all small and discrete, so that explicit probability tables (CPTs) can be developed, the entire process of evaluating the expected utility for each act and recommending an optimal act given any observed (or assumed) observations can be automated. Bayesian network (BN) and influence diagram (ID) software, introduced in Chap. 2, automates the required calculations, producing optimal statistical decision rules mapping observations to decisions (y to a) and calculating their expected utilities. *Value of information* (VOI) calculations of how expected utility changes if more data are collected to inform decisions can also be used to optimize sample size decisions and experimental designs. Chapter 2 provides some simple examples of optimal statistical decision-making using influence diagram software for causal modeling and Bayesian inference from observations to optimize decisions.

Among the very many practical applications of Bayesian statistical decision theory are the following:

- *Statistical estimation:* The observations y consist of sample data. The unobserved state, s , is a population parameter of interest, such as the mean value of an attribute or the proportion of individuals with a certain characteristic in the sampled population. The act a is an estimate of s . An optimal decision-rule selects a to minimize the expected value of a loss function $L(a, s)$ or, equivalently, to maximize the expected value of a utility function that depends on the act and state. Formulating statistical estimation problems in these Bayesian decision-theoretic terms provides an explicit rationale for many traditional statistical procedures. For example, if the loss function from the error in a point estimate of an uncertain quantity is quadratic, $L(a, s) = (a - s)^2$, where s is the true but unknown quantity and a is the point estimate decided on, then the Bayesian optimal statistical decision is to use the mean of the posterior distribution of s , conditioned on all available data, as the point estimate, a . This is the best estimate, in the sense that it minimizes expected loss, i.e., mean square prediction error. If the loss function is the absolute value of the difference between the true and estimated values, then the optimal statistical decision is to use the median of the posterior distribution as the best (expected loss-minimizing) estimate. Such results show conditions under which different traditional estimators, such as the mean and median, are optimal.
- *Change-point analysis (CPA):* The hidden state, s , describes a data-generating process, and the inference task is to infer from data whether, when, and how the data-generating process has changed. The data y consist of time series observations. The decision variable could specify a classification of the process (e.g., changed vs. not changed) or an intervention to be taken when it is concluded from the data that the process being observed has (probably) changed.
- *Statistical quality control:* The hidden state is the true quality (e.g., probability of producing a conforming item), the data consists of observed results of inspections, and the decision rule specifies whether to perform a costly intervention (e.g., replacing or recalibrating a machine used in the production process) based on the sample data.

- *Lot acceptance sampling:* The hidden state is the true proportion of defectives in a lot or consignment of items received for approval. The act a specifies a sampling plan and a decision rule specifying whether to accept or reject the lot (or, in multi-stage plans, whether and how to sample further before making a final accept/reject decision) based on the results of sampling so far, y .
- *Diagnosis:* The unobserved state, s , is the true state of a patient or system being examined. The act a specifies a label (e.g., a disease name or a classification or description of the system's state). An optimal decision rule minimizes the expected cost from misclassification errors. In adaptive decision support systems, the optimal decision rule also specified what tests to perform next, if any, given the results seen so far and the costs and statistical characteristics of available tests.
- *Pattern recognition:* This is a generalization of diagnosis. Pattern recognition, or optimal statistical classification, seeks to classify or label the underlying state of a system or situation based on observations whose probabilities depend on the state. The ability to correctly classify or recognize the state from the observed data with high probability has applications in speech, face, and handwriting recognition, fraud detection, oil exploration, and countless other areas of science, technology, and engineering, in addition to medical diagnosis.
- *Statistical forecasting:* Observations up to the present constitute the data, y . The act or decision consists of a prediction for one or more future observations. An optimal prediction rule minimizes the expected loss from prediction errors. Some forecasting systems, called state space models, use the observations y to estimate the current state. If they are not already known, the state dynamics equations describing the evolution of the state over time, such as (1.6a), (1.7) or the state transition probabilities of a Markov chain, depending on the system being studied, are also estimated from data, along with the observation Eq. (1.8). The state dynamics equations are then applied to the estimated state to predict future states. This approach only coincides with the Bayesian statistical decision theory approach under special conditions, such for “linear quadratc Gaussian” (LQG) systems with state dynamics described by linear ODEs (or difference equations, in discrete time), normally distributed (“Gaussian”) additive observation errors, and quadratic loss functions. Moreover, identification of a unique estimated state (and, if they are not already known, dynamic evolution equations for the state) from available observations is not always possible (the “identification problem” for systems). Even when the state estimation problem is well posed and uniquely solvable from the data, the computational challenge of calculating the posterior probability distribution for states—or even its mean—can be formidable. Efficient recursive updating formulas are available in special cases (most notably, the celebrated Kalman filter for LQG systems widely used in aeronautics, electrical engineering, automatic control, and many other areas (Zarchan and Musoff 2015)). But in general, tractable computation of posterior probability distributions for the hidden state requires using approximation techniques such as Sequential Monte Carlo (SMC) sampling methods. These include popular “particle filtering”

algorithms that draw weighted random samples (“particles”) from the current distribution, propagate them through the system dynamics equations, update the samples and weights by conditioning on observations, and then resample using sequential importance sampling (SIS) and sequential importance resampling (SIR) techniques to yield computationally manageable samples representing the posterior distribution (Doucet and Johansen 2009). For practitioners, particle filtering can be easily implemented using software such as the packages pomp and SMC in R.

- *Adaptive optimal control* (Sutton et al. 1992): Observations that are affected by the underlying state of a system are fed into a feedback control decision rule that, in turn, determines what sampling and control actions to undertake next. The decision rule is adjusted over time in light of observations to try to optimize an objective function, e.g., maximizing expected utility or minimizing expected loss.
- *POMDPs*: A POMDP has a hidden state that evolves according to a stochastic causal law such as (1.7c) and that is observed via probabilistic signals as in Eq. (1.8). An optimal decision rule specifies what act to choose in each period, given the observations, to optimize an objective function.
- *Sequential comparison and selection*: The hidden state is the true average reward or success probability for each of multiple alternatives. These could be treatment plans, advertising campaigns, product designs, or other alternatives being compared in order to select the best one, i.e., the one that generates the highest average reward or success probability. The observations consist of the history of trials so far, i.e., choices made and resulting rewards; these constitute samples from the true but unknown reward distributions for the different alternatives. A decision rule specifies what to try next, given the results so far. An optimal decision rule minimizes the expected sum of costs or losses from sampling and from the opportunity costs (or “regret”) of trying less-than-optimal alternatives. A/B testing in marketing science, clinical trials in medicine, and multi-arm bandit (MAB) problems in statistics and machine learning are examples of such sequential comparison and selection problems.

The huge variety of important and useful applications of Bayesian statistical decision theory all have the following common simple causal structure. A hidden state that is not directly observed causes observations, in the sense that the probabilities of the observations are affected by the state. Also, the current state together with a choice of current act cause (i.e., determine the values of, or the probabilities of) outcomes of interest, typically the next state and rewards or losses. Given observed histories of acts and observations, the d.m. seeks to choose next acts to optimize the expected value of an objective function. The expected value is calculated using posterior probabilities that are conditioned on observations via Bayes’ rule. This combination of concepts and methods, which model causality via conditional probability relationships among acts, states, observations, and outcomes and then use it to draw inferences and optimize actions, powers much of modern signal processing, artificial intelligence, automated control, telecommunications, and industrial engineering technology.

Simulation-Optimization

In many applications, conditional probabilities of outcomes caused by different acts are not known or specified as part of a decision problem. They must instead be estimated from experience or from a model. For a system that is understood in enough detail so that its behavior can be modeled accurately via a discrete-event simulation (DES) or continuous stochastic simulation model, the probabilities of outcomes caused by different combinations of values for the controllable inputs can be estimated via simulation. Optimization algorithms can then be applied to the simulation model to seek the best policies or courses of action. For example, a discrete-event simulation model might be composed by linking stochastic models for the transition rates between “working” and “failed” states for the components and subsystems of a reliability system, with failure and repair rates for each component and subsystem being determined by the states (working or failed) of its predecessors in a dependency graph (DAG) model. The input data for such a model include:

- Spontaneous failure rates and repair rates—or, more generally, state transition intensities—for the “basic events,” i.e., those without predecessors;
- Conditional failure and repair rates, or state transition intensities, for each component or subsystem, given the states of its predecessors.

As outputs, such a model can be used to simulate the probability distribution of the random time until occurrence of an undesirable “top event” such as catastrophic system failure, given different maintenance and repair policies as controllable decision inputs. More generally, simulation can be used to estimate the conditional probabilities of failure times for the system, or for the remaining random time until its first passage into a specified target set of states, given the occurrence or non-occurrence of various events such as failures of its components or subsystems. Such conditional probabilities, in turn, can be used to help figure out which subset of costly improvements in component-level reliabilities would most improve the reliability of the system as a whole for the dollars spent, although true optimization usually requires special techniques of combinatorial optimization for reliability systems. Special techniques, such as importance sampling and subset simulation, have been developed for efficient simulation of rare events such as failures in highly reliable systems (Beck and Zuev 2017).

Simulation and optimization are often combined into powerful *simulation-optimalization algorithms* that automatically vary controllable inputs (the values of “decision variables” or “control variables”), estimate expected utilities or other performance measures via simulation, and then adjust the controllable inputs to increase expected utility—or, equivalently, to reduce expected loss—until no further improvement can be found (Amaran et al. 2016; Fu 2015). A relatively recent class of simulation-optimalization algorithms, *Monte Carlo Tree Search* (MCTS), has led to breakthrough advances in difficult decision problems such as those that arise in large POMDPs and in world champion-level games of Go. MCTS first evaluates potential next acts by running multiple probabilistic (Monte Carlo) simulations to see what

might happen after each is selected, and then chooses the act that has the best evaluation averaged over multiple simulations (Silver and Veness 2010; Fu 2016). Causal knowledge about what can happen, as represented by the rules of a game or by the state transitions probabilities of a POMDP, for example, constrains the simulated possible futures. MCTS typically proceeds via the following systematic four-step process, carried out in a search tree in which the nodes represent possible states and the branches (arcs) leaving a node represent possible acts that can be taken from it (Browne et al. 2012):

- Select a node to “expand,” i.e., to evaluate further by generating and evaluating new child nodes for it;
- Expand the search tree by generating additional acts and new child nodes, i.e., states, to which they lead, following the ones that have already been considered;
- Simulate: Evaluate a newly added node by simulating multiple random feasible trajectories (sequences of states and acts) forward from it, choosing subsequent acts via a default policy (possibly just via random selection) where needed, and using the causal model to simulate outcomes. Stop each simulation when it reaches a terminal state (e.g., a winning or losing position in a game or an absorbing state of an MDP or POMDP) or when an evaluation horizon—that is, a maximum desired depth for the search tree—is reached. Then average the resulting rewards from the simulated trajectories following the new act being evaluated. This phase of the MCTS is also sometimes called “policy rollout” since it evaluates a node by simulation (“rolling forward”) using a default policy to select acts and using a causal model to generate resulting states, observations, and rewards.
- Backpropagation: Update the estimated values of the predecessor acts that lead to the newly evaluated act to reflect the simulated average reward from that point forward.

At the end of this process, the next act chosen is the one with the highest estimated value. Although this basic structure is simple, combining Monte Carlo simulation with decision tree and game tree search and evaluation principles that have been known for decades, it has proved extremely powerful for solving problems that were previously computationally intractable (Fu 2016). In practical implementations, quite sophisticated algorithms and heuristics may be used to carry out each step. For example, in applications to large-scale POMDPs, it is common to use particle filtering to represent and update beliefs about the hidden state, and ideas from reinforcement learning and multiarm bandit problems to optimize search priorities (Katt et al. 2017). Some of these ideas are discussed later in the section on causality in learning analytics, which addresses the problem of decision-making with *unknown causal models*.

At the other end of the spectrum of complexity for simulation-optimization are problems in which both simulation and optimization are straight-forward and no sophisticated algorithms are needed, other than the sampling methods already incorporated into basic Monte Carlo uncertainty analysis software. Such relatively simple, straightforward modeling can still be extremely useful for

evaluating and improving proposed policies when enough is known to develop credible simulation models to assess the (approximate) probabilities of different outcomes for each policy alternative via Monte Carlo simulation.

Example: An Influence Diagram (ID) DAG Model for Optimizing Emissions Reductions

Influence diagrams (IDs) provide a well-developed prescriptive analytics framework that is appropriate for many decision analysis problems. Figure 1.5 shows an example of an ID drawn using the *Analytica* software package (Morgan and Henrion 1990). This ID model illustrates considerations that might be used to support a policy decision about how much to reduce air pollution emissions. It consists of *nodes*, representing variables, and *arrows* between nodes representing dependencies between the corresponding variables. Several different types of nodes are allowed in an ID, as follows.

- *Decision nodes*. The green rectangular node at the upper left of Fig. 1.5 labeled “Emissions Reduction” is a *decision node* representing a *decision variable*. The decision-maker chooses the value for such a variable from a set of possible values. In general, an ID may have multiple decision nodes, although Fig. 1.5 has only one. Decision nodes are also called *choice nodes*.
- *Value node*. The pink hexagon at the lower right of Fig. 1.5 labeled “Total Cost” is a *value node*. This is the ultimate output of the decision problem: a measure of the overall impact of the decisions made. Because it is an output, it has only

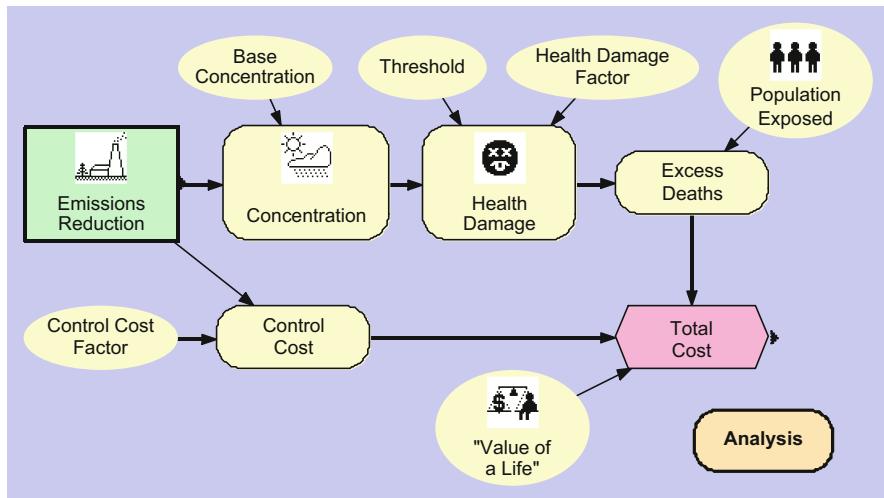


Fig. 1.5 An *Analytica*® influence diagram (ID) for optimizing emissions reduction. *Source:* *Analytica*® web site (http://wiki.analytica.com/index.php?title=Example_Models_and_Libraries)

inward-pointing arrows in the diagram. The decision-maker seeks to set the values of decision variables to optimize the *expected value* of the value node, meaning its average value over many simulation runs of the decision model. Using expected values allows for the possibility of a random relationship between choices and outcomes. This is realistic if other uncertain quantities (modeled as random variables) also affect the outcome. If the value node represents total cost or loss, as in this case, then the challenge for decision analysis is to find values for the decision variables to minimize the expected value of the value node. If the value node represents utility, then the goal is to make decisions to maximize its expected value. In standard operations research and optimization terminology, the value node represents the *objective function* to be optimized in the decision problem.

- *Chance nodes* (yellow ellipses) represent uncertain quantities, modeled as random variables. For example, several of the input nodes (defined as nodes with only outward-pointing arrows, also called *exogenous* nodes) around the periphery of Fig. 1.5, such as *Control Cost Factor* and *Base Concentration*, may be uncertain quantities. In this case, a probability distribution for the value of each uncertain input quantity is specified as part of the ID model. A special case of a chance node is a deterministic node that assigns probability 1 to a specific value: constants are included in this way.
- *Derived nodes*. Nodes (i.e., variables) with inward-pointing arrows have values that depend on the values of the variables that point into them. For example, in Fig. 1.5, the value of *Health Damage* depends on the values of *Concentration*, *Threshold*, and *Health Damage Factor*, each of which may be a random variable. Rectangular boxes with rounded edges are used to represent derived variables in Fig. 1.5. The value of a derived node may be a deterministic function of the values that point into it, but a useful generalization is to allow a derived node to represent a random variable with a conditional probability distribution that depends on the values of the variables pointing into it. The variables that point into a given variable are called its *parents*, and the given variable is called a *child* of any of its parents; thus, an arrow runs from each parent into each of its children. In Fig. 1.5, *Excess Deaths* has two parents, *Population Exposed* and *Health Damage*, and it is a child of each of them. If there are only a few discrete possible values for each of the variables involved, then a *conditional probability table* (CPT) giving the conditional probability of each possible value of a node for each possible combination of the values of its parents provides a way to describe the probabilistic relationship between them. The table for an input node with no parents is called its *marginal probability table*. Chapter 2 discusses other methods of specifying conditional probability relationships that also apply to continuous variables such as age or temperature and to variables with many discrete values, such as daily mortality count in a large population.

The nodes and arcs in an ID are arranged so that there are no directed cycles: it is not possible to start at a node and follow a sequence of arrows that eventually would lead back into the same node. This reflects a constraint that the value of each variable

depends only on the values of other variables, and not on its own value. Such networks composed of nodes joined by arrows that do not form any directed cycles are called *directed acyclic graphs* (DAGs). DAG models are used and discussed extensively in Chap. 2.

A fully quantified ID model specifies a DAG structure together with a marginal probability distribution for the value of each input node and a conditional probability table (CPT) or some other probability model for each derived node giving the probability (or probability density, for continuous variables) of each of its possible values as a function of the values its parents. Thus, once the values of its parents are known, the conditional probability distribution for the value of the node is known. The probability distribution for an output such as *Total Cost* can be quantified for any selected values of the decision inputs by a simple process called *forward Monte Carlo uncertainty analysis*. The process begins by sampling a value from the marginal distribution for each input chance node. Then it samples a value (i.e., draws a value from its conditional probability distribution, given the values of its parents) for each as-yet unevaluated node all of whose parents have been evaluated, thereby evaluating it, i.e., providing a value for it. This process of using sampling to evaluate the as-yet unevaluated nodes with evaluated parents continues until a value has been sampled for the output. Repeating many times for given values of the decision nodes builds up an entire distribution for the output variable, given the selected values of the decision variables. Metaphorically, we might say that the uncertainty probability distributions for the inputs have been “pushed through” the DAG to yield a probability distribution for the output. Varying the values of the decision variables then allows the decision, or setting of decision variable values, with the lowest expected loss or the greatest expected utility to be identified; if necessary, in a problem with many decision variables, optimization algorithms can be used to automate the search for the best values.

Figure 1.6 illustrates the output from this process for the ID in Fig. 1.5. Values of the decision input variable, *Emissions Reduction Factor* (a longer name for

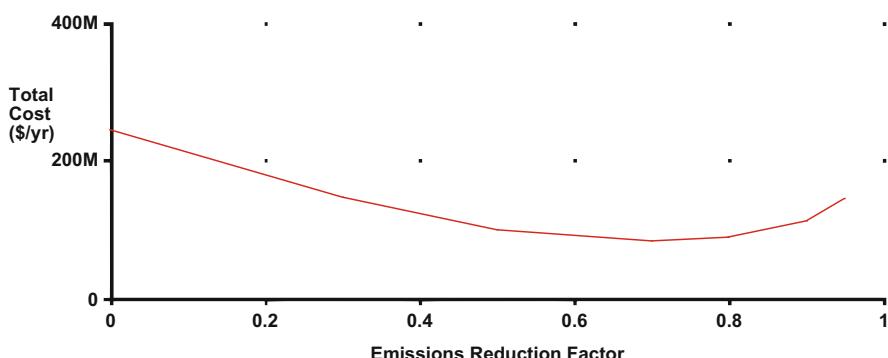


Fig. 1.6 Output from the *Analytica*® influence diagram in Fig. 1.5 showing the expected value of *Total Cost* as a function of the decision variable *Emissions Reduction Factor*

Emissions Reduction in the diagram in Fig. 1.5) are shown on the horizontal axis, and corresponding expected values of the *Total Cost* output variable are shown on the vertical axis. This plot was generated by successively incrementing the values of *Emissions Reduction Factor*, simulating a distribution of *Total Cost* values for each value of *Emissions Reduction Factor*, and then plotting the mean of the simulated distribution of values for *Total Cost* on the vertical axis for each value of *Emissions Reduction Factor* on the horizontal axis. The Analytica® software incorporates sophisticated simulation techniques (such as Latin Hypercube sampling) to increase computational efficiency. It generates the output curve in Fig. 1.6 in a fraction of a second even though each point on the curve represents the average value of *Total Cost* over hundreds of simulation runs. It can equally quickly display upper and lower uncertainty bands around these mean values if desired, such as the high and low values between which 95% of the simulated values of *Total Cost* fall for each value of *Emissions Reduction Factor*.

The output in Fig. 1.6 is easy to interpret prescriptively: the expected value of *Total Cost* is minimized by setting *Emissions Reduction Factor* to a value of 0.7. As often happen when the objective function involved trading off two or more desired attributes, such as low control cost and low excess deaths in this example, the optimal value of the objective function is not very sensitive to the exact choice of value for the decision variable: the expected *Total Cost* curve is flat in the vicinity of the optimal decision variable value, *Emissions Reduction Factor* = 0.7. The prescriptive analysis shows that *if* the ID model in Fig. 1.4 is correct, *then* the best decision of to set *Emissions Reduction Factor* = 0.7.

The directed acyclic graph (DAG) concept illustrated in Fig. 1.5 is of central importance in causal modeling, and it is used extensively in later chapters. Its importance stems from the fact that a DAG shows statistical dependencies and independence relationships between variables. Causation is one way for such statistical dependencies to be created; conversely, effects are not expected to be statistically independent of their direct causes. In a DAG model, each variable is *conditionally independent* of its more remote ancestors, given the values of its direct parents. For example, in Fig. 1.5, although *Excess Deaths* depends indirectly on *Concentration* via the effect of *Concentration* on *Health Damage*, the DAG structure implies that the value of *Excess Death* is conditionally independent of the value of *Concentration* given the value of *Health Damage*. Thus, if we had a data set with three columns of numbers for values of the three variables *Concentration*, *Health Damage*, and *Excess Death* for a large number of cases (e.g., days of observation), then even though each column might be correlated with the other two, the correlation between *Concentration* and *Excess Death* would be zero when only cases with a given value of *Health Damage* are considered. More generally, each node in a DAG is conditionally independent of the values of its more remote ancestors, given the values of its parents. Such conditional independence relationships are implications of a DAG model structure that can be tested statistically using data on the levels of the different variables. Statistical tests for conditional independence tests provide one basis for automatically discovering the DAG structure of variables from data, as discussed in Chap. 2.

In practice, decision-makers must confront the possibility that the model used to produce recommendations is not trustworthy. For example, the DAG model in Fig. 1.5 makes some assumptions about how excess deaths depend on exposure concentration, via a “Health Damage” function involving a threshold and a health damage factor for exposure concentrations above this threshold. A practitioner could double-click on the “Health Damage” node to view this assumed function but might be uncertain about its validity, and hence uncertain about the validity of the model’s conclusions and recommendations from Fig. 1.6. What is often wanted is a reliable way to learn causal models directly from data; to characterize uncertainties about the validity of their predictions; and then to use the models to draw inferences and make recommendations for what actions to take to make preferred outcomes more probable. Figure 1.7 sketches this idealized process. Causal analytics algorithms provide the mapping from data to models of the data-generating process, such as the ID in Fig. 1.5. Monte Carlo simulation then generates probabilistic predictions and expected utilities and uncertainty bands from these models, given any set of values for the controllable inputs, i.e., the decision variables. Specialized simulation-optimization algorithms can vary the values of the decision variables to seek choices that maximize expected utility or minimize expected total cost or loss as predicted by the simulation model. This simulation-optimization step is not shown in Fig. 1.7, since the solution to the optimization problem in Fig. 1.6 is visually obvious. Interested readers can find decision optimization algorithms specifically for IDs in Shachter and Bhattacharjya (2010). More general discussions of simulation-optimization methods are available in many on-line surveys and tutorial papers; in the *Handbook of Simulation Optimization* (Fu 2015); at commercial software sites such as www.solver.com/simulation-optimization; and in the documentation of specialized free software packages such as <http://crantastic.org/packages/scaRabee>.

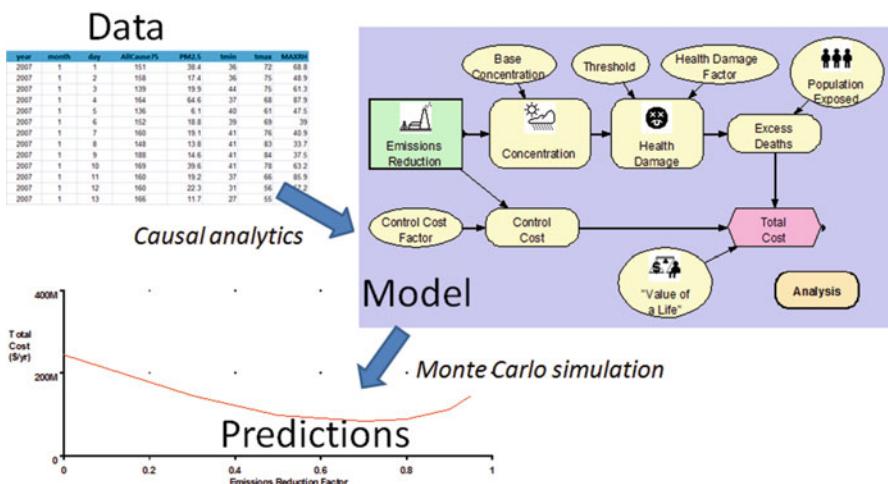


Fig. 1.7 The role of causal analytics. Causal analytics provides algorithms to develop causal models from data and to use them to quantify effects of risk factors and interventions

Example: Forecasting Policy Impacts—Invariance, Causal Laws, and the Lucas Critique in Macroeconomics

Not knowing the causal relationships among variables—especially, which variables affect which others, as revealed by the arrows in a DAG model—undermines ability to use statistical associations between variables to predict how changing one variable would change another. Simpson’s Paradox and similar examples show that even a highly statistically significant association between an outcome variable Y and another variable X need not reveal how, if at all, changing X would change Y . A similar point has been made in macroeconomics, warning that empirically observed relationships that are not causal (or “structural,” in econometrics jargon) cannot be used to predict correctly the effects of policy interventions if the interventions change the conditions that generated the empirical relationships. For example, if it is observed that higher inflation rates are associated with lower unemployment rates, and even if this empirical relationship between them is found to be stable and reliable for making predictions of unemployment from observations of inflation, this still would not imply that intervening to increase inflation (e.g., by printing more money) would reduce unemployment. The vexed history of the original Phillips curve model in macroeconomics illustrates this point. A more general warning against using empirically derived macroeconomic models to predict the effects of policy changes, known as the *Lucas critique* after Nobel Laureate Robert Lucas, helped to convince many economists that only macroeconomic models with strong microeconomic foundations describing causal relationships that remain invariant as policies change could provide a sound basis for predicting consequences of changes in macroeconomic policy (Hoover 2014).

In contrast to empirical models, a “structural” model expresses equations or constraints that stay the same (or remain “invariant,” in more technical terminology) as decision variables are changed. For example, in chemistry, the ideal gas law $PV = nRT$ (where P = pressure, V = volume, T = temperature, n is the number of moles of gas, and R is a constant) expresses a constraint that always holds among these variables in equilibrium. In a system configured so that T is the decision variable and V and n are fixed, this structural law implies that exogenously doubling the temperature by heating the gas would cause its pressure to double. Such laws are useful because they continue to hold even if the experimental conditions under which they were discovered are changed. This makes them useful for descriptive, predictive, and prescriptive analyses. Philosophers of science and causality have noted that invariance of causal laws in the face of policy changes can be viewed as a defining characteristic of causality (Cartwright 2003). Such causal laws are often represented mathematically by *structural equations* that link the values of variables in such a way that a change in a variable on the right side of the equation (such as T in $P = nRT/V$) is understood to cause the dependent variable on the left side (here, P) to change until the equation is once again satisfied. Each such equation can be thought of as representing a *causal mechanism* determining the value of the dependent variable on its left side from the values of the variables on its right side (Simon and Iwasaki 1988). In a DAG model representing this structure, arrows would point from the

variables on the right into the variable on the left. If the relationship between the parent variables on the right and the child variable on the left is probabilistic instead of deterministic, then a random variable can be included on the right-hand side.

The equivalent of such a structural model in an ID is a conditional probability table (CPT) that is invariant across settings and policies. Such a CPT expresses a probabilistic causal law: *if* the parents of a node have specified values, *then* the value of the node has the probability distribution specified by the CPT, no matter what the surrounding context of values of other variable (including decision variables) may be. Seeking such invariant CPTs or structural equations in data provides a different basis for causal discovery algorithms than the tests for conditional independence tests mentioned earlier (Peters et al. 2016; Heinze-Deml et al. 2017). Causal discovery algorithms based on invariance principles are discussed further in Chap. 2.

Causal Study Design and Analysis in Evaluation Analytics

Once a decision has been made and implemented, the next important task for analytics is to collect and analyze data to determine how well the decision has worked—or, more generally, what its effects, intended or otherwise, have been. We call this *evaluation analytics*. It clearly involves causal analysis, insofar as it involves attributing effects to decisions or policies that caused them. It also usually requires modeling how impacts occur over time, since, in the real world, effects of decisions, interventions, and policy changes take time to reveal themselves. By contrast, time is conspicuously missing from many predictive and prescriptive models, including the influence diagram model and results in Figs. 1.5 and 1.6. These relate emissions reductions to predicted resulting changes in total costs without specifying how long those changes will take to occur. As discussed further in Chap. 2, dynamic simulation models can be constructed, using *Analytica*® or other modeling software packages, provided that sufficient knowledge of the system’s dynamics is available so that the required simulation formulas can be specified. The time courses of responses to changes in the inputs to a system can then be simulated. However, such detailed knowledge of dynamic adjustment processes is often lacking, and it is common practice to use models that relate decisions to their projected consequences when equilibrium is achieved without specifying how long this will take. Figure 1.5 is an example. In general, using data to evaluate the consequences caused by a decision requires considering whether enough time has passed for effects to reveal themselves and whether the data collection effort has been designed in such a way that it can support valid causal inferences.

Randomized Control Trials (RCTs)

The gold standard for evaluating the effects caused by policy interventions in both medicine and the social sciences is often considered to be the *randomized control*

trial (RCT). A RCT evaluates the effects of some “treatment” or intervention on responses in populations of experimental “units.” The units could be individual patients when testing the effects of a new drug on high blood pressure; homes or families when testing the effects of bed netting or chlorine tabs for home drinking water on reducing child mortality in a developing country; villages when testing the effects of a microfinance program on subsequent economic indicators; or companies when testing the effects of a new incentive or behavioral “nudge” program on enrollment in employee retirement of health plans. The key aspect of a RCT is that units are *randomly assigned* to the treatment or control groups, or, more generally, to different treatment groups. Systematic, statistically significant differences observed in responses across groups of units receiving different treatments (or none) can then be confidently attributed to the differences in treatments, since randomization removes any other systematic differences among the recipients of different treatments.

Limitations of RCTs include practical and ethical constraints on the possibility of making random assignments and difficulties in generalizing beyond the specific populations or groups studied to other populations or groups to which treatments might be applied. For example, suppose that a well-conducted RCT establishes conclusively that, at a certain hospital, patients randomly selected to receive a new treatment have a significantly higher success rate than similar control group patients not selected to receive the treatment. Even such a strong finding does not guarantee that a similar benefit from the new treatment would hold at other hospitals. This is because other factors that influence the success rate might differ between hospitals, or between the populations that they serve. The challenge of generalizing findings beyond the specific population studied goes by several names, including “transportability” of causal relationships across settings, “external validity” of the findings from a study, and “generalizability” of results. It is important to meet this challenge so that, instead of being limited to drawing narrow conclusions such as “Treatment A worked better than *B* for the patients in the specific hospital studied during the particular time interval of the study,” one can draw more useful and general conclusions such as “Treatment A works better than treatment *B*”—or, if qualification is needed, “Treatment A works better than treatment *B* for people of type *Z*.” Indeed, a field of pragmatic RCTs has been developing to meet this need, since many results from traditional RCTs have turned out not to generalize well beyond the specific populations and circumstances studied (Patsopoulos 2011).

How to generalize correctly from particular study results, including the results of particular RCTs, pragmatic RCTs, or field trials, to arrive at valid general causal laws and conclusions has long troubled philosophers of science. It is a version of the notorious problem of induction (Cartwright 2003). One constructive partial answer makes use of the invariance property of causal laws previously discussed in the context of the Lucas critique of the use of empirical macroeconomic models to predict effects of policy interventions (Hoover 2014). Suppose that *Y* is a response or outcome variable of interest; *X* is a decision variable indicating a policy, intervention, or decision that affects the probability distribution of values for *Y*; and *Z* is a vector of covariates that also affect the response. A causal graph (DAG model)

succinctly indicates these dependencies as $X \rightarrow Y \leftarrow Z$, where the conditional probability table (CPT) for Y specifies the conditional probability of each of its possible values as a function of the values of its parents. In notation, this CPT may be denoted by $P(y | x, z)$ to signify the conditional probability that Y takes value y , given that $X = x$ and that $Z = z$. If the CPT represents a universal causal law, with all the causal parents of Y included in X and Z , then $P(y | x, z)$ must be the same in all settings. With sufficient data, this *homogeneity* or *invariance* implication can be tested statistically in several different ways, e.g., by using statistical tests for homogeneity, latent variables, or mixture distributions. Discovering that the same CPT (or, more generally, the same model) holds in a wide diversity of particular data sets provides a possible basis for extrapolating it to new situations, and thus offers a solution to the problem of inductive inference: the invariant law or CPT becomes the generalization that is learned from particular instances. In short, one basis for modern causal discovery algorithms is to use data from a wide range of particular experiments or studies, possibly involving a wide range of different interventions and conditions, to seek causal laws, often expressed as structural equations or CPTs that are invariant across the diverse studies (Peters et al. 2016). These invariant laws provide the sought-for generalization of the particular evidence from which they are derived. They can be used to derive straight-forward adjustments, or transport formulas, for generalizing (or “transporting”) causal inferences from one set of conditions and interventions to another (Bareinboim and Pearl 2013; Lee and Honavar 2013). This is simpler than it may sound: just as a validated simulation model for a system can be applied to new input scenarios to discover (via simulation) what outputs they would be likely to produce under changed conditions, so networks of CPTs, i.e., causal models, can be applied to new input conditions to calculate corresponding output probabilities. Software for deriving and applying transport formulas is starting to become available, e.g., via the *causaleffect* R package at <https://cran.r-project.org/web/packages/causaleffect/causaleffect.pdf>.

Example: Invariant CPTs, Generalization, and Transportability of Causal Laws

The following example illustrates the arithmetic of how to generalize results from learning or experimental settings in which a probabilistic causal law has been discovered to a different target setting in which it is to be applied. Suppose that large RCTs have been conducted at three hospitals, A, B, and C, to test the success rate for a new treatment compared to an old one. The treatment is for a disease that is never cured without treatment and the old treatment has a 50% success rate in curing people; this rate holds in all hospitals and for all types of patients. By contrast, less is known about the new treatment, but in RCTs at hospitals, A, B, and C, it has been found to have cure rates of 0.82, 0.55, and 0.37, respectively. (For simplicity, we assume that the RCTs are so large that these rates can be treated as accurate, without having to worry about sampling error.) An initial causal DAG model summarizing this very incomplete knowledge of the new Treatment’s effectiveness is as follows:

$$Treatment \rightarrow Cure \leftarrow Hospital$$

The outcome variable *Cure* is coded for each individual so that 1 = successful cure, 0 = not successful cure. Based on the preceding description, the CPT for *Cure* is as follows:

$$\begin{aligned} P(Cure = 1 | Treatment = Old, Hospital = A) &= 0.5 \\ P(Cure = 1 | Treatment = New, Hospital = A) &= 0.82 \\ P(Cure = 1 | Treatment = Old, Hospital = B) &= 0.5 \\ P(Cure = 1 | Treatment = New, Hospital = B) &= 0.55 \\ P(Cure = 1 | Treatment = Old, Hospital = C) &= 0.5 \\ P(Cure = 1 | Treatment = New, Hospital = C) &= 0.37 \end{aligned}$$

These equations use the usual notation of conditional probability. The first one, for example, states that the conditional probability that a randomly selected individual from hospital A who received the old treatment will be cured (have *Cure* = 1) is 0.5. (Since this probability does not depend on the hospital in which the old treatment is administered, we could combine the first, third, and fifth of the above equations into $P(Cure = 1 | Treatment = Old) = 0.5$, which would be a more efficient way to summarize the same information.) The corresponding conditional probabilities for *Cure* = 0 are found by subtracting from 1 these conditional probabilities that *Cure* = 1. Together these 12 conditional probabilities comprise the CPT for *Cure* for the three hospitals studied so far. However, this does not help to predict what will happen if the new treatment is tried in a new hospital, D. The new hospital would represent a new, as-yet unobserved value for the *Hospital* variable. How to extend the current CPT to handle new values of its variables cannot be determined from the data already collected. This is the challenge of generalizing from particular study results to generally applicable findings.

What the data on hospitals A–C do show is that something is missing from the DAG model. That success probabilities for the new treatment differ significantly across hospitals A–C invites the question of *why* they are different—what factors differ across populations in different hospitals that can explain the difference in success rates? As long as a difference remains, this question can always be asked. Only when enough causal parents have been included so that the conditional probabilities for treatment success, given the values of the causal parents, are the same regardless of location does the question no longer arise. This is why invariance is so useful for causal discovery: an adequate causal model must contain the information needed to explain systematic differences in outcomes in terms of invariant conditional probabilities. If it cannot do so, then the remaining unexplained heterogeneity in CPTs limits ability to predict outcome probabilities from the factors that are included in the model.

To help find an invariant causal law, suppose that data are collected on the individuals who participated in the RCTs in hospitals A–C, such as their ages, sexes, ethnicities, medical histories, and so forth. For purposes of a simple illustration, let's assume that only one of these variables turns out to be useful for predicting

the value of *Cure*: the sex of the patient. Chapter 2 discusses predictive analytics algorithms, including Classification and Regression Trees (CART) and Random Forest algorithms, that are widely applied in machine learning and causal discovery algorithms to determine which variables are informative about, and hence help to predict, the values of an outcome variable of interest.

Specifically, for this example, suppose that the percentages of male patients in the RCTs were 80% at hospital A, 50% at hospital B, and 30% at hospital C. If the sex of the patient is indeed the only variable on which the success of the new treatment depends, then the overall success rate for the new treatment at a hospital can be described by the following structural equation model (SEM):

$$E(Cure) = a^*male_fraction + b^*female_fraction$$

where a = probability that *Cure* = 1 for men and b = probability that *Cure* = 1 for women. The data from hospitals A and B give the following two equations:

$$\begin{aligned} 0.82 &= a^*0.80 + b^*0.20 \\ 0.55 &= a^*0.50 + b^*0.50. \end{aligned}$$

These can be solved either manually or using an on-line solver such as the one at http://wims.unice.fr/wims/en_tool~linear~linsolver.en.html to deduce that $a = 1$ and $b = 0.1$. We have thus used the data from hospitals A and B to estimate the following SEM:

$$\begin{aligned} E(Cure) &= male_fraction + 0.1^*female_fraction \\ &= male_fraction + 0.1^*(1 - male_fraction), \\ &= 0.1 + 0.9^*male_fraction. \end{aligned}$$

If this SEM is correct, then it can be applied to any hospital, since a causal law holds universally once the dependence of outcomes on their causal parents has been correctly specified. For example, to validate this model, it can be used to predict the value of *Cure* for hospital C:

$$E(Cure) = 0.1 + 0.9^*0.30 = 0.37.$$

This agrees with, and explains, the data value from the RCT for hospital C. Such agreement does not prove with logical certainty that the SEM is correct, but it adds credibility insofar as it is unlikely that the predicted value of 0.37 would agree by chance with the observed value.

The invariant law $E(Cure) = 0.1 + 0.9^*male_fraction$ that we have now discovered not only describes and explains the different RCT results for hospitals A-C, but also it can be used to predict the success rate of the new treatment in any similar future RCTs that might be carried out at other hospitals. The refined DAG model for *Cure* in any hospital is as follows:

$$Treatment \rightarrow Cure \leftarrow male_fraction \leftarrow Hospital,$$

where the CPT for *Cure* is

$$\begin{aligned} P(\text{Cure} = 1 | \text{Treatment} = \text{old}) &= 0.5 \\ P(\text{Cure} = 1 | \text{Treatment} = \text{new}) &= 0.1 + 0.9 * \text{male_fraction}. \end{aligned}$$

This is the appropriate generalization of the particular findings from these RCTs. More importantly, the refined understanding of the causal parents of *Cure* leads to the following DAG model for individual patients:

$$\text{Treatment} \rightarrow \text{Cure} \leftarrow \text{Sex}$$

with CPT

$$\begin{aligned} P(\text{Cure} = 1 | \text{Treatment} = \text{old}, \text{Sex} = \text{male}) &= 0.5 \\ P(\text{Cure} = 1 | \text{Treatment} = \text{old}, \text{Sex} = \text{female}) &= 0.5 \\ P(\text{Cure} = 1 | \text{Treatment} = \text{new}, \text{Sex} = \text{male}) &= 1 \\ P(\text{Cure} = 1 | \text{Treatment} = \text{new}, \text{Sex} = \text{female}) &= 0.1. \end{aligned}$$

This CPT makes it clear that the best treatment strategy at the level of individual patients, rather than hospital populations of patients, is to prescribe the old treatment to women and the new treatment to men.

This example was simplified for ease of exposition, but the following points hold more generally. First, significant differences in conditional probabilities of different outcomes across different RCT locations, studies or settings, even after conditioning on the values of known factors (e.g., which treatment is administered), indicate that a causal model is incomplete: other causal factors remain to be discovered that affect the outcome. Conversely, a causal law can typically be represented by conditional probabilities of outcome values that are always the same, given the values of the parents of the outcome in a causal model (e.g., a DAG or a set of structural equations). Such an invariant conditional probability table or function provides a generalization of particular instances. For purposes of statistical analysis, it is useful to recognize that causal relationships among variables typically constrain observed data points to lie in a small subset of the volume of points that they could occupy if their values were independent of each other. For example, the data points may lie near a line, curve, or surface (in mathematical terms, a low-dimensional manifold within the higher-dimensional space of the data points) determined by the structural equations or CPTs expressing dependencies among the values of different variables. In the example, this invariant manifold—the same for all data points—was the line $E(\text{Cure}) = 0.1 + 0.9 * \text{male_fraction}$. Measurement error, which was ignored in this example, usually makes it necessary to use techniques such as regression to estimate this underlying relationship instead of using the data points to solve for it exactly. Such estimation can be carried out using specialized software programs such as the Invariant Causal Prediction (ICP) package in R (Peters et al. 2016) and its extensions to allow for nonlinear and nonparametric dependencies among variables (Heinze-Deml et al. 2017). However, even with such sophisticated algorithms, a causal DAG model or set of structural equations can still usually be estimated from a subset of the available data points and then validated on one or more different, disjoint subsets. It is standard terminology in much of the machine learning literature to call the data

points used to learn or estimate a model the *training set* and the data points used to validate it the *test set*. (In the example, with no measurement error, RCT data from any two of the three hospitals A, B, and C could be used to solve for $a = 1$, $b = 0.1$, and hence for the underlying linear structural equation model $E(\text{Cure}) = 0.1 + 0.9 * \text{male_fraction}$. This model could then be validated by checking that it also described data from the RCT trial at the third hospital.) The basis of the validation is that it is unlikely that the points in a test set will lie on or near the manifold estimated from the training set by chance alone. Chapter 2 discusses machine learning algorithms for learning and validating causal models from data.

Quasi-Experiments (QEs) and Intervention Time Series Analysis Are Widely Used to Evaluate Impacts Causes by Interventions

In practice, it is often unethical or impractical to carry out RCTs, including pragmatic RCTs. Other ways are then needed to use data to evaluate the effects caused by interventions. Since the 1960s, one popular way to evaluate the impacts of social programs and policies has been to use *quasi-experiments* (QEs) (Campbell and Stanley 1963; White and Sabarwei 2014). These are studies without randomized assignments of units to treatments that compare outcomes in the treated and untreated groups. For example, one common QE design compares (a) changes in outcome measures from before to after the implementation of an intervention in a group or population that the intervention affects to (b) corresponding changes in a comparison group or population that it does not affect. More recent “difference-in-differences” methods in epidemiology are based on the same idea.

Alternatively, the affected population can be used as its own control. For example, *interrupted time series analysis*, also known as intervention time series analysis (ITSA) or simply *intervention analysis*, tests whether the best-fitting time series model describing the data prior to an intervention differs from the best-fitting time series model after the intervention. If so, it attributes the difference to the intervention (Gilmour et al. 2006). Similarly, a recent program from Google called *CausalImpact* uses data from control time series not affected by an intervention to try to forecast how a variable that the intervention does affect would have evolved in the absence of the intervention; differences between observed and forecast values are attributed to the intervention (<https://google.github.io/CausalImpact/CausalImpact.html>). For example, to estimate the causal impact of an advertising campaign on daily clicks at a web site, the number of clicks expected without the advertising campaign might be forecast from data on other web sites in markets not affected by the campaign. Then the difference in clicks per day between the observed values and these forecast values can be attributed to the advertising campaign, assuming that no other cause can be identified.

However, the design and analysis of QE comparisons and the interpretation of causal attributions and effects estimates based on the data they produce require considerable care. Because QEs do not randomly assign individuals to treatment (intervention) and comparison groups, there is no rigorous way to make sure that the effects they estimate are actually *caused* by the intervention instead of merely *coincident* with it. For example, the average differences in responses between the treatment and control groups, defined as those affected by the intervention or policy being evaluated and those not affected by it, might be explained by unmeasured differences between these groups in the distributions of covariates that also affect the response. Other possible *threats to internal validity*, in the terminology of Campbell and Stanley (1963), meaning possible non-causal explanations for observed differences between treatment and control groups in a QE study, include the following:

- *History*: Other events may coincidentally affect responses and cause changes and differences between treatment and control groups following an intervention even if the intervention itself had no effect, or a lesser effect than the observations suggest.
- *Maturation*: Treated individuals get older between the time before an intervention and the time after it. In studies that compare pre-intervention and post-intervention results, using individuals as their own controls, maturation rather than treatment may explain differences over time.
- *Regression to the mean*: Suppose that interventions are assigned to individuals, locations, or groups that appear to be most in need of them due to extreme values of some variables taken as indicators of need. Measurements taken later may show less extreme values simply because extreme values are less likely than less extreme values. Students selected for a tutoring program because of extremely poor performance on a standardized test, for example, might be expected to do better next time even if the program had no effect.
- Awareness of being studied, increased familiarity with survey instruments, investigator biases, biases in missing data or in attrition from the sample, and other non-treatment sources of differences or changes in responses between treated and control groups.

Unless they are carefully tested and refuted using data, such potential alternative explanations threaten the validity of causal interpretations of observed differences between treatment and control groups within a QE study. They are called threats to *internal validity* because they address causal inferences for the studied populations, i.e., within the scope of the study. As previously discussed, there are also threats to *external validity*, i.e., to ability to generalize any causal conclusions beyond the specific populations and circumstances of the QE study. For example, the studied treatment and comparison groups may not be representative of target populations of interest, or study conditions that affected the observed outcomes might not hold elsewhere. As discussed in Chap. 2, causal analytics methods address these challenges by using models that don't rely on the key assumption that observed differences in outcomes between treatment and control groups are caused by the treatment or intervention being evaluated. These models describe more explicitly how changes

in some variables propagate through specific causal mechanisms (represented by conditional probability tables, structural equations, or other validated causal models) to affect probability distributions of other variables.

Ironically, although Campbell and Stanley (1963) were concerned largely with warning against the dangers of using QEs for causal inference, QEs are now increasingly widely used for that purpose. In general, doing so requires strong assumptions, such as that all relevant causes have been measured, that there are no unmeasured confounders, or that observed associations are causal. If these modeling assumptions are violated, then causal inferences drawn from the QE may be mistaken.

Example: Did Banning Coal Burning in Dublin Reduce Mortality Rates?

A famous study in Dublin County, Ireland of the effects on public health of a ban on coal burning reported that both all-cause mortality rates and cardiovascular mortality rate specifically declined substantially from the 6 years before the ban to the 6 years following it (Clancy et al. 2002). This study contributed to policy decisions to extend coal-burning bans in Ireland based on a belief that cleaner air had been found to cause reduced mortality. It was estimated that the ban saved thousands of lives per year, based on an assumption that changes in health risks *following* the intervention were *caused by* it. This causal assumption was tested a decade later by some of the original authors in an updated study that compared mortality rates in areas of Ireland affected and not affected by the bans (Dockery et al. 2013). The updated study concluded that the bans had produced no detectable reductions in either total or cardiovascular mortality rates (Dockery et al. 2013). As explained by Zigler and Dominici (2014), “However, even when studying an abrupt action, threats to causal validity can arise, as illustrated in extended analyses of the Dublin coal ban that revealed that long-term trends in cardiovascular health spanning implementation of the ban—not the coal ban itself—contributed to apparent effects on cardiovascular mortality.” Social statisticians since the 1960s have cited the “one-group pretest-posttest design” used in the original 2002 study as inappropriate for causal inferences, since it leaves uncontrolled the threat of coincident historical change, as well as other threats to valid causal inference (Campbell and Stanley 1963, p. 7). By contrast, a pretest-posttest control group design such as that in the 2013 follow-up study can show that a large reduction in particulate pollution had no detectable effect on total mortality, as in Dublin, if that is the case; or it can provide strong evidence that high pollution levels cause excess mortalities if mortality rate spikes where and when air pollution spikes—such as in London in 1952—but not otherwise. Interestingly, although it has been known since at least 2013 that the bans had no apparent effects on total mortality rates, they are still cited by Irish regulators and policy makers as having created very substantial reductions in total mortality. As of this writing in late 2017, the Department of Communications, Climate Action and Environment web page (<https://www.dccae.gov.ie/en-ie/environment/topics/air-quality/smoky-coal-ban/Pages/default.aspx>) still states that “The smoky coal ban

allowed significant falls in respiratory problems and premature deaths from the effects of burning smoky coal. The original ban in Dublin is cited widely as a successful policy intervention and has become something of an icon of best practice within the international clean air community. It is estimated that in the region of 8,000 lives have been saved in Dublin since the introduction of the smoky coal ban back in 1990. Further health, environmental and economic benefits (estimated at €53m per year) will be realised, if the ban is extended nationwide. We intend to extend the health and environmental benefits of the ban on smoky coal, currently in place in our cities and large towns, to the entire country.” Similar disinformation continues to be spread in media accounts of the economic and life-saving benefits of the bans. Thus the fact that the bans had no detectable benefits in reducing all-cause mortality when properly evaluated using control groups (Dockery et al. 2013; Zigler and Dominici 2014) appears to have had no impact on the political and media narratives used to justify spreading them more widely.

Many technical efforts have been made since the 1970s to try to obtain the benefits of RCTs from QE data. One approach is to try to reduce—or, if possible, eliminate—known differences between treatment and comparison groups by carefully matching their individual members using measured variables. Another is to try to adjust statistically for the assumed effects of differences between treatment and control groups using assumed models and sensitivity analyses for their effects. A third is to exploit “natural experiments” in which unplanned events are considered to affect some people but not others as if by chance. For example, if a labor strike shutters a factory that had been discharging smoke into the air, then comparison of what happens next to the health of people downwind from the factory compared to others upwind or far removed from it might reveal something about the health effects of cleaner air—but only if the affected population and the comparison group are similar in relevant (health-affecting) ways, apart from the change in exposure. Similarly, arbitrary thresholds, such as the age at which a behavior such as driving, drinking, smoking becomes legal, together with an assumption that people a little on either side of the threshold are otherwise similar to each other, allow differences in outcomes (such as car accident rates) to be estimated and attributed to the difference in permitted behaviors.

Counterfactual and Potential Outcome Framework: Guessing What Might Have Been

QE design and data analysis strategies have given rise to a variety of statistical techniques for estimating population-level causal effects from QE data. A common underlying philosophical framework, the *counterfactual/potential outcome framework*—is widely used to interpret their results (Höfler 2005). This holds that causal effects in populations can be estimated by comparing estimates of what *did* happen, e.g., real observed values of illness or death rates, to estimates of what *would have* happened had causes been different. Since what would have happened under different counterfactual conditions is never observed, and since the reasons for the

hypothetical counterfactual conditions are seldom specified in sufficient detail to predict outcomes uniquely, potential outcome methods depend heavily on the use of statistical models and assumptions to predict what would have happened. This makes their estimates of causal impacts dependent on modeling assumptions. Different choices of modeling assumptions can produce very different estimates of causal impacts.

Methods developed within the potential outcome framework include propensity score matching (PSM), marginal structural models (MSMs), instrumental variables (IVs), regression discontinuity designs (RDDs), and related methods (Höfler 2005; White and Sabarwei 2014). Although they are sometimes described as yielding rigorous estimates of average causal effects in populations from QE data, potential outcome methods depend on making strong modeling assumptions whose validity is seldom known in practice. Examples include the assumptions that are that there are no unobserved confounders, that all relevant factors have been observed, and that all types of individuals have received all treatments. In effect, by assuming that average differences estimated from data are causal rather than being coincidental or explained by something else, potential outcome methods do arrive at causal conclusions, but at the price of reliability. Their causal conclusions may well be wrong, and different investigators may reach contradictory conclusions, starting from the same data, by choosing different counterfactual modeling assumptions.

Example: What Were the Effects of a Public Smoking Ban Policy Intervention on Heart Attack Risks?

To illustrate the promise and pitfalls of QEs, consider the two plots in Fig. 1.8. Each one compares estimated mean monthly incidence rates of heart attacks (acute myocardial infarction, AMI) among adults 30–64 years old in Tuscany, Italy in the 5 years before a January, 2005 ban on public smoking and in the year following the ban (shaded area to the right). Estimates with seasonal effects included (solid curves) show that, as in other studies, heart attack risks are greatest in the cold winter months, especially December and January, and are lowest in the hot summer months. The dashed curves show estimates of the baseline incidence rate of heart attacks with the seasonal changes subtracted out and with a linear trend line (upper plot) or a nonlinear trend curve (lower plot) fit to the data points. (These curves were generated using Poisson regression modeling, since the data show counts of AMI cases, but the main points do not depend on this specific modeling technique.) In the upper plot, there is a clear decrease in estimated baseline incidence rates of AMI from before the ban to after it, consistent with a causal hypothesis that the public smoking bans caused a prompt decrease in heart attack risks. The estimated size of the decrease is 5.4%. In the lower plot, however, with a nonlinear trend fit to the same data, there is no significant decrease (and, in fact, a slight increase) in the estimated AMI incidence rate from before to after the ban. This illustrates how the size and direction of the effect of the ban estimated from these data depend on the model selected to analyze and interpret the data.

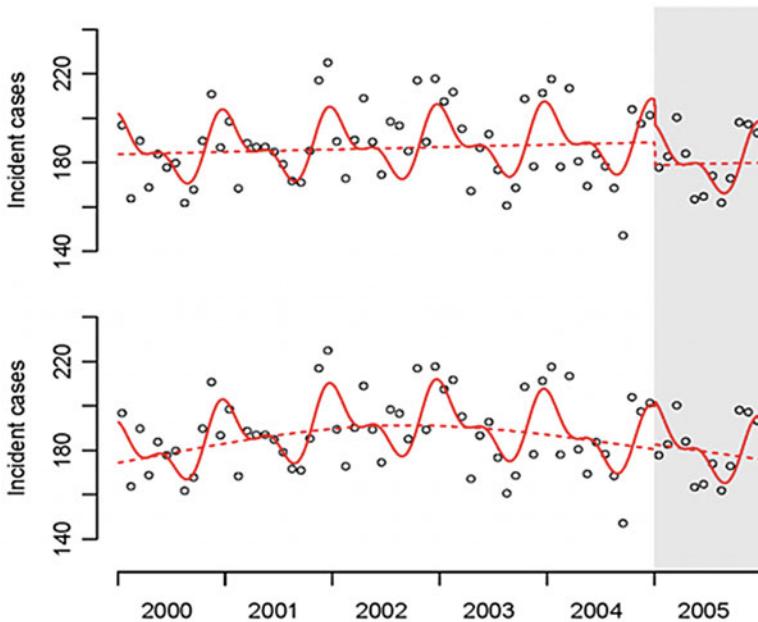


Fig. 1.8 Evaluation results depend on modeling assumptions. The upper and lower panels fit linear and nonlinear models, respectively, to time series data on heart attack (acute myocardial infarction, AMI) incidence in Tuscany, Italy, before and after a January, 2005 ban on smoking in public. *Source:* Gasparrini et al. (2009)

Change Point Analysis (CPA) and Sequential Detection Algorithms

In general, the changes in outcomes caused by policy interventions can be difficult or impossible to determine from time series data alone. Most real-world time series of interest, from economic indicators to disease or mortality counts, result from many factors, not all of which are necessarily known or measured. Hence the statistical characteristics of such time series, including their mean levels and variance around the mean levels, can shift frequently and for unknown reasons. For a volatile, non-stationary time series, comparing values from before an intervention to values after it can be very likely to show statistically significant differences even if the differences were not caused by the intervention.

A partial solution to this challenge is provided by *change point analysis* (CPA) and *sequential detection* algorithms (James and Matteson 2014; Ross 2015; Tartakovsky et al. 2014). CPA algorithms test whether the statistical characteristics of an observed time series changed significantly at one or more points during an interval of observation and, if so, estimate when these “change points” occurred. Some CPA algorithms also estimate the sizes of changes, assuming that they have simple forms such as a step up or down in mean values. Sequential detection

algorithms are similar to CPA algorithms but are applied *as observations come in* to detect a change as soon as possible after it occurs, rather than being applied retrospectively applied in batch mode to look back over an interval and identify changes after the fact. In practice, sequential detection algorithms can be used to trigger alerts or raise alarms (with statistical confidence levels attached) as soon sampling indicates that a process being monitored has changed for the worse; and CPA algorithms can be used later to assess how accurate the warnings probably were and to fine-tune decision thresholds to minimize the sum of costs from false positives and false negatives.

Example: Change-Point Analysis (CPA) Clarifies When and Whether Events Happened

The following example is adapted from Bier and Cox (2017). Since 2001, when a letter containing anthrax led to 5 deaths and 17 other infections from which the victims recovered, the U S Environmental Protection Agency (EPA), the Centers for Disease Control and Prevention (CDC), and the Department of Health and Services have invested over a billion dollars to develop surveillance methods and prevention and preparedness measures to help reduce or mitigate the consequences of bioterrorism attacks should they occur again in future (Grundmann 2014). In practice, detecting a significant upsurge in hospital admissions with similar symptoms may often be the most practical and reliable way to identify that a bioterrorism attack is in progress. The statistical challenge of detecting such changes against the background of normal variability in hospital admissions has motivated the development of methods that seek to reduce the time to detect attacks when they occur, while keeping the rates of false positives acceptably small (Cami et al. 2009; Shen and Cooper 2010).

Such statistical data analysis and pattern detection, carried out in settings for which the patterns for which one is searching are well understood (e.g., a jump in hospitalization rates for patients with similar symptoms that could be caused by a biological agent) and where enough surveillance data are available to quantify background rates and to monitor changes over time, illustrate the types of uncertainty for which excellent, sophisticated techniques are currently available. Figure 1.9 presents a hypothetical example showing weekly counts of hospital admissions with a specified symptomology for a city. Given such surveillance data, the risk assessment inference task is to determine whether the hospitalization rate increased at some point on time (suggestive of an attack), and, if so, when and by how much. Intuitively, it appears that counts are somewhat greater on the right side of Fig. 1.9 than the left, but might this just be due to chance, or is it evidence for a real increase in hospitalization rates?

Figure 1.10 shows the output from a typical statistical algorithm (or computational intelligence, computational Bayesian, machine learning, pattern recognition, data mining, etc. system) for solving such problems by using statistical evidence together with risk models to draw inferences about what is probably

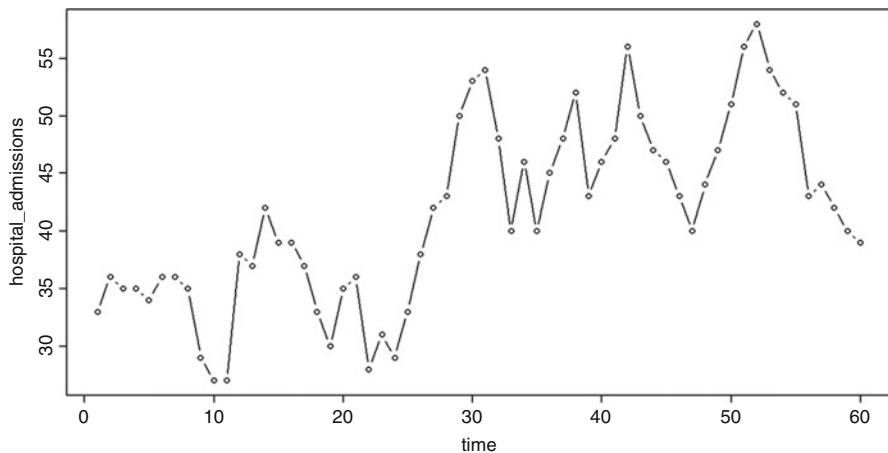


Fig. 1.9 Surveillance time series showing a possible increase in hospitalization rates

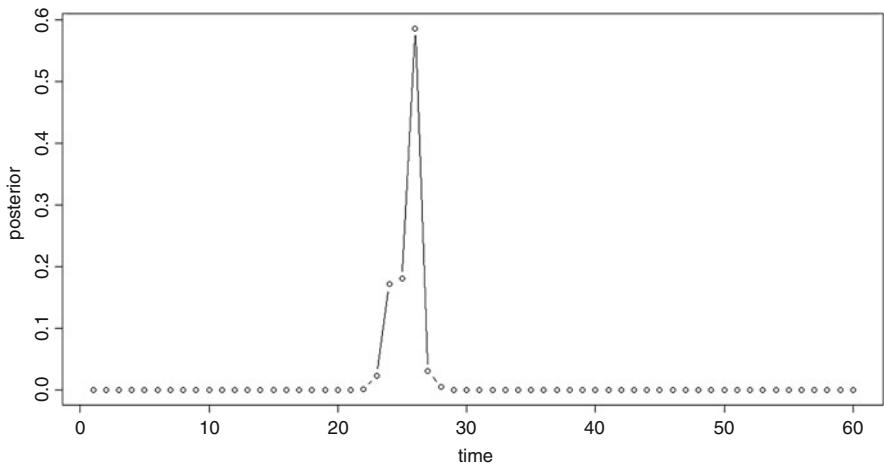


Fig. 1.10 Bayesian posterior distribution for the timing of the increase (change point), if one occurred

happening in the real world from observed data. The main idea is simple: the peak at week 26 indicates the time that is computed to be most likely for when an attack occurred, based on the data in Fig. 1.9. The heights of the points in Fig. 1.10 indicate the probabilities that an attack occurred at different times, if it occurred at all.

The same algorithm that produces this information, described next, also estimates how admission rates increased from before to after the attack. The algorithm identifies the time of the attack (week 26), and estimates the magnitude of its effect (not shown in Fig. 1.10).

The algorithm used in this case works as follows. In Figs. 1.9 and 1.10, the horizontal axes show the possible times, in weeks, at which an attack might have occurred that increased the hospital admission rate from its original level (the baseline or background level) to a new, higher level. For simplicity, we assume that such a one-time, lasting increase in hospitalization rates is known to be the specific form that the observable effect of an attack would have. (More elaborate models would allow for transient effects, multiple waves of attacks, and other complexities, but the highly simplified model of a one-time jump from a lower to a higher level suffices to illustrate key points about change-point detection methods.) On the vertical axis of Fig. 1.10 are scaled versions of the *likelihoods* (discussed next) of the data in Fig. 1.9 for each of 60 different hypotheses, each specifying a week when the attack is hypothesized to have occurred. The likelihoods are rescaled so that they sum to 1; this lets them be interpreted as Bayesian posterior probabilities for the attack time, assuming a uniform (flat) prior. Thus, an algorithm that computes likelihoods (i.e., probabilities of the data, given assumed values for the attack week and other uncertain quantities) also allows the most likely values for these quantities to be inferred from the data.

The likelihoods, in turn, are computed from a *likelihood function*. This gives the probability of the observed data (here, the data on hospital admission counts for all 60 weeks in Fig. 1.9) for each value of the uncertain quantity (or quantities) about which inferences are to be drawn (here, the week of attack, the admission rate prior to the attack, and the admission rate following the attack). The unobserved quantities that affect the joint probability distribution of the observed quantities (i.e., the data) are often referred to generically as the *state* of the system being studied. In this example, the state consists of the week of the attack, the admission rate prior to the attack, and the admission rate following the attack. Symbolically, the likelihood function can be written as $P(\text{data} \mid \text{state})$, denoting the conditional probability of the observed data given the values of the unobserved quantities that they depend on. The likelihood of the data in Fig. 1.9, given the hypothesis that an attack occurred in any particular week, is just the product of the probabilities of the observed numbers of hospital admissions in all of the weeks (i.e., the data in Fig. 1.9), under that hypothesis. It can be calculated by modeling the number of admissions per week as a random variable having a binomial distribution with a mean equal to the product of the number of susceptible people in the community served by the hospital and the admission probability per person per day, which jumps from a baseline value to an increased value when and if an attack occurs. The numerical values of the time and the values of the pre-attack and post-attack admission rates that jointly maximize the likelihood of the data in Fig. 1.9 constitute the *maximum-likelihood estimates* (MLEs) for the attack time and the admission rates before and after the attack. Figure 1.10 shows that the MLE for the attack time—the change-point in the time series in Fig. 1.9—is 26 weeks. If the MLEs for admission rates are 0.1 before the attack time and 0.2 after it, then the MLE for the size of the jump in admission rates caused by the attack would be $0.2 - 0.1 = 0.1$ admissions per person per week.

This example has illustrated how MLE algorithms and modeling assumptions can be used to estimate the time of a change point and the magnitude of the change that

occurred then. The key points are that computational methods are well able to estimate these important unknown quantities from data when: (1) The form of the change to be detected is known (e.g., in this example, the effect of an attack is known to be a one-time increase in admission rates, so that the pattern to look for is known); (2) Plentiful surveillance data are available (which allowed MLE estimates of admission rates to be formed that were highly accurate); and (3) The effect size is large enough to show clearly through the random noise in the surveillance count time series (as indicated by the high peak in Fig. 1.10, which can be shown via simulation to be many orders of magnitude greater than the peaks that occur by chance under the null hypothesis of no real change in admission rates). MLE algorithms detect change points quite quickly (within 1–2 weeks with high confidence in this example) under these conditions. However, their success depends on how well conditions 1–3 satisfied. Modern methods of CPA allow these conditions to be relaxed, as discussed next.

Modern algorithms for CPA and sequential detection, available in free R packages (James and Matteson 2014; Ross 2015), can be applied to multivariate time series, i.e., to series in which more than one quantity is monitored or observed over time. Several of these algorithms apply non-parametric statistical tests to test whether data permit rejection of the null hypothesis that the underlying data-generating process has not changed. These are useful advances over older methods, including maximum likelihood estimation (MLE) methods, that applied only to univariate time series and that had to assume specific parametric models or conditional probability distributions for the observations, such as that they had normal, binomial, or Poisson distributions with means that might have undergone a jump at some time. Modern CPA and sequential detection algorithms using non-parametric tests improve on the situation in Fig. 1.8 by providing more reliable, model-independent answers to the question of whether and when the underlying data-generating process has changed. Analyses such as those in Fig. 1.8 only address whether a selected model estimates different values for the means of observations before and after a user-specified time. Even if the answer is yes, it may simply reflect effects of model specification errors: the selected model may provide different fits to the data before and after some user-specified point without describing either very well, and the difference in fits may simply be an artifact of the model selected to describe it. Non-parametric CPM and sequential detection algorithms avoid these difficulties by making it unnecessary to specify a parametric family of models and by using the data to estimate the times of change points rather than requiring the user to specify them.

Despite these advances, even the best algorithms for change-point analysis, sequential detection, and intervention analysis only address whether and when changes occur (and perhaps their sizes), and not *why* they occur. They are thus suitable for descriptive analytics, describing what changed when, but not for evaluation analytics assessing the causal impact of decisions or interventions on changes in outcomes.

A Causal Modeling Perspective on Evaluating Impacts of Interventions Using CPTs

Although they have been widely applied in efforts to evaluate the effects of social, environmental, economic, and public health initiatives, none of the evaluation methods discussed so far provides a reliable way to figure out by how much an action or intervention has changed an outcome of interest, even though this is the main goal of methods of evaluation analytics. When a valid causal DAG model is known for the system or situation being analyzed, however, calculating the effects of actions becomes straightforward (Ortega and Braun 2014). Consider the following DAG model of a decision:

$$act \rightarrow outcome \leftarrow state.$$

Here, *act* represents a decision variable summarizing the choices, controls, policies, courses of action, or interventions whose effects on the outcome random variable are to be evaluated; and *state* is a random variable summarizing all of the other factors that, together with *act*, determine the conditional probability of the *outcome* variable. If *act* and *state* both have only a few possible values or levels, then the conditional probability table (CPT) for *outcome* in this model can be displayed as an array of probability numbers with generic element $P(o|a, s)$, indicating the probability that the outcome value is *o* if the act chosen is *a* and if the state is *s*. The marginal probability table for the input variable *state* consists of values of $P(s)$, the probability that *state* has level or value *s*, for each of its possible values. Now, suppose that the decision-maker intervenes to choose the value for *act* to be, say, *a'* instead of some other default or status quo value *a*. This choice changes the predictive probability distribution of the outcome from the old distribution given by

$$P(o|a) = \sum_s P(o|a, s)P(s)$$

to a new distribution given by

$$P(o|a') = \sum_s P(o|a', s)P(s)$$

The effect on the outcome variable caused by the decision to set $act = a'$ is this change in the probability distribution of its values. The data elements needed to compute it are the contents of the marginal probability table or marginal probability distribution for the *state* input, $P(s)$, and the CPT elements $P(o|a, s)$ and $P(o|a', s)$.

More generally, suppose that the DAG model lets probabilities for values of the state variables depend on the intervention, like this:

$$act \rightarrow outcome \leftarrow state.$$

Then changing the act from a to a' changes the probability distribution for the outcome from

$$P(o|a) = \sum_s P(o|a, s)P(s|a)$$

to the following new distribution:

$$P(o|a') = \sum_s P(o|a', s)P(s|a').$$

Again, the CPTs in the DAG model provide all the information needed to compute the effect of the intervention, which can be defined as the change that it causes in outcome probabilities, i.e., the shift from $P(o|a)$ to $P(o|a')$. If the DAG model applies to each individual in a population and the outcome variable is numerical (e.g., measured on a ratio scale or an interval scale), then a popular population-level measure of the effect of the intervention is the change in the average value of the outcome over individuals in the population, $E(o|a') - E(o|a)$, where E denotes the expected value operator, $E(o|a) = \sum_o P(o|a)o = \sum_0 \sum_s P(o|a, s)P(s|a)o$.

Example: Calculating Causal Impacts of an Intervention

Returning to the earlier RCT example with the DAG

$$Treatment \rightarrow Cure \leftarrow Sex,$$

the CPT for the outcome variable, $Cure$, can be written as follows:

Outcome, o	Act, a	State, s	Conditional probability, $P(o a, s)$
Cure	Old treatment	Male	0.5
Cure	Old treatment	Female	0.5
Cure	New treatment	Male	1
Cure	New treatment	Female	0.1
No cure	Old treatment	Male	0.5
No cure	Old treatment	Female	0.5
No cure	New treatment	Male	0
No cure	New treatment	Female	0.9

The bottom half of this table is redundant since the conditional probabilities for the outcomes *cure* and *no cure* must sum to 1 for each set of input conditions, but the values for no cure are included in the CPT for completeness.

Problem: Calculate the population-level effect of an intervention that changes from treating all patients with the old treatment to treating all patients with the new treatment at a hospital in which half the patients are men and half our women. Also calculate the individual-level effects for men and for women.

Solution: The population-level effect can be calculated as follows. Define the outcome variable to have value $o = 1$ for a patient who is cured and $o = 0$ otherwise. Then the change in the fraction of patients cured caused by the intervention is:

$$\begin{aligned}
 E(o|a') - E(o|a) &= E(o|new\ treatment) - E(o|old\ treatment) \\
 &= \sum_o \sum_{sex} P(o|new\ treatment, sex)P(sex)o - \sum_o \sum_{sex} P(o|old\ treatment, sex)P(sex)o \\
 &= \sum_{sex} [P(o=1|new\ treatment, sex) - P(o=1|old\ treatment, sex)]P(sex)*1 \\
 &= [P(o=1|new\ treatment, male) - P(o=1|old\ treatment, male)]P(male) + \\
 &\quad [P(o=1|new\ treatment, female) - P(o=1|old\ treatment, female)]P(female) \\
 &= (1 - 0.5)*0.5 + (0.1 - 0.5)*0.5 = 0.25 - 0.20 = 0.05,
 \end{aligned}$$

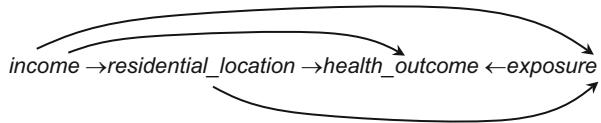
where in the last line, appropriate conditional probability values from the CPT and the assumed marginal probabilities of 50% women and 50% men are substituted for the symbolic expressions. The net result is that changing from the old to the new treatment increases the average cure probability by 5% if the composition of the treated population remains half men and half women.

Rational and informed individuals will care more about individual-level outcome probabilities than about population averages. In this case, the individual-level effects for men and women were averaged in the last line of the above calculation using the fractions of the patient population that were men and women, respectively. The individual-level effects were that the intervention improved the cure rate for men by 0.50, from 0.50 with the old treatment to 1.0 with the new; but it reduced the cure rate for women by -0.40, from 0.50 with the old treatment to 0.10 with the new treatment. Presumably, such an intervention would only be made for the whole population of patients if it were not yet realized that it harmed women. A better individual-level intervention would be to apply the new treatment only to men.

Using Causal Models to Evaluate Total Effects vs. Direct Effects

From the perspective of causal modeling, the best way to use data to estimate the causal impact of an intervention on an outcome is often to estimate and validate a causal model from the data and then use it to quantify how much difference an intervention makes for affected populations and individuals. In general, the effects of policy changes on outcomes of interest can be predicted and evaluated correctly only by modeling the network of causal relationships by which effects exogenous changes propagate among variables. Figure 1.11 illustrates this point via a directed acyclic graph (DAG) model in which *income* has not only a direct causal effect on *health_outcome*, signified by the arrow between them, but also indirect effects transmitted via pathways that include *residential location* and *exposure*.

Fig. 1.11 Direct and indirect paths in a causal DAG model network



Predicting how *health_outcome* probabilities would change if *income* were exogenously changed requires quantifying the effects transmitted via these multiple pathways. This cannot be accomplished by statistical methods that ignore the network of relationships among variables, e.g., by regressing *health_outcome* against *income* and other variables. Such a regression model does not predict how an exogenous change in *income* would change the joint frequency distribution of values for *residential_location* and *exposure* in the affected population. But these changes must be known to predict how the distribution of *health_outcome* in the population will change, since it depends on them. Therefore, methods that go beyond regression modeling are needed to quantify how exogenous changes in one or more variables will affect the joint probability distribution of the remaining variables and, specifically, the distribution of outcomes of interest, such as *health_outcome*.

Causal network analysis methods address this challenge. Moreover, they provide useful distinctions among different kinds of causal effects. For example, the *total causal effect* of a change in *income* on *health_outcome* is the change in the probability distribution of *health_outcome* that occurs when changes in *income* propagate via all pathways, causing changes in *residential_location* and *exposure* that, in turn, contribute to changes in the conditional probability distribution of *health_outcome*. By contrast, the *direct causal effect* of a change in *income* on *health_outcome* is the change in the probability distribution of *health_outcome* that occurs when changes in *income* propagate only via the direct link from *income* to *health_outcome*. Thus, to compute the direct causal effect on health of an intervention that exogenously changes *income*, the CPT for *health_outcome* would be used to calculate the difference in the conditional probability distribution for *health_outcome* made when only the value of *income* changes from its old to its new value. By contrast, the total causal effect of the intervention is found by using the CPT for *health_outcome* to compute how its conditional probability distribution changes when the distributions of all variables affected by the change in *income* also change as it changes, in accord with their own CPTs. Chapter 2 introduces algorithms for evaluating total and direct causal impacts of interventions on outcomes in DAG models using software to facilitate the calculations. In general, no single number, such as a regression coefficient, can represent both the direct and the total effect of a change in an explanatory variable on a dependent variable. Causal network analysis allows such distinctions to be drawn with clarity and provides straightforward ways to define and evaluate both kinds of effects, as well as related concepts such as indirect effects and effects mediated by a specific variable or pathway. Chapter 2 explains these additional concepts and shows how to evaluate quantitatively each type of effect in DAG models. The same ideas and principles carry over to more sophisticated causal models, such as dynamic simulation models.

Using MDPs and DES Causal Models to Evaluate Policies and Interventions

Evaluating the effect of a past action, policy, or intervention on observed outcomes is in general no easier than developing a full causal model for how actions affect outcome probabilities: they are essentially the same problem. Understanding how past actions have affected current outcomes (or their probabilities, if the causal relation is not deterministic) requires essentially the same causal analysis techniques as predicting how current actions will affect future outcomes or their probabilities. When enough knowledge and data are available to develop and validate them, discrete-event simulation (DES) models and Markov decision process (MDP) models can be very useful for evaluating effects of past programs and comparing the effects and the cost-effectiveness of alternative policies, treatments, and interventions. They have been widely used for this purpose in healthcare program evaluations. For example Goehler et al. (2011) survey the results of 27 Markov models, 3 discrete-event simulation models, and 4 mathematical models taken from the literature on evaluation of different interventions to reduce risks of chronic heart failure (CHF). Interventions evaluated by these models included efforts to improve screening, diagnosis, treatment with drugs or devices, disease management programs, and a study of heart transplants. The dynamic simulation models are able to simulate how probably distributions of outcomes over time in a treated population shift in response to different interventions. Together with implementation cost and timing information, this provides the crucial information about the probable consequences caused by different interventions needed for well-informed decision-making based on methods such as cost-effectiveness and risk-cost-benefit analysis.

Simulation-based evaluations of effects of interventions are vulnerable to misuse if the simulations are not based on valid causal models. Comparing probability distributions for outcomes under real and hypothetical conditions can lead to quantified “effects” estimates that cannot be verified and that rest on the validity of the hypothesized conditions, which may be unknown. Different hypothetical modeling assumptions could produce different answers. Chapter 2 discusses further counterfactual and potential-outcomes causal modeling.

Causality in Learning Analytics

The preceding sections describe how causal analytics and causal models can be used to describe and interpret patterns of associations among observed variables, predict outcome probabilities for different decisions or for the *status quo*, optimize decisions, and evaluate how interventions have changed or will change outcome probabilities. Table 1.6 summarizes models and methods for carrying out these tasks, as discussed in this chapter and Chap. 2.

Table 1.6 Models and methods for analytics tasks

Analytics task	Models for delivering results	Principles, methods, algorithms
<i>Describe, interpret, and explain observed data</i>	<ul style="list-style-type: none"> Dashboard displays for status, histories, events Correlations, clusters, interaction plots, exploratory data analysis Bayesian networks (BNs) and other DAG models showing probabilistic dependencies Simulation model representing the data-generating process 	<ul style="list-style-type: none"> Deep learning for extracting meaningful features, patterns, and clusters from data Most probable explanations in BNs (see Chap. 2) Aggregate, display, and interpret associations and patterns in data Use transport formulas to generalize particular findings
<i>Predict outcome probabilities if no intervention or change in policy</i>	<ul style="list-style-type: none"> DAG models, Dynamic Bayesian Networks (DBNs, see Chap. 2) Time series simulation models State space forecasting models Forecasting model ensembles 	<ul style="list-style-type: none"> Calculate conditional probabilities, $P(\text{output} \mid \text{input})$ Predictive analytics Time series forecasting Particle filtering
<i>Understand how choices affect outcomes or their probabilities; predict how they change for new interventions or changes in policy</i>	<ul style="list-style-type: none"> Causal DAG models MDP, POMDP System dynamics continuous simulation models Discrete-event simulation (DES) Optimal control models 	<ul style="list-style-type: none"> Causal DAG learning algorithms Markov decision process (MDP), partially observable MDP (POMDP), and system dynamics model estimation and simulation DES simulation algorithms
<i>Prescribe what to do next</i>	<ul style="list-style-type: none"> Decision tables Influence diagram (ID) models Statistical decision theory Simulation-optimization models MDP and POMDP models Optimal control and reinforcement learning models 	<ul style="list-style-type: none"> Influence diagram (ID) algorithms Simulation-optimization algorithms Monte Carlo Tree Search (MCTS)
<i>Evaluate how actions or policies affect outcome probabilities</i>	<ul style="list-style-type: none"> Quasi-experiments, including time series intervention studies Counterfactual models Causal DAG; simulation models 	<ul style="list-style-type: none"> Simulation-based what-if analyses Counterfactual comparisons: Show how outcome probabilities change as decision inputs are varied

The causal DAG models (including influence diagrams) and simulation models shown in the middle column can support the analytics tasks shown in the left-most column *if* the required models are known.

But suppose that no trustworthy causal model of how outcome probabilities depend on decisions is available when decisions must be made. Then, how should a decision maker proceed? This commonly happens when the system or situation of interest is not yet understood well enough to credibly simulate or model how it will (probably) respond to changes in inputs. Learning from experience how to act effectively in such an uncertain world is the central challenge for *learning analytics*.

The usual context for machine learning algorithms that address this challenge is a sequence of repeated opportunities to choose an act, perhaps informed by accompanying observations, resulting in outcomes to which rewards are assigned (Sutton et al. 1992). The decision maker (d.m.) makes a sequence of decisions, such as which treatments to prescribe to patients with certain medical histories; which loan applications to approve; how much inventory to order each day; or how to adjust the position of an autonomous vehicle on the road. Outcomes are gradually revealed: patients improve or worsen or die, borrowers repay their loans or don't, vehicles stay on the road or stray off it, and so forth. Some of these outcomes earn rewards or penalties; generically, we use “rewards” to refer to the values assigned to the outcomes, with larger rewards being preferred. The d.m. seeks to make decisions to maximize an objective function such as average reward per unit time, total reward by a termination time, or discounted reward.

This set-up is, of course, extremely reminiscent of those for statistical decision theory, MDPs, POMDPs, and simulation-optimization. The big difference is that now, since no causal model is initially known, the d.m. must act on, and learn from, the real-world environment rather than from a model of it. This implies that certain outcomes, such as patient deaths or autonomous vehicle crashes, which might be informative in a simulation should be avoided as well as possible while learning. This is an aspect of the famous *exploration-exploitation* tradeoff between choosing to reap the rewards from what are currently expected to be the best or safest decisions based on experience so far; or instead to take some risks to discover whether other decisions might yield even better rewards. The following approaches can help to decide what to do next in such settings, even in the absence of a causal model.

- *Random guessing.* One possibility is to choose decisions at random and hope for the best. Unsatisfactory as this is, it can actually outperform some widely used techniques such as applying “risk matrices” to guide risk management intervention priorities based on estimated frequencies and severities of outcomes, rather than on how alternative actions would change outcome probabilities and expected rewards (Cox 2008).
- *Reinforcement learning.* A much more promising approach for many applications is to allow for some randomization in selecting actions, but also to keep track of what was tried and what rewards resulted and to systematically modify action selection probabilities based on this experience to favor those yielding larger average rewards. This basic idea is exploited in modern reinforcement learning and adaptive control algorithms used in machine learning, artificial intelligence, and control engineering (Sutton et al. 1992). An example is *SARSA-learning*, which learns from experiences, encoded as “state-action-reward-state-action” (SARSA) sequences, by updating the estimated value of taking a specific act in a specific state (the first state-action (SA) pair in the SARSA sequence) to reflect the difference between predicted and actual outcomes (the immediate reward, R , and the next state encountered and act taken, say s' and a' , comprising the rest of the sequence). For a Markov decision process (MDP), the SARSA update equation is the following simple modification of Eq. (1.4):

$$Q(s, a) \leftarrow Q(s, a) + \alpha(R + \gamma Q(s', a') - Q(s, a)) \quad (1.12)$$

In words, the estimated value of taking act a in state s , denoted by $Q(s, a)$, is updated by taking its current value on the right side of the update arrow \leftarrow and adding an increment proportional, with proportionality constant α , to the difference between the estimated experienced reward, $R + \gamma Q(s', a')$, and the reward actually received. Under certain conditions, such as that each state of the MDP is visited infinitely often as the time horizon for the MDP becomes infinitely long (no absorbing states), the SARSA updating rule leads to the same optimal policy as the Bellman equation for stochastic dynamic programming, but with the considerable practical advantage that the rewards and transition probabilities need not be known initially: SARSA learns the optimal policy via well-chosen trial-and-error learning. Reinforcement learning software is available at these sites:

- <https://cran.r-project.org/package=ReinforcementLearning> for R
- <http://www.wildml.com/2016/10/learning-reinforcement-learning/> for Python.
- *Model-based Bayesian reinforcement learning* (Ross et al. 2011). A different approach to adaptively optimizing decisions in the absence of a known model is to make one up and then use it to make decisions. For example, if the transition rates among states in an MDP are modeled as having a Dirichlet prior distribution, then this prior can be updated by conditioning it on observed transitions, thereby obtaining an updated Dirichlet posterior, and decision can be made to maximize expected reward with respect to the updated distributions. Ross et al. (2011) develop this approach in detail for MDPs and POMDPs.
- *Optimism under uncertainty* (Auer et al. 2002). A useful paradigm for many adaptive decision optimization problems under uncertainty, including clinical trials, is the *multi-armed bandit* (MAB) class of decision problems. A “one-arm bandit” is a slot machine: its single arm takes one’s money. Putting in a coin may result in a random payoff. The probability distribution for this reward is initially unknown, but it can be learned about gradually, for a cost, by continuing to play. A MAB problem generalizes this setup by having multiple machines. In applications, these might represent the different arms or treatments of a clinical trial, variants of ads or of web pages in A/B testing, performances of different subcontractors or vendors in a supply chain, and so forth. The decision problem is to choose which machine to play next, given the history of tries and rewards so far. A relatively simple approach provides a solution with provably low regret, measured as the cumulative difference between the expected reward if the true best (highest average reward) option were always selected and the expected reward from the decisions actually made. The simple rule, called the *upper confidence bound 1 (UCB1) algorithm*, prescribes playing each machine or action once and then using available data to compute an upper confidence bound on the average reward from each machine or action j via the formula: $UCB_j = m_j + \text{sqrt}(2 \log t / n_j)$, where sqrt is the square root function, t is the total number of trials for all machines so far, and n_j is the number of times action or machine j has been

tried. It can be proved that the regret from using UCB1 grows no faster than roughly logarithmically—about as much in the first 10 decisions as in the next 90, and then in the next 900, and so on—and that no decision rule (including the optimal one) can guarantee regret that grows more slowly than this (Auer et al. 2002). The UCB1 algorithm illustrates the useful principle of *optimism under uncertainty*, i.e., always choose next the action that might be best, meaning that it has the highest upper confidence bound. Doing so is a low-regret strategy.

- *Probability matching and Thompson sampling* (Ortega and Braun 2014). Another simple but powerful decision rule for MAB problems and more general classes of adaptive decision and control problems, including Markov decision processes (MDPs), is *probability matching*. This selects each action according to the current estimated probability that it is the best one, meaning the one giving the highest average reward. Thus, if an action is known to be best, it will always be selected; if it is known not be best, it will never be selected; and if it might be the best, it is selected with probability equal to the probability that it is best. A straight-forward software implementation initially assigns each action a uniform prior probability distribution for the unknown probability that it is the best (i.e., has highest average reward). Its distribution for the probability of being best is then updated by conditioning on experience. Finally, on each move, the action to be chosen is decided by sampling from the current distributions for all actions and then taking the action with the highest sample value. In slightly more detail, for each action, the prior probability of being best is initialized to the unit normal distribution $U[0, 1]$. Its posterior distribution after it has generated S successes and F failures is the beta distribution with parameters $S + 1$ and $F + 1$. If the actions generate binary rewards ($1 = \text{win}$, $0 = \text{lose}$), then the number of successes S for an action is just the number of wins it has generated so far. For an arbitrary reward distribution normalized to run from $0 = \text{lowest reward}$ to $1 = \text{highest reward}$, the observed reward on any trial (a number between 0 and 1 on this scale) is scored as a success with probability equal to the reward, and as a failure otherwise (Agrawal and Goyal 2012). This idea, known as Thompson sampling, was proposed in the 1930s to minimize the number of patients who receive the less effective drug during a clinical trial. It has been rediscovered many times since, recently gaining prominence in artificial intelligence; it is also related to animal behaviors such as optimal foraging (Averbeck 2015). Despite its simplicity, the performance of Thompson sampling, as measured by cumulative regret and speed of convergence to the optimal policy, compares favorably to UCB1 on many test problems.

MAB software for computing the probability of each action being best after any history is available at the following link:

– <https://cran.r-project.org/web/packages/bandit/bandit.pdf>

- *On-line decision-making using a model ensemble* (Kalai and Vempala 2005). An “on-line” decision problem refers to a sequence of decisions that must be made one at a time, without knowledge of the future or ability to backtrack and revise an earlier decision. Applications might include approving or rejecting loan

applications, choosing which treatment to prescribe to patients at a walk-in treatment center, or deciding which advertisement or version of a web page to display to visitors arriving at a web site. When no single best model describing the relation between acts and consequence (i.e., reward or loss) probabilities is known initially, an ensemble of possible models can be used instead. These might correspond to different statistical models, machine-learning algorithms, experts, or other sources of advice and recommendations. The *follow-the-perturbed-leader* (FPL) strategy slightly modifies the idea of always choosing the action recommended by the model with the best empirical performance so far, by perturbing the cumulative reward for each model by a random amount before selecting the best one (i.e., the one with the greatest cumulative reward or the least cumulative loss). In many settings, the decisions recommended by FPL converge rapidly to those recommended by the best of the considered models, implying that little is lost in the long run by learning which model(s) to use based on their observed performance (Kalai and Vempala 2005). Since forecasting can be viewed as a special type of decision-making in which the forecaster decides what prediction to make and the objective is to optimize some performance metric (such as minimizing mean squared prediction error or a weighted sum of type 1 and type 2 errors for classification), FPL and other algorithms with performance guarantees (e.g., provably low cumulative regret) can also benefit predictive analytics.

- *Response-surface methods and Evolutionary Operations (EVOP)*. If experimentation is possible and relatively inexpensive, then a complimentary tactic is to systematically vary the decision variables and study resulting responses using industrial design-of-experiments (DOE) techniques. The resulting data can then be used to fit an approximate statistical model, called a *response surface* model, showing the model-predicted expected value of the response or outcome variable that one wants to maximize for each combination of values of the decision variables. Quadratic regression models are a popular choice for this purpose. The response surface model can then be used to estimate the combination of decision variables that will maximize the expected value of the response variable. Repeating this process sequentially by (a) collecting further data around the estimated optimum using a designed experiment that varies the input factors slightly around their current levels; (b) fitting a new response surface model; and (c) re-optimizing the combination of inputs can improve the response surface model and the optimization of decision variables. This approach, called Evolutionary Operation (EVOP), has been used since the 1950s in chemical and manufacturing industries to improve ongoing production operations without interrupting them (Box 1957). When experiments are relatively costly or time-consuming, an alternative is to use *stochastic approximation* algorithms (Dupac 1965) or more recent *stochastic gradient* adaptive optimization algorithms (Lange et al. 2014) to adjust the inputs in order to move toward the optimum of the initially unknown (and perhaps gradually changing) response surface. The guiding metaphor for these approaches is to climb a hill of initially unknown shape by collecting data and then moving in the direction of the estimated greatest

improvement. Response surface methodology (RSM) is supported by freely available software such as the *rsm* package in R, available at the following link:

- <https://cran.r-project.org/web/packages/rsm/rsm.pdf>

These methods for deciding what to do when no causal model is available can be combined and extended in various ways. Doing so has led to a recent burgeoning of productive new methods for adaptive decision-making and learning in artificial intelligence and machine learning. For example, the UCB1 algorithm has been combined with Monte Carlo tree search (MCTS), yielding the “UCT” (UCB1 for trees) algorithm for learning to play games. UCT has been streamlined and further improved by incorporating value-of-information (VOI) heuristics, and has been applied successfully to multi-armed bandits, Markov decision processes, and adversarial games (e.g., Tolpin and Shimony 2012; Pepels et al. 2014). The original multi-arm bandit problem has been generalized to settings where the reward distributions from different actions are not independent, are correlated with observable information (“contextual bandits”), change over time (“restless bandits”), or are manipulated by an intelligent opponent or attacker (“adversarial bandits”). Deep learning methods for succinctly describing and estimating state-action values have been combined with variations on SARSA by several investigators (e.g., Ganger et al. 2016; Van Seijen et al. 2009). There are many other examples of the confluence of technical ideas leading to ever-more powerful and robust optimization of decisions without models. These procedures are so effective that they are often used for simulation-optimization when models are known. Both adaptive optimization of action-selection probabilities and EVOP-like design-of-experiments and iterative improvement of estimated statistical models and optimization of decisions implied by the estimated models are widely used in current simulation-optimization algorithms and software (Amaran et al. 2016).

The success of these model-free algorithms for adaptively learning to optimize decisions raises the questions of why and whether causal models are really needed in analytics. After all, if a learning algorithm can learn fairly quickly to make effective decisions without them, then why use them? Answers include the following.

- For *descriptive analytics*, causal models help to describe, explain, and interpret observed data. The learning methods just described are black-box methods, in the sense that they map data to decisions without trying to explain or interpret why the decisions make sense. Their sole goal is to discover how to optimize some objective, such as minimizing cumulative regret. By contrast, causal models and causal analytics methods, such as the Bayesian networks and algorithms introduced in Chap. 2, help to understand and visualize the probabilistic interdependencies (and independencies) among variables. They can also provide a most probable explanation (MPE) for observations in terms of the likely value of unobserved variables. This allows descriptive analytics to go well beyond simply providing statistical summaries or compact descriptions and visualizations of what is observed, but also to provide possible explanations, e.g., diagnoses for anomalies observed in patients or in systems. Such explanations carry implications for action that mere description without causal interpretation does not.

- For *predictive analytics*, causal models can predict what will happen if interventions are undertaken in new settings. Neither the interventions nor the conditions under which they are applied need to have been seen before to make accurate probabilistic predictions using transport formulas (Bareinboim and Pearl 2013; Lee and Honavar 2013). Transport formula software is available at <https://cran.r-project.org/web/packages/causaleffect/causaleffect.pdf>.
- Similarly, for *prescriptive analytics*, causal models such as influence diagrams can optimize decisions for new situations and new interventions. This is something that black-box data-driven methods that do not generalize beyond the data-generating process for the observations at hand cannot do.
- From the standpoint of *causal analytics*, algorithms that learn the conditional probability distributions or conditional expected values of rewards for different actions taken (or, more generally, different actions taken in the context of available observations) are in fact doing a form of causal discovery. This point of view is articulated further specifically for generalizations of Thompson sampling by Ortega and Braun (2014).
- For *evaluation analytics*, causal models provide a principled way to compare what has actually happened to the full probability distribution for what could have happened, both under the conditions of the actions, policies, treatments, or interventions that were actually taken and under alternative, counterfactual conditions.

Thus, despite the exciting advances in machine learning applied to efficient discovery of low-regret decision rules, causal modeling and inference methods have valuable roles to play throughout these other parts of the risk analytics sequence.

Causality in Collaborative Analytics

Commercially available collaborative analytics information technologies (IT) platforms enable multiple data collectors, analysts, and decision-makers to share data and insights, both throughout an organization and in physical or virtual “war rooms” and operations centers where information is displayed, analyzed, and acted upon by teams collaborating in real time. Causal analytics models can provide useful content and organizing structures to coordinate the collection, analysis, and display of information on such platforms. Causal graph models such as causal Bayesian networks and influence diagrams provide useful conceptual and visual aids to help teams of collaborating analysts combine, coordinate, and synthesize the research and expertise of different subject matter experts (SMEs). They allow SMEs with complementary expertise about different parts of a problem—such as emission reductions technologies, costs, and health effects of pollution in Fig. 1.5—to develop the parts of the diagram and the corresponding conditional probability tables or models that each understands best, while understanding how the pieces fit together.

Causal Models in Game Theory: Normal and Extensive Forms, Characteristic Functions

More generally, effective cooperation across and collaboration within analytics centers requires principles and insights into how multiple agents can best work together to accomplish shared goals such as managing a complex system—whether a logistics network, a nuclear power plant, a commercial forest or fishery, or a multi-divisional firm—to reduce risks and increase the production of desired outcomes. Many socioeconomic, technological, and business systems are governed by the interacting choices of multiple decision-makers. The decision-makers may have different information, different opportunities to take actions, and different individual preferences and values, beliefs, and incentives. For example, a business's pollutant emissions may be regulated by a combination of Federal, state, county, and city authorities. It may also reflect past capital budgeting and investment decisions about which production and pollution-abatement technologies to purchase; ongoing operating and production decisions from multiple divisions; and remediation and clean-up decisions by various vendors and subcontractors hired by the business. The confluence of these many decision threads over time affects how much pollution is actually produced.

To understand how to improve decision-making by multiple decision-makers in groups, teams, or organizations, it is necessary to extend normative principles for prescriptive analytics beyond the expected utility formalism (Eq. 1.1) to allow for multiple agents. Normative principles that seem unexceptionable for individual decision makers need to be rethought for groups of decision makers. For example, an individual decision maker should never choose a strictly dominated strategy, meaning one that always yields a lower-utility outcome than some other feasible choice no matter how uncertainties are resolved. But the notorious Prisoner's Dilemma shows that in situations with multiple decision-makers and negative externalities, if each decision-maker chooses an undominated strategy, then these choices can cause each to receive lower-utility outcomes than if each had instead chosen a dominated strategy (Thomas 2003).

Cooperation and coordination of individual choices are required to avoid such Pareto-inferior outcomes. With the need for coordination and cooperation come a host of new technical, practical, and conceptual issues, including how to model information and incentives; trust and reputation; credible commitments and signals of intent; asymmetric and private information; altruism and free riding; costly private or public monitoring and enforcement of agreements; and sharing of information, control, and rewards or losses. Game theory addresses these and related topics in detail (Thomas 2003; Myerson 1991; Osborne 2004; Shoham and Leyton-Brown 2009). It provides both theory and algorithms for fair division of jointly produced gains; clarifies the existence, uniqueness, computational complexity, and computability of various proposed solutions to the joint decision problem for multiple agents

(e.g., finding sets of choices such that no agent or coalition can increase its own expected reward by unilaterally changing its own decisions); and establishes impossibility results for design of collective choice mechanisms that simultaneously satisfy several desired properties, such as fairness (e.g., symmetric treatment of participants), efficiency, budget balance, and incentive-compatibility (e.g., voluntary participation and cooperation). In each of these areas, causal models relate the joint decisions of multiple agents to the joint probability distributions of their rewards. These models are essential for analysis and applications. Important special cases that focus on multiple agents cooperating to accomplish shared goals include the mathematical theories of teams and organizations (Marschak and Radner 1972; Yüksel and Saldi 2017) and of risk and reward sharing in investment syndicates (Wilson 1968).

Game theory uses causal models that generalize normal form (decision table) and extensive form (decision tree) models from decision analysis to allow for multiple decision makers. These extensions yield payoff matrix models mapping n -tuples of player choices to n -tuples of resulting rewards or utilities, where n is the number of players; and game tree models in which different players are able to choose moves (branches) at different nodes, based on the information available to them at the time. The decision makers are usually called “agents” in the artificial intelligence and machine learning literature, or “players” in classical game theory terminology. They may have different initial information, including information about each other’s utilities, beliefs, intentions, capabilities, constraints, and resources (or, more generally, each other’s “types”), which they update over time by conditioning on their own observations. Private information arises when observations differ for different agents. Their actions jointly determine the probabilities of their individual rewards (often called “payoffs” in the older game theory literature) via the causal model. In addition to normal form payoff matrices and extensive form game trees used in analysis of non-cooperative games, cooperative game theory uses causal models called “characteristic functions” that specify how much reward each subset or coalition of agents can obtain if it acts by itself (Myerson 1991; Shoham and Leyton-Brown 2009). In most game-theoretic modeling, the relevant causal models (or “mechanisms”) mapping player choices to their outcome or reward probabilities are exogenously specified, sometimes by a social or institutional mechanism designer attempting to elicit desired behaviors from agents. In evolutionary game theory, there is no centralized mechanism designer, but mechanisms emerge and evolve from the interactions of agents game theory. Most of the game-theoretic analysis then focuses on what the agents should choose to do, or on what behaviors and decision rules governing agent interactions (including shared norms and conventions) will emerge and persist, given the specified causal relations between choices (or behaviors) and their probable consequences. In practice, causal analytics methods and models applied to available data and knowledge may be needed to create usefully realistic game-theoretic models for supporting predictions and decisions.

Causal Models for Multi-agent Systems

The causal models discussed earlier in this chapter for prescriptive analytics also have generalizations for multiple cooperating decision-making agents. For example,

- *Influence diagrams* can be generalized to multi-agent influence diagrams (MAIDs), with game-theoretic reasoning then being executed by agents within the MAID graph (Koller and Milch 2003). At present, the literature on MAIDs is relatively small.
- *Markov decision processes (MDPs) and partially observable MDPs (POMDPs)* have been extended to multi-agent systems (MAS). Multi-agent planning and coordination problems, as well as multi-agent team reinforcement learning problems, are often formulated as large-scale MDPs or POMDPs, augmented with models describing observations and communication among agents, that can be solved using a combination of special-purpose techniques (e.g., DAG-based factoring and function approximation methods) (Amato and Oliehoek 2015; Koller and Parr 1999).
- *Optimal control* problems, both deterministic and stochastic, have corresponding *distributed control* formulations when multiple sensors and controllers (agents consisting of software and actuators that they control) are distributed throughout a system and cooperate to manage its performance. The agents communicate with each other via a communication network to share information generated by their local sensing and control activities and to coordinate their actions to achieve common goals. Control is often arranged in a hierarchy. In a two-level control hierarchy, a centralized planner collects information from each local controller and allocates tasks and goals (and in some cases, resources). The local controllers then use their own local information, resources, and control opportunities to pursue their assigned goals. Such hierarchical control architectures have been applied to control of electric power networks with randomly varying local loads and conditions (Ilic and Liu 1996); to control of traffic signals which, in turn, provide distributed control of traffic flows in cities (Abdoos et al. 2014); and, more recently, to control of swarms of autonomous vehicles or drones engaged in search-and-rescue or other missions (Gómez et al. 2016). In the latter application, when the swarms operate in uncertain and variable environments (e.g., with unpredictable wind gusts that can suddenly push a drone from its planned trajectory), the local controllers must respond quickly and competently to their changing local conditions. Bearing in mind the uncertainties and performance constraints the local controllers must confront, e.g., in navigating narrow spaces in the presence of disruptions from wind or from obstacles and from other drones, the centralized planner may assign longer but safer flight paths to some members of the swarm to reduce risks (Gómez et al. 2016). Conditions under which hierarchical control is more effective or less effective than purely decentralized control, in which local control agents communicate with each other and form their own plans directly instead of going through a centralized planning and coordinating controller, are still being investigated in a variety of applications

(Feng et al. 2017). The annual RoboCup competitions for robot soccer provide a useful experimental forum for evaluating different communication, coordination, and control protocols for teams of cooperating agents in uncertain, dynamic, and adversarial environments.

- *Reinforcement learning* algorithms have been extended to multi-agent reinforcement learning (MARL) algorithms for multiple cooperating agents (software programs) communicating with each other as they sense, act, and learn. Applications with substantial technical literatures include distributed sensing and control of urban traffic via communicating traffic signals; intelligent control of electric power production, distribution, storage, and routing in smart grids; real-time control and efficiency improvement of industrial processes based on distributed sensors and actuators; and management of multiple autonomous unmanned aerial or ground vehicles, including drone swarms. MARL algorithms are included as components of many other hierarchical and distributed control systems for such applications.
- *Multi-armed bandit* problems have been studied for teams of cooperating players who can try different arms and share information about what they have learned. For example, Chakraborty et al. (2017) compare centralized and decentralized control for teams of such agents when communication is costly and each agent in each period must choose whether to collect more information before broadcasting what it has discovered. They show that both centralized and decentralized control architectures lead to decisions with no-regret properties for the communication cost model considered.
- *Simulation and simulation-optimization* programs today routinely include simulation of the behaviors of multiple interacting intelligent agents. Applications are as diverse as managing aquaculture and protecting ecosystems more efficiently; countering piracy in commercial shipping by coordinating convoys and mutual help among ships; optimizing detection and responses to cyberattacks and fraud; designing and managing supply chains and logistics networks to achieve and maintain desired feasible levels of risk, profitability, performance, resilience to disruption; and other objectives; and changing layouts and operations to improve patient experiences in hospitals, customer experiences in shopping malls, and driver experiences on roads. Theory, practical software, and deployed applications of multi-agent simulations and system design improvements based on simulation-optimization of multi-agent systems (MAS) can be found in many sources. The annual *Proceedings of the Winter Simulation Conference* (WSC) (<https://informs-sim.org/>) is one of the best.

The artificial intelligence, machine learning, and control engineering literatures on cooperative multi-agent control of uncertain systems and on coordination of team actions and communications in uncertain environments are vast. They include many elegant and profound results and frameworks, including exact forms of optimal control laws for each agent in some models and proposed formulations that apply to human organizations (Marschak and Radner 1972) as well as to engineered systems (Ho and Chu 1972). Causal modeling contributes several important

components to this rich body of knowledge. First, causal models, when they are available, permit simulation (or, in special cases, analytic solutions) for predicting the outputs or output probabilities of a system for different inputs. Simulation, in turn, permits simulation-optimization of system designs, decision rules, and operating decisions to improve performance metrics and make preferred outcomes more likely, even in uncertain environments and even when multiple intelligent agents are involved in the operation or use of the system. Thus, causal models provide a foundation for improving the design and operation of many real-world systems. Second, the study of collaboration in multi-agent systems (MAS) suggests new ways to manage a variety of risks using distributed sensors and controllers. The opportunities to use distributed communication and control systems to monitor and reduce risks and to improve performance in areas from production and logistics to transportation and energy infrastructures to improved healthcare and patient and customer experiences are already enormous. They are increasing as sensor networks, the internet of things, and ubiquitous access to data become increasingly widespread realities. Finally, the study of collaboration among software agents, each collecting and processing local data, updating its own beliefs about the causal relation between its actions (and perhaps also the actions of others) and resulting probabilistic changes in the world it acts on, taking actions based on its own information and beliefs, and communicating results within a team, may yield insights that help to improve distributed analytics and cooperative goal-seeking, control of enterprises and systems, and risk management in human organizations. This was a goal of mathematical team theory more than half a century ago (e.g., Marschak and Radner 1972; Ho and Chu 1972). Its fruition is starting to be seen today in applications.

Conclusions: Causal Modeling in Analytics

This chapter has surveyed numerous causal models and methods for representing and calculating conditional probabilities of outcomes in response to changes in the inputs to a system or situation. They are summarized in Table 1.7. Their inputs can usefully be partitioned into those controlled by a decision-maker—the *decision variables* or *control variables*, usually with values to be set or chosen by the decision-maker to optimize expected utility or some other criterion, subject to feasibility constraints—and those not controlled by the decision-maker. The uncontrolled inputs represent random variables, such as exogenous noise, shocks, or uncertain quantities, that affect the outcomes but that are not selected or set by the decision-maker. Each causal model can be thought of as mapping its inputs into conditional probabilities for outputs, which include outcomes and rewards in all models, and also observations and changes of state in dynamic models.

A causal model links actions to their probable consequences and shows how other variables affect each other and the outcomes. Probabilistic dependencies among variables are expressed through conditional probability table (CPTs) or other conditional probability models. Causal models differ from statistical models by clearly

Table 1.7 Causal models relating decisions to probabilities of consequences

Causal model	Controlled inputs (decisions)	Random outcomes
Decision table (normal form)	Row	Column (state of nature) and outcome
Decision tree (extensive form)	Branch at choice node	Branches taken at chance nodes; terminal node and reward received
Influence diagram	Decisions at choice nodes	Value of random variable at chance node
Markov decision process (MDP)	Act selected from current state	Next state, reward
Semi-Markov decision process	Act selected from current state	Dwell time in current state, next state, reward
Partially observable MDP (POMDP)	Act selected given current conditional probabilities of states	Next state, observation, reward
Deterministic optimal control	Input at each moment	State, output, and reward trajectories (deterministic)
Stochastic optimal control	Input at each moment	Output, state, and reward trajectories
Statistical decision theory	Decision rule or policy mapping observations to actions	State, observations, rewards
Simulation model (continuous or discrete-event)	Initial conditions, controllable inputs at each moment	Outcome trajectories
Multi-arm bandit	Which act (arm) to choose next	Rewards
Response surface model	Next settings of inputs	Next output (e.g., yield)
Game theory models	Each player's choice of strategy	Rewards to players
Multi-agent systems (MAS)	Agent decision rules	Behaviors, rewards

distinguishes between effects of *doing* something (setting a controlled variable to a certain value) and of *seeing* something (observing that a variable has a certain value): acting on the world can cause the probability distributions of variables to change, but observing the world only allows us to update our beliefs about it (Pearl 2009). This fundamental distinction is discussed further in Chap. 2.

Once a causal model for how actions affect outcome probabilities has been formulated, it can be used for a variety of purposes, including describing how variables depend on each other; identifying the most probable explanations for observations; comparing outcome probabilities under different counterfactual scenarios; predicting probabilities of outcomes for current or assumed initial conditions and decision rules; and recommending courses of action that make preferred outcomes more likely to occur. To extract useful advice from causal models on what to do next, it is necessary to have solvers that use them to determine the feasible settings of decision variables that cause the most desirable probability distribution of outcomes. A variety of algorithms are available for using the causal models in

Table 1.7 to identify acts or decision rules that optimize various criteria, such as maximizing expected utility or average reward per period or expected net present value of future rewards; or minimizing expected loss or cumulative regret. Approaches introduced in this chapter include the following:

- *Static optimization*: Suppose that $u(a, s)$, represents the expected utility of the consequence of choosing act a if the state of the world is s , where a is an act or decision rule in a feasible choice set A and s is in a set S of possible states with a known probability distribution or probability measure (which may in general depend on the choice of a). Given this model, optimization techniques can be used to search for a choice in A that maximizes the expected value of $u(a, s)$, i.e., expected utility. In the simplest case of a small decision table with only a few rows and columns, where A is the set of rows and S is the set of columns, brute-force evaluation of the expected utility for each row reveals the optimal choice. In many operations research and risk management problems, however, the choice set A includes one or more continuous decision variables, and perhaps some discrete ones as well, and the set A is specified via a set of constraints determining jointly feasible choices of their values. Mathematical programming algorithms for constrained optimization, such as linear, nonlinear, or mixed integer programming solvers, are used to solve such problems.
- *Stochastic dynamic programming (SDP)*: In a decision tree or game tree representing a sequential decision problem, classical (backward) dynamic programming, beginning with the rewards at the terminal nodes (“leaves”) of the tree and working back by calculating expected utilities at chance nodes and maximizing them at choice nodes, provides an effective procedure for identifying the optimal initial choice, i.e., the one that maximizes expected utility when all subsequent decisions are made optimally. Dynamic programming is also used to formulate the Bellman equations for optimizing decisions in Markov decision processes (MDPs) and partially observable MDPs (POMDPs) and the Hamilton-Jacobi-Bellman equations for optimal control of deterministic and stochastic dynamic systems.
- *Optimal control, dynamic and adaptive optimization, and reinforcement learning*: Mathematical analysis using dynamic optimization methods leads to closed-form optimal decision rules for many optimal control problems, for both deterministic and stochastic dynamic systems. It also yields insights into the qualitative characteristics of optimal policies, such as the existence of thresholds or control limits for observed attribute values that should trigger specific actions such as sparing or harvesting. Numerical optimal control algorithms and software packages are increasingly available for optimizing multiple controlled inputs together over time when the equations representing a causal model of a dynamic system are known. When they are not, reinforcement learning approaches such as the SARSA algorithm for MDPs or various POMDP solvers can be used to discover effective decision rules for systems with uncertain initial conditions and state transition dynamics. Optimal control methods and MDP and POMDP

models and solvers have a wide range of practical applications, such as the following:

- Harvesting and replenishing a forest or other renewable resource over time
- Managing a perishable inventory with time-varying demands and values
- Timing the storage and release of water behind a hydroelectric dam to take advantage of changing hourly electricity prices
- Scheduling inspections of components in a complex industrial plant or infrastructure network to increase reliability and reduce failure risks
- Scheduling the screening, treatments, and release of patients from a hospital.

Adaptive optimization principles such as optimism under uncertainty (e.g., the UCB1 algorithm), probability matching (via Thompson sampling), and “follow-the-perturbed-leader” (FPL) algorithms for on-line prediction and decision problems have proved surprisingly effective in many settings for yielding high-quality, low-regret approximations to the optimal control laws that would be used if valid causal models were known. These adaptive optimization techniques can be viewed as learning causal models on the fly for the effects of acts on outcome probabilities in different situations.

- *Markov Chain Tree Search (MCTS)*: For decision or game trees that are too large to be fully generated and evaluated by dynamic programming, MCTS provides a practical heuristic alternative. MCTS emphasizes forward search into a partly known tree, rather than backward optimization starting from the tips of a fully known one. MCTS heuristics have been incorporated into algorithms for solving large-scale POMDPs.
- *Influence diagram solvers* use graph-theoretic algorithms (e.g., bucket elimination, which is closely related to non-serial dynamic programming) to solve for the optimal decisions at choice nodes.
- *Simulation-optimization algorithms* combine numerical optimization techniques such as gradient descent, statistical techniques such as response surface methodology, smoothing and aggregation techniques such as deep learning, and computational Bayesian sampling and updating methods such as particle filtering, to seek combinations of controllable inputs over time that optimize the performance metrics for a simulated system.
- *Multiagent Systems (MAS)* algorithms extend simulation and simulation-optimization concepts to teams or systems of interacting agents, each of which uses its own information to decide what to do and what to communicate. Studying the behavior of such systems in response to different communication and control protocols and different simulated environmental conditions can suggest system design and operating principles to improve the performance, reliability, and safety of real systems.

Mastering any of these techniques provides valuable skills for formulating and solving certain types of analytics problems using appropriate causal models. Their learning curves are substantial, however, and thorough understanding usually requires experience applying them. The references at the end of this chapter provide

useful points of entry to the relevant technical literatures. Fortunately, modern object-oriented software makes it possible to encapsulate much of the detailed knowledge and expertise needed to apply these algorithms correctly to appropriate causal models or data.

Causal modeling is used throughout the rest of risk analytics to help describe and interpret data and to predict, optimize, and evaluate the impacts of risk management decisions (Table 1.6). Chapter 2 examines more closely the multiple meanings of “cause” and “causality” and discusses the senses that are most useful for analytics tasks such as prescription and evaluation. It also introduces techniques and algorithms for learning causal models from data and for using them to carry out other analytics tasks. The emphasis is largely, but not exclusively, on causal Bayesian networks (BNs) and other models that can be represented as BNs. Such models turn out to be both relatively tractable to learn from data in many practical applications, and also highly useful for insightful description, prediction, prescription, and evaluation of the effects of decisions and interventions. Chapter 2 introduces the main ideas of causal BNs and related causal modeling methods and describes free software implementations to enable practitioners to apply them.

References

- Abdoos M, Mozayani N, Bazzan ALC (2014) Hierarchical control of traffic signals using Q-learning with tile coding. *Appl Intell* 40(2):201–213
- Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. *J Mach Learn Res* 23(39):39.1–39.26. <http://proceedings.mlr.press/v23/agrawal12/agrawal12.pdf>
- Akhavan-Tabatabaei R, Sánchez DM, Yeung TG (2017) A Markov decision process model for cervical cancer screening policies in Colombia. *Med Decis Mak* 37(2):196–211. <https://doi.org/10.1177/0272989X16670622>
- Amacher GS, Ollikainen M, Koskela E (2009) Economics of forest resources. The MIT Press, Cambridge, MA
- Amato C, Oliehoek (2015) Scalable planning and learning for multiagent POMDPs. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence: 1995–2002. www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewFile/9889/9495
- Amaran S, Sahinidis NV, Bikram S, Bury SJ (2016) Simulation optimization: a review of algorithms and applications. *Ann Oper Res* 240(1):351–380
- Ashcroft M (2013) Performing decision-theoretic inference in Bayesian network ensemble models. In: Jaeger M, Nielsen TD, Viappiani P (eds) Twelfth scandinavian conference on artificial intelligence, vol 257, pp 25–34
- Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Mach Learn* 47(2/3):235–256. <https://doi.org/10.1023/A:1013689704352>
- Averbeck BB (2015) Theory of choice in bandit, information sampling and foraging tasks. *PLoS Comput Biol* 11(3):e1004164. <https://doi.org/10.1371/journal.pcbi.1004164>
- Bala MV, Mauskopf JA (2006) Optimal assignment of treatments to health states using a Markov decision model: an introduction to basic concepts. *PharmacoEconomics* 24(4):345–354
- Bareinboim E, Pearl J (2013) Causal transportability with limited experiments. In: Proceedings of the 27th AAAI conference on artificial intelligence, pp 95–101. ftp://ftp.cs.ucla.edu/pub/stat_ser/r408.pdf

- Beck JL, Zuev KM (2017) Rare event simulation. In: Ghanem R, Higdon D, Owhadi H (eds) Handbook of uncertainty quantification. Springer, New York. <https://arxiv.org/pdf/1508.05047.pdf>
- Bennett CC, Hauser K (2013) Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. *Artif Intell Med* 57(1):9–19. <https://doi.org/10.1016/j.artmed.2012.12.003>
- Bertsekas DM, Shreve SE (1996) Stochastic optimal control: the discrete-time case. Athena Scientific, Belmont, MA
- Bier VM, Cox LA Jr (2017) Coping with uncertainty in adversarial risk models. In: Abbas A, Tambe M, von Winterfeldt D (eds) Improving homeland security decisions. Cambridge University Press, New York
- Browne CB, Powley E, Whitehouse D, Lucas SM, Cowling PI, Rohlfsen P, Taverne S, Pere D, Samothrakis S, Colton S (2012) A survey of Monte Carlo tree search methods. *IEEE Trans Comput Intell AI Games* 4(1):1–43
- Box GEP (1957) Evolutionary Operation: a method for increasing industrial productivity. *J R Stat Soc Ser C Appl Stat* 6(2):81–101. <https://doi.org/10.2307/2985505.JSTOR2985505>
- Cami A, Wallstrom GL, Hogan WR (2009) Measuring the effect of commuting on the performance of the bayesian aerosol release detector. *BMC Med Inform Decis Mak* 9(Suppl 1):S7
- Campbell DT, Stanley JC (1963) Experimental and quasi-experimental designs for research. Houghton Mifflin Company, Boston, MA
- Cao Q, Buskens E, Feenstra T, Jaarsma T, Hillege H, Postmus D (2016) Continuous-time semi-Markov models in health economic decision making: an illustrative example in heart failure disease management. *Med Decis Mak* 36(1):59–71. <https://doi.org/10.1177/027989X15593080>
- Cartwright N (2003) Two theorems on invariance and causality. *Philos Sci* 70:203–224. <https://doi.org/10.1086/367876>
- Chakraborty M, Chua KYP, Das S, Juba B (2017) Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI), Melbourne, Australia, pp 164–170. <https://doi.org/10.24963/ijcai.2017/24>
- Choi E, Schuetz A, Stewart WF, Sun J (2017) Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 24(2):361–370. <https://doi.org/10.1093/jamia/ocw112>
- Clancy L, Goodman P, Sinclair H, Dockery DW (2002) Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 360(9341):1210–1214
- Cox LA Jr (2008) What's wrong with risk matrices? *Risk Anal* 28(2):497–512
- Dayer MJ, Jones S, Prendergast B, Baddour LM, Lockhart PB, Thornhill MH (2015) Incidence of infective endocarditis in England, 2000–13: a secular trend, interrupted time-series analysis. *Lancet* 385(9974):1219–1228. [https://doi.org/10.1016/S0140-6736\(14\)62007-9](https://doi.org/10.1016/S0140-6736(14)62007-9)
- Dockery DW, Rich DQ, Goodman PG, Clancy L, Ohman-Strickland P, George P, Kotlov T, HEI Health Review Committee (2013) Effect of air pollution control on mortality and hospital admissions in Ireland. *Res Rep Health Eff Inst* 176:3–109
- Dorfman R (1969) An economic interpretation of optimal control theory. *Am Econ Rev* 59 (5):817–831
- Doucet A, Johansen AM (2009) A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan D, Rozovsky B (eds) The oxford handbook of nonlinear filtering. Oxford University Press, Oxford
- Dupac V (1965) A dynamic stochastic approximation method. *Ann Math Statist* 36(6):1695–1702
- Feng X, Shekhar A, Yang F, Hebner RE, Bauer P (2017) Comparison of hierarchical control and distributed control for microgrid. *Elect Power Comp Syst* 45(10):1043–1056. <https://doi.org/10.1080/15325008.2017.1318982>

- Fu MC (2016) AlphaGo and Monte Carlo tree search: the simulation optimization perspective. In: Proceedings of the winter simulation conference (WSC), 11–14 December 2016. IEEE, Washington, DC. <https://doi.org/10.1109/WSC.2016.7822130>
- Fu MC (ed) (2015) Handbook of simulation optimization. Springer, New York. www.springer.com/us/book/9781493913831
- Ganger M, Duryea E, Hu W (2016) Double Sarsa and double expected Sarsa with shallow and deep learning. *J Data Anal Inf Process* 4:159–176. <https://doi.org/10.4236/jdaip.2016.44014>
- Gasparrini A, Gorini G, Barchielli A (2009) On the relationship between smoking bans and incidence of acute myocardial infarction. *Eur J Epidemiol* 24(10):597–602
- Gilmour S, Degenhardt L, Hall W, Day C (2006) Using intervention time series analyses to assess the effects of imperfectly identifiable natural events: a general method and example. *BMC Med Res Methodol* 6:16
- Goehler A, Geisler BP, Manne JM, Jahn B, Conrads-Frank A, Schnell-Inderst P, Gazelle GS, Siebert U (2011) Decision-analytic models to simulate health outcomes and costs in heart failure: a systematic review. *PharmacoEconomics* 29(9):753–769
- Gómez, V, Thijssen, S, Symington, A, Hailes, S, Kappen, HJ (2016) Real-time stochastic optimal control for multi-agent quadrotor systems. In: Proceedings of the 26th international conference on automated planning and scheduling (ICAPS'16), June 12–17. AAAI Press, London, UK. <https://arxiv.org/pdf/1502.04548.pdf>
- Grundmann O (2014) The current state of bioterrorist attack surveillance and preparedness in the US. *Risk Manag Healthc Policy* 7:177–187
- Heinze-Deml C, Peters J, Meinshausen N (2017) Invariant causal prediction for nonlinear models. <https://arxiv.org/pdf/1706.08576.pdf>
- Ho Y-C, Chu K-C (1972) Team decision theory and information structures in optimal control problems—part I. *IEEE Trans Autom Control* 17(1):15–22
- Höfler M (2005) Causal inference based on counterfactuals. *BMC Med Res Methodol* 5:28
- Hoover KD (2014) Reductionism in economics: causality and intentionality in the microfoundations of macroeconomics. CHOPE Working Paper, No. 2014–03. <https://www.econstor.eu/bitstream/10419/149715/1/chope-wp-2014-03.pdf>
- Ilic MD, Liu S (1996) Hierarchical power systems control: its value in a changing industry. Springer, Heidelberg
- James NA, Matteson DS (2014) ecp: an r package for nonparametric multiple change point analysis of multivariate data. *J Stat Softw* 62(7):1–25
- Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. *J Artif Intell Res Arch* 4(1):237–285. www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/r1-survey.html
- Kalai A, Vempala S (2005) Efficient algorithms for online decision problems. *J Comput Syst Sci* 71:291–307. www.microsoft.com/en-us/research/wp-content/uploads/2016/11/2005-Efficient_Algorithms_for_Online_Decision_Problems.pdf
- Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R (2015) Causal phenotype discovery via deep networks. *AMIA Annu Symp Proc* 2015:677–686
- Kamien MI, Schwartz NL (2012) Dynamic optimization: the calculus of variations and optimal control in economics and management, 2nd edn. Dover Publications, Mineola, NY
- Katt S, Oliehoek FA, Amato C (2017) Learning in POMDPs with monte carlo tree search. In: Proceedings of the 34th international conference on machine learning, Sydney, Australia, PMLR 70. <http://proceedings.mlr.press/v70/katt17a/katt17a.pdf>
- Koller D, Parr R (1999) Computing factored value functions for policies in structured MDPs. In: Proceedings of the sixteenth international joint conference on artificial intelligence (IJCAI99), Stockholm, Sweden, July 31–August 6. Morgan Kaufmann, San Francisco, CA
- Koller D, Milch B (2003) Multi-agent influence diagrams for representing and solving games. *Games Econ Behav* 45(1):181–221. https://ai.stanford.edu/~koller/Papers/Koller+Milch_GEB03.pdf

- Król A, Saint-Pierre P. (2015) SemiMarkov: an R package for parametric estimation in multi-state semi-markov models. *J Stat Softw* 66(5). www.jstatsoft.org/article/view/v066i06
- Lange K, Chi EC, Zhou H (2014) A brief survey of modern optimization for statisticians. *Int Stat Rev* 82(1):46–70
- Lee S, Honavar V (2013) m-transportability: transportability of a causal effect from multiple environments. In: Proceedings of the twenty-seventh AAAI conference on artificial intelligence. www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/viewFile/6303/7210
- Liao L, Ahn H (2016) Combining deep learning and survival analysis for asset health management. *Int J Prognost Health Manag* 7:7. www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2016/ijphm_16_020.pdf
- Lopiano KK, Smith RL, Young SS (2015) Air quality and acute deaths in California, 2000–2012. <https://arxiv.org/abs/1502.03062>
- Luce RD, Raiffa H (1957) Games and decisions: introduction and critical survey. Wiley, New York
- Marschak J, Radner R (1972) Economic theory of teams. Cowles Foundation for Research in Economics at Yale University, Monograph 22. Yale University Press, New Haven
- Morgan MG, Henrion M (1990) Chapter 10 of uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press, New York, reprinted in 1998. www.lumina.com/images/uploads/main_images/Analytica%20A%20Software%20Tool%20for%20Uncertainty%20Analysis%20and%20Model%20Communication.pdf
- Myerson RB (1991) Game theory: analysis of conflict. Harvard University Press, Cambridge, MA
- Ortega PA, Braun DA (2014) Generalized Thompson sampling for sequential decision-making and causal inference. *Complex Adapt Syst Model* 2:2. <https://doi.org/10.1186/2194-3206-2-2>
- Osborne MJ (2004) An introduction to game theory. Oxford University Press
- Patsopoulos NA (2011) A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci* 13 (2):217–224
- Pearl J (2009) Causal inference in statistics: an overview. *Stat Surv* 3:96–146. https://projecteuclid.org/download/pdfview_1/euclid.ssu/1255440554
- Pepels T, Cazenave T, Winands MHM, Lanctot M (2014) Minimizing simple and cumulative regret in Monte-Carlo tree search. In: Cazenave T, Winands MHM, Björnsson Y (eds) Computer Games. CGW 2014. Communications in Computer and Information Science, vol 504. Springer, Cham. www.lamsade.dauphine.fr/~cazenave/papers/PepelsCGW2014.pdf
- Peters J, Bühlmann P, Meinshausen N (2016) Causal inference using invariant prediction: identification and confidence intervals. *J R Stat Soc Ser B* 78(5):947–1012. <https://arxiv.org/abs/1501.01332>
- Ross GJ (2015) Parametric and nonparametric sequential change detection in R: the cpm package. *J Stat Softw* 66(3)
- Ross S, Pineau J, Brahim C, Kreitmann P (2011) Bayesian approach for learning and planning in partially observable markov decision processes. *J Mach Learn Res* 12(2):1729–1770
- Shachter RD, Bhattacharjya D (2010) Solving influence diagrams: exact algorithms. In: Cochran J et al (eds) Wiley encyclopedia of operations research and management science. Wiley, New York. www.it.uu.se/edu/course/homepage/aism/st11/Shachter10.pdf
- Shackleton M, Sødal S (2010) Harvesting and recovery decisions under uncertainty. *J Econ Dyn Control* 34(12):2533–2546. <https://EconPapers.repec.org/RePEc:eee:dyncon:v:34:y:2010:i:12:p:2533-2546>
- Shan G, Pineau J, Kaplow R (2013) A survey of point-based POMDP solvers. *Auton Agent Multi-Agent Syst* 27(1):1–51. <https://doi.org/10.1007/s10458-012-9200-2>
- Shen Y, Cooper GF (2010) A new prior for Bayesian anomaly detection: application to biosurveillance. *Methods Inf Med* 49(1):44–53
- Shoham Y, Leyton-Brown K (2009) Multiagent systems: ALgorithmic, game-theoretic, and logical foundations. Cambridge University Press. www.masfoundations.org/download.html
- Silver D, Veness J (2010) Monte-Carlo planning in large POMDPs. *Advances in Neural Information Processing Systems* 23 (NIPS)

- Simon HA, Iwasaki Y (1988) Causal ordering, comparative statics, and near decomposability. *J Econ* 39:149–173. <http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=34081>
- Simpson KN, Strassburger A, Jones WJ, Dietz B, Rajagopalan R (2009) Comparison of Markov model and discrete-event simulation techniques for HIV. *PharmacoEconomics* 27(2):159–165. <https://doi.org/10.2165/00019053-200927020-00006>
- Sutton RS, Barto AG, Williams RJ (1992) Reinforcement learning is direct adaptive control. *IEEE Control Syst* 12(2):19–22. www.ieeecss.org/CSM/library/1992/april1992/w01-ReinforcementLearning.pdf. Accessed 9 Oct 17
- Taghipour S, Caudrelier LN, Miller AB, Harvey B (2017) Using simulation to model and validate invasive breast cancer progression in women in the study and control groups of the Canadian National Breast Screening Studies I and II. *Med Decis Mak* 37(2):212–223. <https://doi.org/10.1177/0272989X16660711>
- Tartakovsky A, Nikiforov I, Basseville M (2014) Sequential analysis: hypothesis testing and changepoint detection. Chapman and Hall/CRC, Boca Raton
- Thomas LC (2003) Games, theory and applications. Dover Publications, Mineola, NY
- Tolpin D, Shimony S (2012) MCTS based on simple regret. In: Proc. Assoc. Adv. Artif. Intell. pp 570–576. www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/viewFile/4798/5240
- Tsoukalas A, Albertson T, Tagkopoulos I (2015) From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR Med Inform* 3(1):e11. <https://doi.org/10.2196/medinform.3445>
- Van Seijen H, Van Hasselt H, Whiteson S, Wiering M (2009) A theoretical and empirical analysis of expected Sarsa. In: 2009 I.E. symposium on adaptive dynamic programming and reinforcement learning, Nashville, 30 March–2 April 2009, pp 177–184. <https://doi.org/10.1109/ADPRL.2009.4927542>
- Wilson RB (1968) The theory of syndicates. *Econometrica* 36(1):119–132
- Yüksel S, Saldi N (2017) Convex analysis in decentralized stochastic control, strategic measures, and optimal solutions. *SIAM J Control Optim* 55(1):1–28
- White H, Sabarwei S (2014) Quasi-experimental design and methods. UNICEF Office of Research. Methodological Briefs Impact Evaluation No. 8. UNICEF Office of Research—Innocenti, Florence, Italy. https://www.unicef-irc.org/publications/pdf/brief_8_quasi-experimental%20design_eng.pdf
- Zarchan P, Musoff H (2015) Fundamentals of kalman filtering: a practical approach, 4th edn. American Institute of Aeronautics and Astronautics, Reston, VA
- Che Z, Purushotham S, Khemani R, Liu Y (2016) Interpretable deep models for ICU outcome prediction. *AMIA Annu Symp Proc* 2016:371–380
- Zigler CM, Dominici F (2014) Point: clarifying policy evidence with potential-outcomes thinking—beyond exposure-response estimation in air pollution epidemiology. *Am J Epidemiol* 180(12):1133–1140

Chapter 2

Causal Concepts, Principles, and Algorithms



It is an important truism that association is not causation. For example, people living in low-income areas may have higher levels of exposure to an environmental hazard and also higher levels of some adverse health effect than people living in wealthier areas. Yet this observed association, no matter how strong, consistent, statistically significant, biologically plausible, and well documented by multiple independent teams, does not necessarily tell a policy maker anything about whether or by how much a proposed costly reduction in exposure would reduce adverse health effects. Perhaps only increasing income, or something that income can buy, would reduce adverse health effects. Or maybe factors that cannot be changed by policy interventions increase both the probability of living in low-income areas and the probability of adverse health effects. Whatever the truth is about opportunities to improve health by changing policy variables, it typically cannot be determined by studying correlations, regression coefficients, relative risks, or other measures of association between exposures and health effects (Pearl 2009). Observed associations between variables can contain both causal and non-causal ("spurious") components. In general, the effects of policy changes on outcomes of interest can only be predicted and evaluated correctly by modeling the network of causal relationships by which effects of exogenous changes propagate among variables. The chapter reviews current causal concepts, principles, and algorithms for carrying out such causal modeling and compares them to other approaches.

Many different concepts of causality were proposed in the twentieth century and earlier by philosophers (Suppes 1970; Hausman and Woodward 1999), geneticists (Wright 1921), statisticians and social statisticians (Neyman 1923; Campbell and Stanley 1963; Blalock 1964; Rubin 1974); epidemiologists (Robins and Greenland 1992), mathematicians and physicists (Wiener 1956; Schreiber 2000), economists and econometricians (Simon 1953; Granger 1969), artificial intelligence and machine learning researchers, and computer scientists (Charniak 1991; Druzdzel and Simon 1993). They expressed, with varying degrees of rigor and precision, intuitions such as that effects regularly and predictably follow their causes; that

causes make their effects different from what they otherwise would have been; that causes are informative about and help to predict their effects; that expected values or probability distributions for effect sizes can be determined from the values of their causes; and that changing causes changes the probabilities of their effects. By the year 2000, these strands of thought on how to define, measure, and estimate causal relationships and effects had largely been unified in a framework that emphasizes the use of diagrams with nodes representing variables and arrows between nodes representing causal dependencies (Pearl 2009). This framework includes the popular “directed acyclic graph” (DAG) models introduced in Chap. 1, as well as more general models with cycles and undirected arcs (representing dependency with unknown causal direction) allowed. We shall use the DAG models in the following sections.

This chapter explores what it means to say that one thing *causes* another and reviews key ideas about causality that have proved useful in interpreting a broad variety of data and estimating causal impacts of interventions on outcomes. It discusses how to represent different types of causal knowledge using diagrams and mathematical, statistical, and computational models to facilitate explanation, communication, and computation of causal inferences. Finally, this chapter surveys principles and algorithms for using causal models to answer practical questions requiring causal inference. These include questions of attribution and diagnosis, prediction and prognosis, explanation, prescriptive optimization of decisions, and evaluation of their impacts.

Learning goals for this chapter are as follows:

- Distinguish between (a) statistical associations, inferences, and models; and (b) causal models to support/evaluate/improve policy decisions.
- Introduce, explain, and show how to apply several different types or concepts of causality to improve predictions, decisions, and learning. The main types discussed in this chapter are associational, attributive, counterfactual, structural (computational), predictive, manipulative, and mechanistic or explanatory causation.
- Explain the main concepts and software tools currently available to solve causal analytics problems. These include techniques for identifying causal network models from data and for using them to predict, infer, attribute, and explain effects based on observations; optimize decisions; and quantify partial (“direct” and “indirect”) and total causal relationships.
- Introduce algorithms and principles for identifying approximately correct causal models from data using relatively objective (assumption-free, investigator-independent) machine-learning methods where possible, together with knowledge-based constraints where necessary (e.g., that effects do not precede their causes, or that weather can be a cause but not an effect of illnesses).
- Illustrate how to use freely available software for applying causal analytics methods and specific causal discovery and inference algorithms to data. Air pollution health effects research is used as an example for illustrating state-of-the-art causal analytics algorithms.

The chapter is relatively long and introduces many technical concepts and terms needed to take advantage of current causal analytics methods and software. By the end of this chapter, the reader will be conversant with the main ideas and methods of modern causal analytics and will understand their potential and limitations for practical applications in risk analysis. To minimize the burden on readers who are mainly interested in applications, subsequent chapters briefly recapitulate key concepts and techniques where they are used, leaving a fuller exposition of concepts and methods to this chapter. On the other hand, for readers who wish to delve further into the technical methods surveyed in this chapter, an extensive and up-to-date list of references gives access to the primary research literature and to several outstanding surveys, tutorials, and software packages. As in so much of the current practice of data science, the exposition here is targeted mainly at readers who seek to understand technical concepts and methods well enough to use them correctly and effectively and to provide a relatively accessible point of entry to the large and exciting recent technical and research literatures that are transforming how artificial intelligence, machine learning, and data science are being used to learn about cause and effect and to improve understanding and control of the behaviors of a broad range of uncertain systems that affect human health and wellbeing.

Multiple Meanings of “Cause”

The claim that one event or condition *causes* another has meant different things to different people and organizations. Modern causal analysis clarifies these different meanings, allowing more precise expression of what questions a causal study addresses and how the answers should be interpreted. For example, in public and occupational health risk analysis, the causal claim “Each extra unit of exposure to substance X increases rates of an adverse health effect (e.g., lung cancer, heart attack deaths, asthma attacks, etc.) among exposed people by R additional expected cases per person-year” can be interpreted in at least the following ways:

1. *Probabilistic causation* (Suppes 1970): The conditional probability of the health response or effect occurring in a given interval of time is greater among individuals with higher exposures compared to otherwise similar-seeming individuals with lower exposures. In this sense, probability of response (or observed age-specific hazard rate for occurrence of response) increases with observed exposure levels. On average, there are R extra cases per person-year per unit of exposure. The main intuition is that causes (exposures) make their effects (responses) more likely to occur within a given time interval, or increase their occurrence rates—not necessarily in the sense of manipulative causation (discussed below), meaning that an exogenous increase in exposure increases the probability of response, but in the sense that the conditional probability of observing a response is higher when conditioned on observations of higher levels of exposure.

2. *Associational causation* (IARC 2006): Higher levels of exposure have been observed in conjunction with higher risks in one or more studies, although not necessarily in all. This association is judged by selected experts to satisfy desired criteria such as being strong, consistent across multiple studies and locations, and biologically plausible. (Other experts may disagree, and may be correct to do so; the judgments made are usually matters of subjective opinion, rather than objective scientific facts with which all well-informed observers must agree.) The slope of a regression line between estimated historical levels of risk and exposure in the exposed population of interest is R extra cases per person-year per unit of exposure. The main intuition is that causes are positively associated with their effects. Relative risk (RR) ratios—the ratios of responses per person per year in exposed compared to unexposed populations—and quantities derived from them discussed later, such as burden-of-disease metrics (Murray and Lopez 2013), attributable risks, population attributable fractions, probability of causation formulas, and closely related metrics, are widely used in epidemiology and public health to quantify associational causation.
3. *Attributive causation* (Murray and Lopez 2013): Authorities attribute R extra cases per person-year per unit of exposure to X ; equivalently, they blame exposure to X for R extra cases per person-year per unit of exposure. In practice, such attributions are usually made based on measures of association such as the ratio or difference of estimated risks between populations with higher and lower levels of exposure. Differences in risks between the populations are attributed to their differences in exposures without further analysis of other possible explanations. The main idea is that if people with higher exposures have higher risks for any reason, then the increased risk is attributed to the higher exposure. (If many risk factors differ between low-risk and high-risk groups, then the difference in risks can be attributed to each of them separately; there is no consistency constraint preventing multiples of the total difference in risks from being attributed to the various factors.)
4. *Counterfactual and potential outcomes causation* (Höfler 2005; Glass et al. 2013; Lok 2017; Li et al. 2017): In a hypothetical world with 1 unit less of exposure to X , expected cases per person-year in the exposed population would also be less by R . Usually, such counterfactual numbers are derived from modeling assumptions, and reasons for the counterfactual reduction in exposure are not discussed. The main intuition is that differences in causes make their effects different from what they otherwise would be.
5. *Predictive causation* (Granger 1969; Kleinberg and Hripcsak 2011; Papana et al. 2017): Time series data show that the observation that exposure has increased or decreased is predictably followed, perhaps after a lag, by the observation that average cases per person-year have also increased or decreased, respectively, by an average of R cases per unit of change in exposure. The main intuition is that changes in causes help to predict changes in their effects. More generally, causes are informative about their effects, so effects can be predicted better with information about their causes than without it.

6. *Structural causation* (Simon 1953; Simon and Iwasaki 1988; Hoover 2012): In valid mathematical or computational simulation models, the number of cases per person-year is derived at least in part from the value of exposure. Thus, the value of exposure must be determined before the value of yearly case count can be determined. Moreover, the average calculated or simulated value of the case count per person-year decreases by R for each unit of decrease in exposure. The main intuition is that effects depend on, and are calculated from, their causes.
7. *Manipulative causation* (Voortman et al. 2010; Hoover 2012; Simon and Iwasaki 1988): Reducing exposure by one unit reduces expected cases per person-year by R . The main intuition is that changing causes changes their effects. More generally, changing a manipulative cause changes the probabilities of the outcomes that it affects.
8. *Explanatory/mechanistic causation* (Menzies 2012; Simon and Iwasaki 1988): Increasing exposure by one unit causes changes to propagate through a biological network of causal mechanisms. When all changes have finished propagating, the new expected value for case count per person-year in the exposed population will be R more than before exposure was increased. The main intuition is that changes in causes propagate through a network of law-like causal mechanisms to produce changes in their effects. Causal mechanisms are usually represented mathematically by structural equations or by conditional probability tables (CPTs) that are invariant across settings (Pearl 2009).
9. *Producing causation, but-for causation*: A cause (such as exposure) suffices to create a response that would not otherwise have occurred. “Suffices” here means that it suffices in the context of the other exposures and conditions that are present.

For risk managers and policy makers, manipulative causation is key, since the goal of decision-making is to choose acts that cause desired outcomes, in the manipulative sense of making them (or causing them to become) more probable. But-for or producing causation is used primarily in tort law. It requires manipulative causation, since the cause (e.g., exposure) must suffice to produce the effect, which lower or absent exposure would not have sufficed to produce: thus, changing the cause changes the effect (manipulative causation). Manipulative causation is implied by mechanistic causation: if there is a network of mechanisms by which changes in exposure (or other causes) change the probabilities of outcomes (mechanistic causation), then changing the causes will indeed change the probabilities of outcomes (manipulative causation). But neither manipulative causation nor mechanistic causation is implied by associational, attributive, counterfactual, or predictive concepts of causation (Pearl 2009). Understanding and appropriately applying these distinctions among concepts of causation, and making sure that associational concepts are not misrepresented or misunderstood as manipulative causal ones in policy deliberations and supporting epidemiological analyses, provides a crucial first step toward improving current practice in epidemiology (Petitti 1991).

The following sections examine these different concepts of causality more closely and discuss how they are related. *Probabilistic causal models*, which are common to

all of these concepts of causation, are emphasized. In particular, we explain how Bayesian network (BN) models can be used to represent probabilistic dependencies among variables, manipulate probabilities to make predictions, and draw probabilistic inferences. They also provide a useful unifying framework and generalization of many well-known probabilistic risk assessment (PRA) and decision analysis techniques. Modern software makes it relatively easy to build and use BNs. Several examples show how to use current BN software to create simple BN models and use them to draw inferences and make predictions. BN algorithms can also be extended to networks with decisions, i.e., influence diagrams (IDs), and used prescriptively to solve for optimal statistical decisions; additional examples illustrate these methods. The final sections of the chapter consider how to learn causal models from data and conclude with a brief description of selected milestones in the historical development of modern causal analysis and a summary of the main themes and lessons from the chapter.

Probabilistic Causation and Bayesian Networks (BN)

Perhaps the simplest intuition relating probability and causation is that *causes make their effects more probable*. To sharpen this intuition and use it to draw quantitative inferences, it is necessary to be more explicit about how one observation, action, or event can make another more probable. The assumed technical background for this discussion is elementary probability theory, especially the concept of a random variable and the definitions and notations for joint, marginal, and conditional probability distributions.

Technical Background: Probability Concepts, Notation and Bayes' Rule

Uncertain quantities in this chapter are represented by random variables. Most of this chapter assumes that the random variables in question are discrete. The notation $P(x)$ will be used as an abbreviation for the probability that random variable X has specific value x . Thus, $P(x)$ is a short-hand for $P(X = x)$, or, as it is sometimes more explicitly denoted, $P_x(x)$, where the subscript shows the particular random variable for which probabilities of values are being given. $P(x)$ is often called the probability mass function, or, for continuous random variables, the probability density function of the random variable X . When X is just one of several random variables being considered, $P(x)$ is also called its *marginal distribution*. In such a multivariate context, where the particular random variable being referred to might be unclear, the notation $P_X(x)$ for the marginal distribution of X can preserve clarity. We will use the simpler $P(x)$ when the random variable being referred to is clear from context. Likewise, $P(x, y)$ will

denote the *joint probability* that random variable X has specific value x *and* that random variable Y has specific value y ; thus, $P(x, y)$ is short for $P(X = x, Y = y)$ and for the more explicit notation $P_{X,Y}(x, y)$ for the joint probability that $X = x$ and $Y = y$. The *conditional probability* that $X = x$, given that $Y = y$, will usually be written as $P(x | y)$ in preference to the longer and more explicit notations $P(X = x | Y = y)$ or $P_{X|Y}(x | y)$. Recall the definition of conditional probability:

$$P(x|y) = P(x, y)/P(y) \quad (2.1)$$

when the denominator is greater than zero. This definition follows by rearranging the identity

$$P(x, y) = P(y)P(x|y) \quad (2.2)$$

i.e., the probability that both $X = x$ and $Y = y$ is the probability that $Y = y$ times the conditional probability that $X = x$ given that $Y = y$. With equal validity, the joint probability $P(x, y)$ can be factored as a product of a marginal and a conditional probability in a different way, as follows:

$$P(x, y) = P(x)P(y|x) \quad (2.3)$$

The marginal distribution for a random variable X can always be calculated from its conditional probabilities, given each of the values of one or more other variables, and from the marginal probabilities of those values, via the *law of total probability*. This states that the total probability of an event (such as that X has specific value x) is the sum of the probabilities of all of the ways in which it can occur in conjunction with each of a set of mutually exclusive, collectively exhaustive events (such as that Y has each of its possible specific values).

Applying the law of total probability to two random variables X and Y to obtain the marginal distribution of Y from the marginal distribution of X and the conditional probability distributions of Y given each value of X yields the following prediction formula for Y values:

$$P(y) = \sum_x P(y|x)P(x) \quad (2.4)$$

Here, the sum is taken over each of the distinct possible values, x , of X ; if X is a continuous random variable, then the sum must be replaced by an integral. Equating the right-hand sides of Eqs. (2.2) and (2.3), since they both equal $P(x, y)$, yields

$$P(y)P(x|y) = P(x)P(y|x) \quad (2.5)$$

Dividing both sides by $P(y)$ (assuming it is non-zero) gives the identity

$$P(x|y) = P(x)P(y|x)/P(y) \quad (2.6)$$

Then, expanding $P(y)$ via the law of total probability (2.4), yields *Bayes' Rule*:

$$P(x|y) = P(x)P(y|x)/\sum_{x'} P(y|x')P(x') \quad (2.7)$$

(The primes on the x values in the denominator are inserted to make clear that x' is simply an index for the values of X being summed over, not to be confused with the specific, fixed value x in the numerator and on the left side of the equation.) $P(x)$ is called the *prior probability* that $X = x$, and $P(x \mid y)$ is called the *posterior probability* that $X = x$, given the observation or data that $Y = y$. Bayes' Rule allows data on the marginal probabilities of X values and on the conditional probabilities of Y values given X values to be used to infer conditional probabilities of X values given Y values. We assume familiarity with these aspects of probability theory throughout the remainder of this chapter.

Example: Joint, Marginal, and Conditional Probabilities for Answering Queries

Table 2.1 shows the nine joint probabilities for all possible combinations (i.e., pairs) of values for two discrete random variables, each with three possible values: X with possible values 1, 2, and 3; and Y with possible values 4, 8, and 16. Such a joint probability table can be used to answer any question about the probabilities that the values of X and Y fall in specified sets or satisfy specified constraints.

Problem: Use the joint probability distribution in Table 2.1 to answer the following questions.

- What is the probability that $X + Y < 10$?
- What is the probability that $X + Y \leq 10$?
- What is the probability that $Y > 5$?
- What is the marginal probability distribution of X ?
- What is the marginal distribution of Y ?
- What is the conditional probability $P(Y > 5 \mid X + Y < 10)$?

Answers: All answers can be found by summing the joint probabilities in Table 2.1 that satisfy the specified conditions. The marginal probability distribution for a variable is found by summing all the joint probabilities corresponding to each of its values to get the total marginal probability for that value. (These sums are typically shown in the margins of the table, hence the term “marginal” distribution.)

- $P(X + Y < 10) = 0.0 + 0.25 + 0.15 + 0.05 = 0.45$
- $P(X + Y \leq 10) = 0.0 + 0.25 + 0.15 + 0.05 + 0.15 = 0.60$
- $P(Y > 5) = 0.05 + 0.15 + 0.0 + 0.1 + 0.0 + 0.3 = 0.60$

Table 2.1 A joint probability distribution for two random variables: X with possible values 1, 2, and 3; and Y with possible values 4, 8, and 16. For example, $P(x, y) = (3, 16) = 0.3$

		X values		
		1	2	3
Y values	4	0.0	0.25	0.15
	8	0.05	0.15	0.0
16	0.1	0.0	0.3	

- (d) The marginal probability distribution of X is: $P(X = 1) = 0 + 0.05 + 0.1 = 0.15$; $P(X = 2) = 0.25 + 0.15 + 0.0 = 0.40$; $P(X = 3) = 0.15 + 0.0 + 0.3 = 0.45$.
 (e) The marginal distribution of Y is $P(Y = 4) = 0 + 0.25 + 0.15 = 0.40$; $P(Y = 8) = 0.05 + 0.15 + 0.0 = 0.20$; $P(Y = 16) = 0.1 + 0.0 + 0.3 = 0.40$;
 (f) The conditional probability $P(Y > 5 | X + Y < 10) = 0.05/0.45 = 0.11$.

Discussion: This example illustrates a simple principle for how to answer any query about the probability that a condition or assertion holds, given constraints reflecting observations, knowledge, or assumptions about the values of variables on which its truth depends and a prior joint probability distribution for the variables. Let's call a *query* a condition whose truth or falsity can be determined from the values of variables for which a joint distribution is known, and call *data* a set of observations or constraints reflecting whatever is known about the values of these variables. Examples of query conditions in the previous calculations are $X + Y < 10$ in part (a) and $Y > 5$ in part (f). Corresponding examples of data about X and Y to be used in calculating the probabilities of the query conditions are: nothing in part (a) (i.e., only the prior joint probability distribution of the variables is used); and $X + Y < 10$ in part (f). To calculate the probability that a query condition holds, given some data, one can simply sum the probabilities of all cells (i.e., all combinations of values for the variables) in a joint probability distribution such as Table 2.1 that satisfy both the query condition and the constraints given by the data; sum the probabilities of all cells that satisfy the constraints given by the data; and then divide the former by the latter. This corresponds to the following form of Eq. (2.1):

$$P(\text{query}|\text{data}) = P(\text{query and data})/P(\text{data}). \quad (2.8)$$

The required sums can be calculated in one pass from top to bottom of the joint probability table if it is arranged as in Table 2.2. (This is sometimes called the “long” or “molten” form of the data, as opposed to the “wide” data in Table 2.1 (Wickham 2014).) The joint probability for each row, shown in the right-most column of Table 2.2, is included in the cumulative sum for $P(\text{query and data})$ if and only if that row's values of the variables match both the query condition and the data constraint. The joint probability for the row is included in the cumulative sum for $P(\text{data})$ if and only if that row's values of the variables match the data constraint,

Table 2.2 “Long data” form of data in Table 2.1

X value, x	Y value, y	Joint probability $P(x, y)$
1	4	0.0
2	4	0.25
3	4	0.15
1	8	0.05
2	8	0.15
3	8	0.0
1	16	0.1
2	16	0.0
3	16	0.3

data. When all rows have been visited and both cumulative sums have been computed, their ratio $P(\text{query and } \text{data})/P(\text{data})$ provides the value of $P(\text{query} \mid \text{data})$, i.e., the conditional probability that *query* is true given that *data* is true. For example, $P(X > 1 \mid Y < 16)$ could be computed by this algorithm from the joint distribution in Table 2.2 as follows: $P(X > 1 \mid Y < 16) = (0.25 + 0.15 + 0.15 + 0.0)/(0.0 + 0.25 + 0.15 + 0.05 + 0.15 + 0.0) = 0.917$.

In principle, this method provides an oracle for calculating the conditional probability of any *query* condition given any *data*, including user-specified constraints or assumptions as well as observed values of variables. All that is required is the joint probability table and that the *query* condition and the *data* constraints are specified clearly enough in terms of the values of the variables to determine which rows in the joint probability table match each one. This might appear to provide a useful way to construct a probabilistic expert system for answering queries based on data. Such a query-answering system, returning answers of the form $P(\text{query} \mid \text{data})$, could potentially be useful for many applications, including the following, among very many others:

- Medical diagnosis, quantifying $P(\text{disease} \mid \text{symptoms})$;
- Fault diagnosis systems, quantifying $P(\text{fault_condition} \mid \text{test_results})$ for electronic equipment;
- Credit scores quantifying $P(\text{repayment} \mid \text{credit_history})$;
- Forecasts quantifying $P(\text{future_value} \mid \text{history})$. Special cases include weather forecasts such as $P(\text{rain_tomorrow} \mid \text{observations})$ or hurricane predictions, agricultural forecasts, economic forecasts, and many others.
- Reliability models quantifying $P(\text{system will survive for} > T \text{ more years} \mid \text{current age and condition})$;
- Predictive toxicology models quantifying $P(\text{adverse_response} \mid \text{exposure})$ or $P(\text{chemical is a mutagen} \mid \text{chemical properties})$
- Mineral prospecting expert systems quantifying $P(\text{ore grade} \mid \text{observations})$

But in practice, the sheer number of combinations of different values for the different variables makes this direct approach impractical for all but the smallest problems. For example, in a data set with only seven continuous variables such as age, weight, income, daily maximum and minimum temperature, relative humidity, and so forth, each discretized to have only five levels, the joint probability table analogous to Table 2.2 would have $5^7 = 78,125$ rows. If each variable had ten levels, then there would be ten million rows, and if there were ten variables each with ten levels, there would be ten billion rows. Filling out all of the joint probabilities for such a table, or summing probabilities over all rows that match specified conditions to answer queries, would be impractical. Thus, while the basic idea is promising, a much more efficient way is needed to specify and store the joint probability distribution and to use it to compute conditional probabilities in order to create a practical query-answering technology. Bayesian networks (BNs) provide such a more efficient way.

Example: Sampling-Based Estimation of Probabilities from a Database Using R

Instead of seeking to calculate conditional and marginal probabilities from a full joint probability distribution table such as Table 2.2, a more frequent task in practice is to estimate them from available data. Table 2.3 presents an example of a very small database with six records (rows) and four variables (columns other than the Record ID column). Such a database can be used to answer queries about probabilities for variable values (or for predicates and conditions based on them) for a randomly sampled record, as in the following examples.

For a randomly selected record, what is:

- (a) $P(\text{Smoker} = \text{Yes})$ (Answer: $P(\text{Smoker} = \text{Yes}) = 3/6 = 1/2$)
- (b) $P(\text{COPD} = \text{Yes} \mid \text{Smoker} = \text{Yes})$ (Answer: $2/3$)
- (c) $P(\text{Smoker} = \text{Yes} \mid \text{COPD} = \text{Yes})$ (Answer: $2/2 = 1$)

Likewise, the reader can verify the following:

- (d) The probability of COPD for a male smoker is $1/2$;
- (e) The probability that a COPD patient over 50 years old is a male smoker is $1/2$;
- (f) The probability that a randomly selected person in this database is a smoker with COPD is $2/6 = 1/3$.

Although such calculations are simple for a database of only six records, suppose that the database were much larger, e.g., with 60 million records instead of 6. Then answering queries by brute force application of Eq. (2.8), scanning each record and counting the number of records that satisfy various user-specified conditions, would be time consuming. A faster way to get numerically accurate approximate answers would be to randomly sample the records and then use the random samples to compute the requested probabilities.

For readers familiar with R, we illustrate exact and approximate sampling-based calculations for the data in Table 2.3; the same calculations can be applied to much larger data sets. The following R script 1 creates a data table (a data frame named “table”) from the data in Table 2.3, with M and F recoded as 1 and 0, respectively, and with Yes and No recoded as 1 and 0, respectively:

Table 2.3 A small data base

Record ID	Gender	Age	Smoker	COPD
1	M	31	Yes	No
2	F	41	No	No
3	F	59	Yes	Yes
4	M	26	No	No
5	F	53	No	No
6	M	58	Yes	Yes

```
# R script 1: Create data table, store as a data frame named "table"
  data <- c(1, 31, 1, 0, 0, 41, 0, 0, 0, 59, 1, 1, 1, 26, 0, 0, 0, 53, 0,
  0, 1, 58, 1, 1)
  table <- matrix(data, nrow=6, byrow = T)
  colnames(table) <- c('Gender', 'Age', 'Smoker', 'COPD')
  rownames(table) <- c(1:nrow(table))
  table <- as.data.frame(table)
```

The table looks like this in R Studio:

```
> table
   Gender Age Smoker COPD
1      1   31      1    0
2      0   41      0    0
3      0   59      1    1
4      1   26      0    0
5      0   53      0    0
6      1   58      1    1
>
```

Given this table, answers to the previous probability questions about a randomly selected record can be calculated via the following R script:

```
# R script 2: Calculate marginal and conditional probabilities using P(A | B) = nrow(A & B)/nrow(B)
attach(table)
# Question a: P(Smoker)
nrow(subset(table, Smoker == 1))/nrow(table)
# Question b: P(COPD | Smoker)
nrow(subset(table, COPD == 1 & Smoker == 1))/nrow(subset(table, Smoker == 1))
# Question c: P(Smoker | COPD)
nrow(subset(table, COPD == 1 & Smoker == 1))/nrow(subset(table, COPD == 1))
# Question d: P(COPD | male smoker)
nrow(subset(table, COPD == 1 & Smoker == 1 & Gender == 1))/nrow(subset(table,
Smoker == 1 & Gender == 1))
# Question e: P(male smoker | COPD patient over 50)
nrow(subset(table, Smoker == 1 & Gender == 1 & COPD == 1 & Age >
50))/nrow(subset(table, COPD == 1 & Age > 50))
# Question f: P(smoker with COPD)
nrow(subset(table, Smoker == 1 & COPD == 1))/nrow(table)
detach(table)
```

The template is the same for each of these: to calculate $P(query | data)$, just count the number of records (using the “`nrow`” command) satisfying the *query* and *data* conditions (specified using “`==`” to show that a variable has a certain specific value, or using comparison operators such as “`>`” to select ranges of values, and using “`&`” between conjoined conditions). Then divide by the number of records satisfying the *data* conditions. The ratio is the desired conditional probability, $P(query | data)$.

Running R script 2 in R Studio produces the following output, in agreement with the previous manual calculations:

```

> attach(table)
The following objects are masked from table (pos = 3):
  Age, COPD, Gender, Smoker
> # Question a: P(Smoker)
> nrow(subset(table, Smoker == 1))/nrow(table)
[1] 0.5
> # Question b: P(COPD | Smoker)
> nrow(subset(table, COPD == 1 & Smoker == 1))/nrow(subset(table, Smoker == 1))
[1] 0.6666667
> # Question c: P(Smoker | COPD)
> nrow(subset(table, COPD == 1 & Smoker == 1))/nrow(subset(table, COPD == 1))
[1] 1
> # Question d: P(COPD | male smoker)
> nrow(subset(table, COPD == 1 & Smoker == 1 & Gender == 1))/nrow(subset(table,
Smoker == 1 & Gender == 1))
[1] 0.5
> # Question e: P(male smoker | COPD patient over 50)
> nrow(subset(table, Smoker == 1 & Gender == 1 & COPD == 1 & Age > 50))/nrow(su
bset(table, COPD == 1 & Age > 50))
[1] 0.5
> # Question f: P(smoker with COPD)
> nrow(subset(table, Smoker == 1 & COPD == 1))/nrow(table)
[1] 0.3333333
> detach(table)

```

Such calculations run quite quickly on data sets of several thousand rows and several hundred columns.

Almost the same script can be used to answer questions by sampling rows instead of exhaustively counting how many satisfy the specified conditions. The only change is that instead of creating a table that holds all the data, we create one that holds a random sample of the data. Randomly sampling 100,000 rows using simple random sample with replacement suffices to estimate probabilities to two significant digits. This can be done by running R script 3.

```
# R script 3: Answer questions using random sample of size 10,000
sample <- table[sample(1:nrow(table), 100000, replace = TRUE),]
```

The first few records in this sample are as follows:

```

> head(sample)
   Gender Age Smoker COPD
6       1  58      1    1
6.1     1  58      1    1
4       1  26      0    0
2       0  41      0    0
1       1  31      1    0
6.2     1  58      1    1

```

(Note that record numbers such as 6.1 and 6.2 indicate repeated samples of the same record from the original data set—record 6, in this case.) Overwriting the original data table with the newly created sample using the assignment command “table <- sample” (an admittedly sloppy practice, as one should in general avoid

overwriting original data sources, but adequate for purposes of illustration) and then rerunning R script 2 produces the following output:

```
> attach(table)
The following objects are masked from table (pos = 3):
  Age, COPD, Gender, Smoker
> # Question a: P(Smoker)
> nrow(subset(table, Smoker == 1))/nrow(table)
[1] 0.49704
> # Question b: P(COPD | Smoker)
> nrow(subset(table, COPD == 1 & Smoker == 1))/nrow(subset(table, Smoker == 1))
[1] 0.6650773
> # Question c: P(Smoker | COPD)
> nrow(subset(table, COPD == 1 & Smoker == 1))/nrow(subset(table, COPD == 1))
[1] 1
> # Question d: P(COPD | male smoker)
> nrow(subset(table, COPD == 1 & Smoker == 1 & Gender == 1))/nrow(subset(table,
Smoker == 1 & Gender == 1))
[1] 0.4983275
> # Question e: P(male smoker | COPD patient over 50)
> nrow(subset(table, Smoker == 1 & Gender == 1 & COPD == 1 & Age > 50))/nrow(su
bset(table, COPD == 1 & Age > 50))
[1] 0.5002269
> # Question f: P(smoker with COPD)
> nrow(subset(table, Smoker == 1 & COPD == 1))/nrow(table)
[1] 0.33057
> detach(table)
```

These sampling-based answers are obtained quickly (less than a second to execute R script 2 with 100,000 records in the sample) and they are sufficiently accurate approximations for many purposes. Even if the original data table had contained many millions of records, a simple random sample of size 100,000 records would suffice to answer such questions with this level of speed and accuracy.

This example illustrates how to use random sampling of records to answer questions about the conditional probability that a target condition holds for a randomly selected record in a data set, given some other observed or assumed conditions (with all conditions expressed as constraints on values of the variables). The same idea can be used if the data-generating process is well enough understood so that it can be simulated. Each simulation run produces, in effect, a synthetic record for a potential database consisting of sample values drawn from the joint distribution of the variables. Creating a large number of runs provides a sample that can be used to make probability calculations and answer questions using the techniques just illustrated.

Example: Bayes' Rule for Disease Diagnosis

Setting: Suppose that 1% of the population using a medical clinic have a disease such as HIV, hepatitis, or colon cancer. A medical test is available to help detect the disease. Its statistical performance characteristics are as follows:

$$P(\text{test is positive} \mid \text{disease is present}) = 1 - \text{probability of false negative} = 0.99.$$

$$P(\text{test is positive} \mid \text{disease is not present}) = \text{probability of false positive} = 0.02.$$

Thus, the probability of a false positive is 2% and the probability of a false negative is 1%.

Problem: Suppose that a randomly selected member of the population is given the medical test and the result is positive. Then what is the probability that the individual has the disease?

Solution via Bayes' Rule: Before continuing, the reader may wish to make an intuitive judgment about the probability that the subject has the disease, given the above information (i.e., 1% base rate of the disease in the population, sensitivity of test = 99%, specificity of test = 98%, where sensitivity = true positive rate and specificity = true negative rate). Most people judge that the posterior probability of the disease, given a positive test result, is quite high, since the test has high sensitivity and specificity. The quantitative answer can be calculated via Bayes' Rule (Eq. 2.7), which is repeated here for convenience:

$$P(x|y) = P(y|x)P(x)/\sum_{x'} P(y|x')P(x') \quad (2.7)$$

Applying this rule to the disease data yields the following:

$$\begin{aligned} P(\text{disease}|\text{test positive}) &= P(\text{test positive}|\text{disease})P(\text{disease}) / [P(\text{test positive}|\text{disease})P(\text{disease}) + P(\text{test positive} \mid \text{no disease})P(\text{no disease})] \\ &= 0.99 * 0.01 / (0.99 * 0.01 + 0.02 * 0.99) = 1/3. \end{aligned}$$

Discussion: The posterior probability of disease, given a positive test result, is only 1/3, despite the high sensitivity and specificity of the diagnostic test, because of the low base rate of the disease (only 1%) in this population. Intuitive judgments often neglect the importance of the base rate, and hence over-estimate the posterior probability of disease given the test results.

Bayesian Network (BN) Formalism and Terminology

The assertion “ X causes Y ” or “ X is a cause of Y ” can be made for different kinds of entities X and Y , including events that occur at specific moments; conditions or attribute values or states that persist over an interval of time; transition rates between states, or rates of change of variables; values of variables at specific moments; and time series, histories or sequences of values of time series variables. In the Bayesian network (BN) formalism introduced in this section, the entities between which causal relationships might hold are *random variables*. These are flexible enough to represent events, conditions, states, and deterministic quantities as special cases. The values of the variables can be binary indicators with values 1 if an event occurs or a condition holds and 0 otherwise. They can be names or numbers for the possible

values of state variables, if states are discrete; or they can indicate values of continuous attributes, states, or uncertain quantities, or of a known, deterministic quantity if all probability mass is concentrated on a single value. The variables corresponding to the columns in a data frame or table such as Table 2.3 are usually modeled as random variables, with the specific value in a particular row (representing a *record* or *case*) and column (representing a variable) being thought of as a sample value drawn from a conditional probability distribution that may depend on the values of some or all of the other variables for that case.

A BN consists of *nodes*, representing the random variables, and *arrows* between some of the nodes, representing statistical dependencies between random variables. An input node, meaning one with no arrows pointing into it, has an unconditional probability distribution for its possible values; this is also called its *marginal probability distribution*, as previously noted. A node with one or more arrows pointing into it from other nodes has a *conditional probability distribution* for its possible values that depends on the values of the variables (nodes) pointing into it. The nodes that point into a given node are called its *parents*; the parents of node or variable X in a BN are often denoted by $Pa(X)$.

When each random variable has only a few possible values, the conditional probability distributions for each combination of a node's parents' values can be tabulated explicitly. The resulting data structure is called a *conditional probability table* (CPT) for that node. If a node represents a continuous variable, then its conditional probability distribution, given the values of its parents, may be described by a regression model instead of by a table. Alternatively, modern BN software often discretizes continuous random variables, e.g., rounding age to the nearest year. A coarser discretization (e.g., 5-year age buckets) can be used if none of the other variables is very sensitive to age, in the sense that finer-grained discretization does not change their conditional probability distributions significantly compared to using the coarser-grained description. Algorithms for automatically discretizing continuous variables (e.g., vector quantization or classification and regression tree (CART) algorithms based on information theory) are well developed and are incorporated into some BN software packages, but a simpler approach that is also widely applied is to use the ten deciles of the empirical frequency distribution of each continuous variable to represent it as a discrete random variable. We shall use the term “conditional probability” and its abbreviation “CPT” generically to refer to any representation—whether an actual table, a regression model, a CART tree, a random forest ensemble (discussed later), or some other model—that specifies the conditional probability distribution for the value of a variable given the values of its parents.

The arrows in a BN are oriented to form a *directed acyclic graph* (DAG), i.e., a graph or network with directed arcs that do not form directed cycles, so that no variable is its own parent or its own ancestor (where an ancestor is defined recursively as a parent of a parent or a parent of an ancestor). The BN's DAG structure represents a way to decompose the joint probability distribution of the variables into marginal probability distributions for the input nodes and conditional probabilities for other nodes. For example, the two-node BN

$$X \rightarrow Y \quad (2.9)$$

has only one input variable, X , and one output variable, Y . Input X has a marginal probability distribution $P(x)$ and output Y has a CPT, $P(y | x)$. Thus, the BN (2.9) decomposes the joint probability distribution for the pair of variables X and Y as shown in Eq. (2.3):

$$P(x, y) = P(x)P(y|x) \quad (2.3)$$

Conversely, BN (2.10) represents the decomposition in Eq. (2.2):

$$X \leftarrow Y \quad (2.10)$$

$$P(x, y) = P(y)P(x|y) \quad (2.2)$$

Both decompositions represent the same joint probability distribution, but which is more useful depends on what can be measured and what questions are to be answered. To predict conditional probabilities of Y values from measured values of X , we would use BN (2.9). The prediction comes straight from the CPT for Y , $P(y | x)$. If x values are not measured, but are known to be described by (or “drawn from”) a marginal probability distribution $P(x)$, then the prediction formula for y values is given by Eq. (2.4):

$$P(y) = \sum_x P(y|x)P(x) \quad (2.4)$$

The probability of y is just the probability-weighted average value of $P(y | x)$, weighted by the probabilities for each x value. For example, in the previous example of disease diagnosis, the predictive probability that a randomly selected patient will have a positive test result is

$$\begin{aligned} P(\text{positive test}) &= P(\text{positive test}|\text{disease})P(\text{disease}) \\ &\quad + P(\text{positive test}|\text{no disease})P(\text{no disease}) \\ &= 0.99 * 0.01 + 0.02 * 0.99 = 0.0297. \end{aligned}$$

Conversely, to assess the probability of disease given an observed test result, one would use BN (2.10) and Bayes’ rule (2.7), as already illustrated.

These ideas extend to BNs with more than two nodes. A joint distribution of n random variables can be factored by generalizing the idea that “joint = marginal x conditional” for two variables, as follows:

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \dots P(x_n|x_1, x_2, \dots, x_{n-1}) \quad (2.11)$$

To apply this to a DAG model, the variables are numbered so that each variable is listed only after all of its parents (if any) have been listed. The input nodes in the DAG appear first in the list of variables, (x_1, x_2, \dots, x_n) . (Such an ordering of the nodes is always possible for a DAG, and it can be produced by running a topological sort algorithm on the DAG.) The key insight that makes BNs so useful is that most of

the terms in the decomposition (2.11) simplify for sparsely connected DAGs, since usually the set of parents of a node, $Pa(X)$, on which its CPT depends is only a (possibly small) subset of all of the variables that have been listed so far. For example, the joint probability distribution $P(x_1, x_2, x_3, x_4)$ of the variables in the DAG model

$$X_1 \rightarrow X_2 \rightarrow X_3 \leftarrow X_4$$

can be factored as follows:

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_4)P(x_2|x_1)P(x_3|x_2, x_4) \quad (2.12)$$

If each of the four variable has two levels, then it requires two numbers (summing to 1, so only one independent number, or degree of freedom) to specify the marginal distribution for X_1 ; another independent number to specify the marginal distribution for X_4 ; 2 independent numbers for the CPT $P(x_2 | x_1)$, one for each level of x_1 ; and 4 numbers for the CPT $P(x_3 | x_2, x_4)$, one for each combination of values for x_2 and x_4 . Thus, a total of 8 independent numbers must be provided as input values to fully determine the joint probability distribution of the four variables using Eq. (2.12). By contrast, the general identity (2.11), which for $n = 4$ is

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)P(x_4|x_1, x_2, x_3)$$

requires considering 1 independent number to specify $P(x_1)$, 2 more for $P(x_2 | x_1)$, 4 more for $P(x_3 | x_1, x_2)$, and 8 for $P(x_4 | x_1, x_2, x_3)$, for a total of 15, in keeping with the fact that the joint probability distribution of n binary variables requires $2^n - 1$ independent numbers to specify in the most general case—one probability number for each possible combination of values, with the constraint that the numbers must sum to 1. Thus, the DAG model represented by Eq. (2.12) requires about half (8/15) as many numbers to fully specify as would be true in the worst case. For DAG models with more variables, the potential savings in specifying the joint distribution using a factored representation in which each variable's CPT depends only on its parents can be dramatic. A fairly sparse DAG with n binary nodes and only a few parents for each node—say, no more than k parents, for some small integer k —requires fewer than $n*2^k$ independent input numbers to specify the joint distribution by means of probability tables representing marginal distributions and CPTs, instead of $2^n - 1$ numbers to list the full joint distribution table. If $n = 12$ and $k = 3$, for example, then the DAG representation requires fewer than 100 input numbers for the probability tables, compared to over 4000 for the joint distribution table.

Using BN Software for Probability Predictions and Inferences

BN software products and methods allow the following standard approach to formulating and solving probabilistic inference problems for BNs with any number of nodes.

1. Create a BN consisting of a node for each random variable and a DAG (directed acyclic graph) in which arrows between variables represent dependencies between them.
2. Specify a marginal probability distribution for each input node.
3. Specify a CPT for each node with an arrow pointing into it.
4. Enter observations or assumptions (sometimes referred to generically as “findings”) about the values of some of the variables.
5. Obtain the conditional (posterior) distributions of all other variables, conditioned on the findings entered by the user. BN solver software packages automatically calculate these updated distributions.

The following example illustrates this process for the disease diagnosis example previously solved via Bayes’ rule, using the BN software package Netica (downloaded from www.norsys.com/netica.html) to perform the calculations.

Example: A Two-Node BN for Disease Diagnosis Using Netica

Setting: Suppose again that 1% of the population using a medical clinic have a disease and that a diagnostic test for the disease has these statistical performance characteristics:

$P(\text{test is positive} \mid \text{disease is present}) = \text{test sensitivity} = 0.99$.

$P(\text{test is negative} \mid \text{disease is not present}) = \text{test specificity} = 0.98$.

Problem: What is the probability that an individual has the disease, given a positive test result?

Solution via Netica Bayesian Network Solver: The Netica software can be downloaded for free from www.norsys.com/download.html. After running the installation package, double-click on Netica.exe in the folder to which it has been downloaded to open the software. When prompted, click on “Limited Mode” to run the free version for small problems. Under *File*, select *New* and then *Network* to open a new network (Fig. 2.1):

Starting with a blank network, select a “Nature Node” (yellow oval) from the Netica network drawing toolbar and then click on the blank area to create a node: it will appear as a rectangular box with a title at the top. By default, Netica will label this “Node A.” (“Nature Node” is Netica’s term for a random variable.) Click again to create a second node, which will automatically be labeled “Node B.” Join the nodes by an arrow using the “Add Link” arrow in the Netica toolbar. Double-clicking on each of these newly created nodes will let its name be edited and values or names for its possible states be entered.

Dialogue boxes will open that let the user click on “New” to create new values for the “States” (possible values) of a node, and that also allow a new name for the node to be entered. Rename nodes A and B as “Disease_state” and “Test_result” and create and name two possible states for each: “Yes” for disease present and “No” for disease not present; and “Positive” for positive test result and “Negative” for

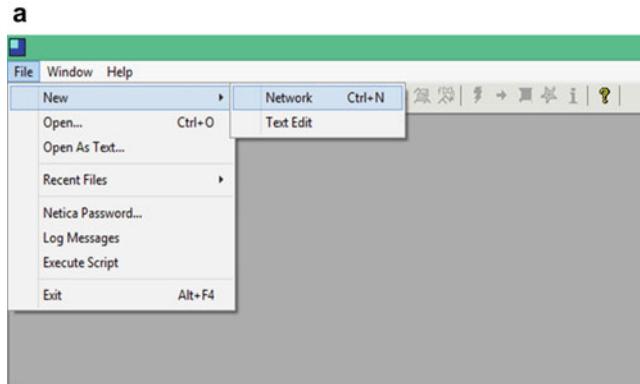
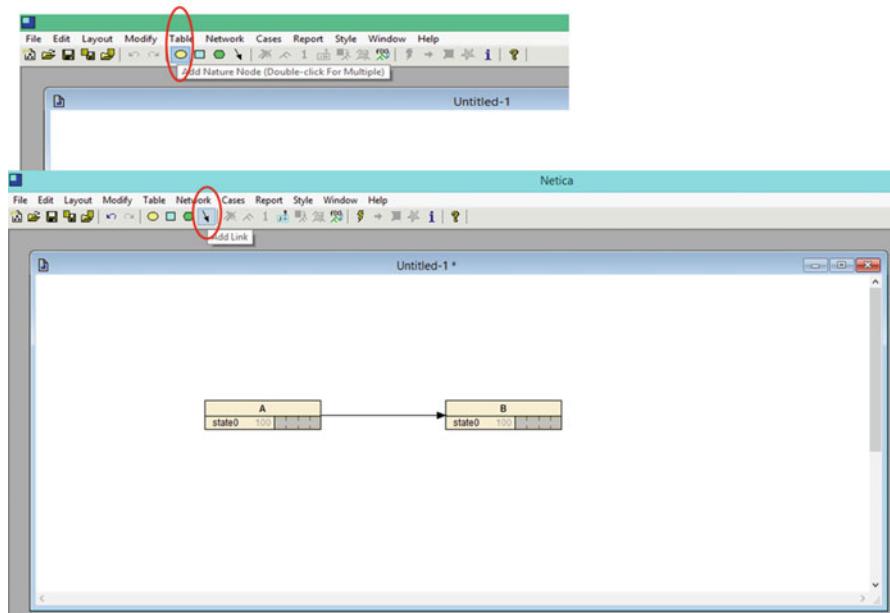
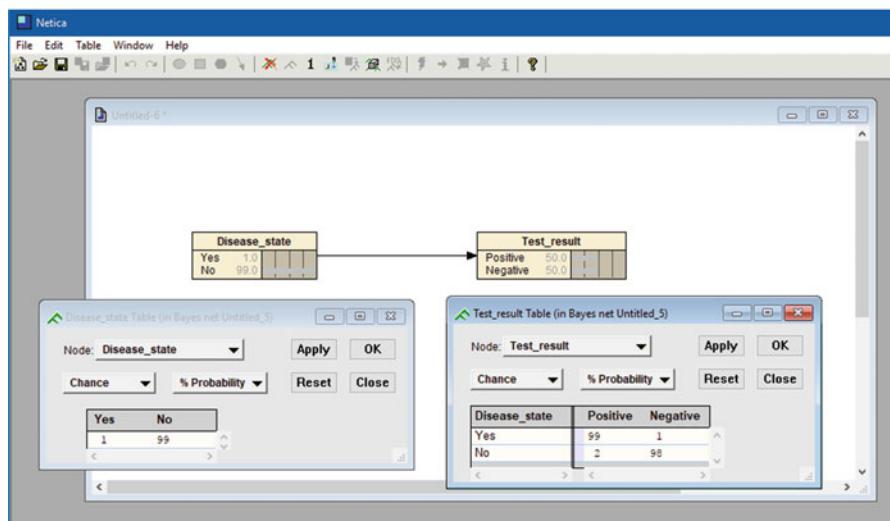


Fig. 2.1 (a) Creating a new Netica network. (b) Adding nodes to the newly created network and linking them by an arrow. (c) DAG model drawn, with the node probability tables filled in, using Netica. (d) Complete Netica model, ready to use to draw inferences. (e) Netica model with the finding “Disease_state = Yes” entered by clicking on “Yes” in the node for Disease_state. (f) Netica model with the finding “Test_result = Positive” entered by clicking on “Positive” in the node for Test_result

negative test result, respectively. Clicking on “OK” in the dialogue box makes these changes, which henceforth will be displayed in the rectangle for each node. This completes the construction of the two-node DAG. The whole model can now be named and saved (using *Save As* under *File*); we will call the model “Disease.” To complete the model, it is necessary to fill in a probability table for each node. Double-clicking on a node opens a dialogue box that contains “Table” as a button, and clicking on this button opens another dialogue box into which the probability table for the node can be entered. For an input node such as *Disease_state*, this probability table is the marginal probability distribution of the states (or values) of the node. For a node with incoming arrows, such as *Test_result*, the probability table is the node’s CPT.

When the probability tables have been specified, click on “OK” in each open dialogue box to make these changes and close the dialogue boxes. The data have now been entered in the DAG, but the DAG display will still be displaying default values—the marginal distribution for *Disease_state* (1.0% for Yes and 99.0% for No, and a uniform distribution such as 50.0% for Positive and 50.0% for Negative for *Test_result*, assuming that the “% Probability” display is being used. This is the default option in the dialogue box for the node. A decimal probability can be displayed instead if desired.) To start using the network to make calculations, select “Compile” under the “Network” drop-down menu on the main Netica toolbar. The nodes of the model will now start displaying correct probability values and corresponding bar charts. For example, if no findings are entered, then the probability for a Positive test result will be displayed as 2.97%.

This agrees with the manual calculation using predictive probability Eq. (2.4),
 $P(\text{positive test}) = P(\text{positive test} \mid \text{disease})P(\text{disease}) + P(\text{positive test} \mid \text{no disease})$

b**c****Fig. 2.1** (continued)

$P(\text{no disease}) = 0.99 \times 0.01 + 0.02 \times 0.99 = 0.0297 = 2.97\%$. On the other hand, entering a finding that the patient has the disease will increase the probability of a positive test to 99%. The easiest way to do this in Netica is simply to click on "Yes" in the *Disease_state* node.

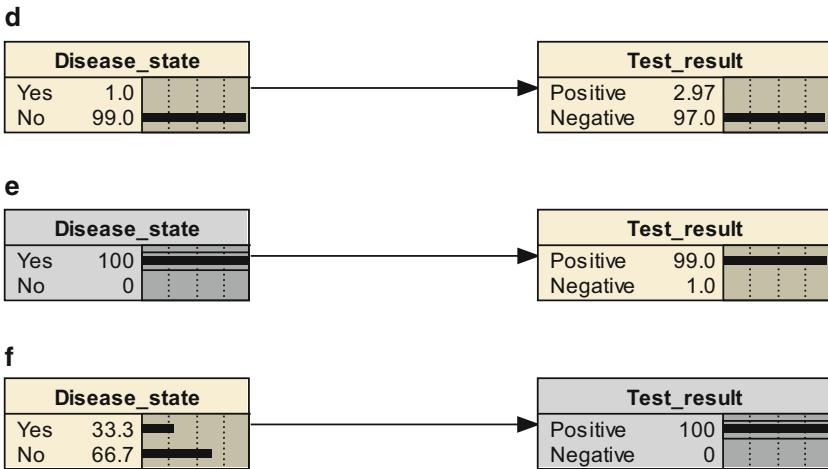


Fig. 2.1 (continued)

Retracting this finding (by clicking on the blank space in the *Disease_state* node or by right-clicking on the node and selecting “Unknown” under the “Finding” option that then appears on a menu) restores the model to its original state, ready to process another set of findings. Clicking on “Positive” in the *Test_result* node will then result in the posterior probability distribution for the *Disease_state* node being displayed. There is a 1/3 probability, displayed as 33.3%, that the patient has the disease, given a positive test result, as calculated previously via Bayes’ rule.

This example has illustrated the steps of building a DAG model, populating its probability tables, and using it to draw probabilistic inferences in the simplest case of only two nodes (random variables), each with only two values. The same steps can be followed to build and use much larger BNs.

Example: Bayesian Inference in a Small BN—The Family Out Problem

Setting: The following example is based on Charniak (1991), which provides a nice introductory tutorial on BNs. Suppose that I am driving home and, as I approach my house, I see that the outdoor light is on, and I wonder whether my family is at home. Figure 2.2a shows a BN for what I know about the relationships among five variables, including the light being on. In the absence of other information, there is a prior probability of 15% that the family is out. If the family is out, then there is a 60% probability that the outdoor light will be on; otherwise, there is a 5% probability that the light is on. (Thus, the prior probability of the light being on is $P(\text{family out}) \cdot P(\text{light on} \mid \text{family out}) + P(\text{family in}) \cdot P(\text{light on} \mid \text{family in}) = 0.15 \cdot 0.60 + 0.85 \cdot 0.05 = 0.1325$.) The probability that I hear the dog bark is 70% if the dog is out and 1% if it is in. The dog may be out because it has a bowel problem, but this has a prior probability of only 1%. The probability that the dog is out is 99% if it has

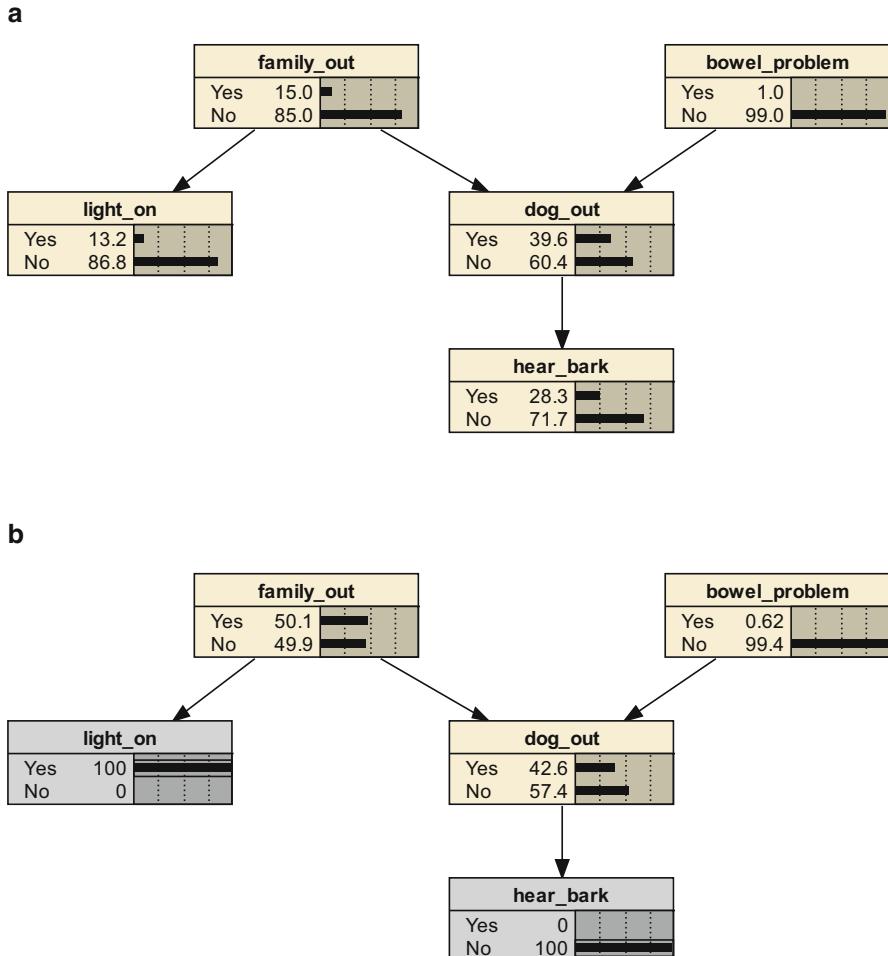


Fig. 2.2 (a) A 5-node Bayesian network for the family out problem. (b) Propagation of evidence through the 5-node Bayesian network

a bowel problem and the family is out; 90% if there is no bowel problem but the family is out; 97% if there is a bowel problem but the family is in; and 30% if there is no bowel problem and the family is in.

Problem: If I observe that the light is on but I do not hear the dog bark, then what is the probability that the family is out?

Solution using Netica: Figure 2.2b shows the solution using Netica. The evidence of the light being on and no bark being heard has been entered as two findings (shaded nodes). The posterior probability that the family is out is 50.1%. Note that the probability that the dog is out has increased from 39.6% in the absence of evidence to 42.6% given the evidence of the light being on but no bark being

heard. This shows that that evidence of the light being on increases the probability of the dog being out more than the evidence of no bark being heard reduces it.

These examples illustrate that BN software packages such as Netica make it relatively easy to assemble DAG models representing factored joint distributions of variables. Several popular packages are available for this purpose (e.g., Bayesia, BayesiaLab, Bayes Server™, HUGIN, GeNIE and SMILE; see www.kdnuggets.com/software/bayesian.html). To support quantitative inference, each node in a BN model has a corresponding *probability table*. For an input node (i.e., a node with no arrows directed into it), the probability table is the prior marginal probability distribution; for any other node, the probability table is its CPT. Once a BN model has been created by specifying a DAG structure and a probability tables for each node, it can be used to draw inferences and to answer questions for which the answers are expressed as posterior probabilities for the values of variables. Given a user-specified set of observed or assumed values (“findings”) for some variables, special algorithms calculate the posterior marginal distributions for the other variables, conditioned on the findings. These are the same distributions that would be obtained by explicitly storing the full joint probability table, as in Table 2.2, and then using it first to condition on the findings, by discarding rows that do not match them, and then to calculate the marginal distributions of each remaining variable by summing the probabilities for each of its values over all combinations of the other variables. However, the BN representation and calculations are far more efficient than this brute-force procedure.

Inference algorithms used by BN software to calculate posterior distributions conditioned on findings are sophisticated and mature (Koller and Friedman 2009; Peyrard et al. 2015). They include both exact methods and Monte Carlo simulation-based methods for obtaining approximate solutions with limited computational effort. Exact inference algorithms (such as variable elimination, a type of dynamic programming) exploit the factorization of the joint distribution revealed by the DAG structure to calculate and store factors that are then used multiple times in calculating posterior probabilities for variables connected by paths in the DAG. Chapter 9 of Koller and Friedman (2009) presents variable elimination in detail. A simple class of algorithms for approximate inference is based on Gibbs sampling, a special case of Markov Chain Monte Carlo sampling. The main idea is that values of variables specified by “findings” (user-input assumptions or observations) are held fixed at the user-specified values. Values for other input variables are then randomly sampled from their marginal distributions. Values for other variables are successively sampled from the appropriate conditional distributions specified in their CPTs, given the values of their parents. Repeating many times yields samples drawn from the joint distribution of all the variables. As discussed in the example for Table 2.3, a sample size from the joint distribution that is very manageable with modern software (e.g., 100,000 samples) suffices to calculate approximate probabilities for answering queries to about two significant digits. Specialized sampling methods (such as likelihood-weighted sampling and related importance sampling methods) can be used to improve sampling efficiency for rare events. Thus, BNs provide effective

methods for representing joint distributions and for using them to calculate posterior distributions conditioned on findings for a wide range of practical applications.

Both exact and approximate BN inference methods can also find the most probable values for unobserved variables (called their “maximum *a posteriori*” or MAP values) given the findings for observed (or assumed) values of evidence variables, with about the same computational effort needed to calculate posterior probabilities. This enables BN packages such as Netica to generate *most probable explanations* (MPEs) for findings. These are defined as values for the unobserved (non-evidence, non-finding) variables that maximize the probability of the findings. For example, in the family out example, the MPE for the findings in Fig. 2.2b (light on, no bark heard) can be found by selecting the “Most Probable Expl” option under “Network” on the Netica toolbar. Netica displays the MPE by setting to 100% the probability bars for the most probable values of the non-evidence nodes (those with no findings entered). For the evidence in Fig. 2.2b, the MPE is that the family is out, the dog has a bowel problem, and the dog is out.

Part 2 of Koller and Friedman (2009) presents details on inference algorithms for BNs, including computation of MAP/MPE explanations. For many practitioners, it suffices that available BN software includes well-engineered algorithms to quickly produce accurate answers for networks with dozens to hundreds of nodes. Examples of deployed applications include a BN with 671 nodes and 790 edges for automatic fault diagnosis in electrical power networks (Mengshoel et al. 2010) and a BN with 413 nodes and 602 edges for inference about individual metabolism and needed customized adjustments of insulin doses for individual diabetes patients (Andreassen et al. 1991; Tudor et al. 1998). Cossalter et al. (2011) discuss the challenges of visualizing and understanding networks with hundreds of nodes.

Practical Applications of Bayesian Networks

The Bayesian network technology explained so far is mature, and BNs have been successfully applied in hundreds of practical applications in recent decades. The following examples illustrate their wide range of applications:

- *Dispute resolution and facilitation of collaborative risk management.* BNs have been proposed to help importers and exporters agree on how to jointly manage the risks of regulated agricultural pests (e.g., fruit flies) along supply chains that cross national borders (Holt et al. 2017). A BN provides a shared modeling framework for discussing quantitative risks and effects of combinations of interventions at different points along the supply chain. Similarly, Wintle and Nicholson (2014) discuss the use of BNs in trade dispute resolution, where they can help to identify specific areas of disagreement about technical uncertainties.
- *Microbial risk assessment for food manufacturers* (Rigaux et al. 2013). In this application, prior beliefs about microbial dynamics in a food production chain are summarized as a BN and measurements of microbial loads at different points in

the production chain are used to update beliefs, revealing where unexpected values are observed and hence where prior beliefs may not accurately describe the real world. Such comparisons of model-predicted to observed values, followed by Bayesian updating of beliefs and, if necessary, revision of the initial BN model, and keys to model validation and improvement light of experience.

- *Land use planning* for hazardous facilities such as chemical plants, taking into account the potential for domino effects, or cascading failures, during major industrial accidents (Khakzad and Reniers 2015).
- *Monitoring of vital signs* in homecare telemedicine (Maglogiannis et al. 2006).
- *Safety analysis of hazards*, such as underground buried pipelines, during river tunnel excavation (Zhang et al. 2016). In this and many other applications to risk, reliability, and safety analysis, practitioners often find it convenient to condition the variables in a BN on fuzzy descriptions of situations (e.g., “not very deeply buried”) when precise quantitative data are not available. BNs have also been developed for risk assessment and risk management of underground mining operations, bridge deterioration and maintenance, fire prevention and fire-fighting (for both wildfires and building fires), collision avoidance between ships in crowded ports and waterways, and safe operation of drones and autonomous vehicles.
- *Occupational safety in construction projects* (Leu and Chang 2013). BNs can identify site-specific safety risks and their underlying causes and probabilities, thus helping to prioritize costly interventions to reduce accident frequencies.
- *Reliability, fault diagnosis, and operations management*. BNs have been applied to reliability analysis and related areas such as real-time fault diagnosis, local load optimization, and predictive maintenance of complex engineering infrastructures from electric power networks to high-speed trains to dams.
- *Cyber security, fraud detection, and counterterrorism*. Detecting attacks, managing accounts with suspected but not proven fraudulent activity, and understanding contagion and systemic financial risks are increasingly popular areas for BNs.

Non-causal Probabilities: Confounding and Selection Biases

Armed with basic probability theory and Bayesian networks (BNs) for manipulating joint, marginal, and conditional probabilities, it is easy to recognize both strengths and weaknesses of the idea that occurrence of a cause makes its effects more probable. Figure 2.3 shows the BN from Fig. 2.2, about a disease and a test result, and illustrates the effects of two different findings.

The top BN in Fig. 2.3 shows the prior BN, before any findings are entered. The predictive probability of a positive test result is 2.97%. The middle BN shows that the presence of disease in an individual (indicated by Disease_state = Yes) increases the probability of a positive test result from 2.97 to 99%. This illustrates the intuition that a cause (presence of disease) makes its effect (positive test result) more

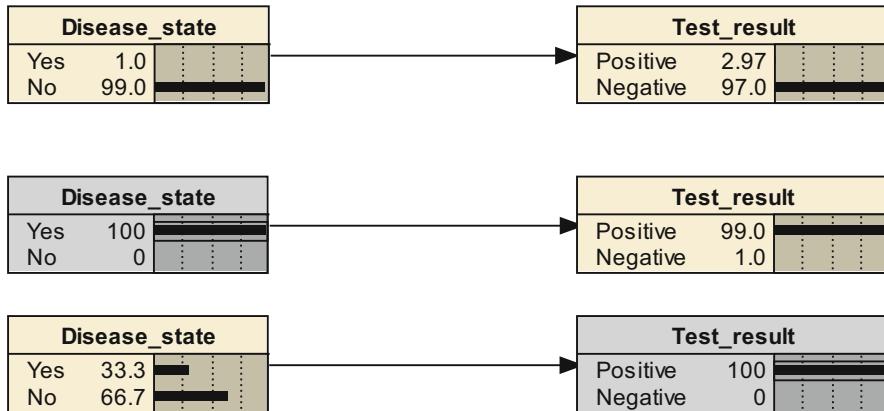


Fig. 2.3 Changes in probability may reflect inference or causation (or both)

probable. However, the bottom BN in Fig. 2.3 illustrates a reason that “ X causes Y ” and “ X increases the probability of Y ” are not synonymous (where random variables X and Y are binary event indicators). In this BN, the finding of a positive test result increases the probability of disease from its unconditional prior of 1% to a posterior value of 33.3% based on (i.e., conditioned on) the finding of positive test result. Such examples make clear that attempted definitions of cause in terms of probabilities, such as “ X is a cause of Y if $P(Y|X) > P(Y)$, i.e., if the conditional probability of Y given that X has occurred is greater than the unconditional prior probability of X ” are not adequate. Indeed, dividing both sides of Eq. (2.6) by $P(x)$ yields the identity in Eq. (2.13), which implies that if finding that $Y = y$ increases the probability that $X = x$ (meaning that $P(x|y)/P(x) > 1$, or $P(x|y) > P(x)$) then, conversely, finding that $X = x$ increases the probability that $Y = y$.

$$P(x|y)/P(x) = P(y|x)/P(y) \quad (2.13)$$

The relationship “increases the probability of” between two events is seen to be symmetric, but the relationship “causes” between two events is generally taken to be asymmetric. Hence, one is not a good model for the other.

Other ways in which “increases the probability of” can fail to coincide with “causes” are important in applied sciences such as epidemiology. For example, the diverging DAG in (2.14) illustrates the concept of *confounding*.

$$X \leftarrow Z \rightarrow Y \quad (2.14)$$

In such a model, seeing a high value of X may make a high value of Y more likely, not because X causes Y , but because Z causes both.

Example: Confounding—Effects of Common Causes

In DAG model (2.14), suppose that high values of Z are associated with high values of both X and Y , and that low values of Z are associated with low values of both X and Y . Specifically, suppose that all three variables are binary (0–1) random variables representing the status of three different conditions in an individual, with 1 indicating presence and 0 indicating absence. Suppose that the marginal distribution for input Z is that it is equally likely to be 0 or 1 and that the CPTs for X and Y are specified by $P(X = 1 | Z = 1) = 0.8$; $P(Y = 1 | Z = 1) = 0.7$; and $P(X = 0 | Z = 0) = P(Y = 0 | Z = 0) = 1$. (The remaining CPT probabilities can be found by subtracting these from 1 where needed, e.g., $P(Y = 0 | Z = 1) = 1 - P(Y = 1 | Z = 1) = 1 - 0.7 = 0.3$.) Then the prior probability that $Y = 1$ is 35%:

$$P(Y = 1) = P(Z = 1)*P(Y = 1|Z = 1) = 0.5*0.7 = 0.35.$$

On the other hand the conditional probability that $Y = 1$, given that $X = 1$, can be calculated with the help of Bayes' rule (or by building the DAG in Netica, compiling it, and entering the finding $X = 1$):

$$\begin{aligned} P(Y = 1|X = 1) &= P(Z = 1|X = 1)*P(Y = 1|Z = 1) \\ &= P(X = 1|Z = 1)*P(Z = 1)*P(Y = 1|Z = 1) \\ &\quad \times/(P(X = 1|Z = 1)*P(Z = 1) + P(X = 1|Z = 0)*P(Z = 0)) \\ &= 0.8*0.5*0.7/(0.8*0.5 + 0 + 0) = 0.70 \end{aligned}$$

Thus, seeing that $X = 1$ doubles the probability that $Y = 1$, from a prior value of 0.35 to a posterior value of 0.70. Yet, this is not because X has any causal effect on Y , but because seeing that $X = 1$ is diagnostic of Z probably being 1, and hence of Y probably being 1. This idea extends to ordered categorical and continuous random variables. For example, if high values of X , Y , and Z represent high exposure to a bad (e.g., contaminated or unclean) environment, poor health (or high mortality and morbidity rates), and low income (or high poverty), then one might expect to observe a positive association between bad environment and poor health even if neither causes the other, but both are caused by the confounder *low_income*. If only X and Y are observed variables, and the confounder Z is an unmeasured (also called a *hidden* or *latent* variable), then there might be no easy way to detect whether $X \rightarrow Y$ or DAG model (2.14) is correct.

$$\text{bad_environment} \leftarrow \text{low_income} \rightarrow \text{poor_health} \quad (2.15)$$

Example: Collider Stratification and Selection Bias—Conditioning on a Common Effect

The converging DAG model (2.16) and its abstract version (2.17) show situations in which two variables point into a third. If arrows are interpreted as reflecting causality

(by any definition), then the third variable can be viewed as a common effect of the two causes. Conditioning on a specific value of the effect variable induces a probabilistic dependency between the values of the variables that point into it.

$$\text{low_income} \rightarrow \text{hazardous_occupation} \leftarrow \text{good_health} \quad (2.16)$$

$$X \rightarrow Z \leftarrow Y \quad (2.17)$$

Z is called a *collider* between X and Y in this DAG, for obvious reasons. A possible interpretation is that only people with good health (indicated by $Y = 1$) or with low incomes (indicated by $X = 1$) are likely to work in a particular hazardous occupation (indicated by $Z = 1$). Treating all variables as binary for simplicity, suppose that the marginal distributions are that each of X and Y is equally likely to be 0 or 1, and the expected value of Z is given by the structural equation $E(Z) = 0.5(X + Y)$ (equivalent to the CPT $P(Z = 1 | X = 0, Y = 0) = 0; P(Z = 1 | X = 1, Y = 0) = 1; P(Z = 1 | X = 1, Y = 1) = P(Z = 1 | X = 0, Y = 1) = 0.5$). Building this network in Netica, entering the finding $Z = 1$, and varying the value of X shows that, with the constraint $Z = 1$, cases with $X = 1$ have a 2/3 probability of having $Y = 1$, but cases with $X = 0$ have a 100% probability of having $Y = 1$. Interpreting this pattern in terms of model (2.16), among workers in a hazardous occupation (corresponding to $Z = 1$), those with low incomes (corresponding to $X = 1$) are less likely to have good health (corresponding to $Y = 1$) than those with high income (corresponding to $X = 0$). Quantitatively, the probability of good health ($Y = 1$) is 2/3 for those with low income, compared to 100% for those with high income. The reason is not that high income is a cause of better health. Rather, the explanation is that having high income implies that a worker is not employed in this occupation because of low income, and therefore the rival explanation for such employment—that the worker has good health—becomes more probable. This is another way in which observing one condition (high income) can make another more likely (good health) even if neither is a cause of the other.

The general pattern illustrated by the foregoing example, in which conditioning on a common effect (or, in less suggestive language, a common child or descendant in a DAG) induces statistical dependencies among their parents or ancestors, has been discussed under different names in epidemiology, including selection bias, collider stratification bias, and Berkson's bias (Cole et al. 2010; Westreich 2012). A practical implication is that statistic modeling can inadvertently create significant associations and dependencies between variables that are not causally related, or even associated with each other in the absence of conditioning on other variables. Either of the following two common research situations can create such non-causal statistical associations:

- *A study design that selects a certain population for investigation*, such as an occupational cohort, a population of patients in a specific hospital or health insurance plan, or residents in a certain geographic area. If the dependent variable

of interest and an explanatory variable of interest both affect membership in the selected study population, then selection bias might create spurious (meaning non-causal) associations between them.

- *A statistical analysis, such as regression modeling, which stratifies or conditions on the observed values of some explanatory variables.*

Stratification and conditioning are usually done to “control for” the statistical effects of the selected explanatory variables on a dependent variable. However, the attempted control can create spurious effects if a variable that is stratified or conditioned on—that is, a variable on the right side of a regression equation—is affected by both the dependent variable and one or more other explanatory variables. Conditioning on it can then induce spurious associations between other explanatory variables that affect it and the dependent variable. For example, if the data-generating process is described by the collider DAG $X \rightarrow Z \leftarrow Y$, then regressing Y against X and Z might produce highly statistically significant regression coefficients for both X and Z even though neither one actually affects Y . If $Z = X + Y$, where X and Y are two independent random numbers and Z is their sum, then fitting the regression model $E(Y|X, Z) = \beta_0 + \beta_X X + \beta_Z Z$ to a large sample of values for the three variables ($x, y, z = x + y$) produces estimated regression coefficients of 1 for β_Z and -1 for β_X , corresponding to the equation $Y = Z - X$, even though Y is unconditionally independent of X . The values of X and Y might have been generated by independent random number generators, but conditioning on their common effect Z by including it on the right side of the regression equation induces perfect negative correlation between X and Y , despite their unconditional independence. Thus, *regression modeling can estimate a significant statistical “effect” of X on Y even if they are independent*, if some other covariate Z that depends on both of them is included in the model.

Causal Probabilities and Causal Bayesian Networks

It is clear from the foregoing examples that observing a high value of one random variable can increase the probability of observing a high value for another random variable even if there is no direct causal relation between them. Reverse causation, confounding, and selection bias provide three different ways for the observed occurrence of one event to make occurrence of another more probable without causing it. However, BNs also provide a simple way to represent direct causal effects and mechanisms. Suppose that the CPT for each node in a BN is interpreted as specifying how its conditional probability distribution changes if the values of its parents are changed. Such a BN is called a *causal* Bayesian network. In a causal BN, the parents of a node are its *direct causes* with respect to the other variables in the BN, meaning that none of them mediates the parents’ effects on the node. Conversely, the children of a node, meaning the nodes into which it points via arrows of the DAG, are interpreted as its *direct effects*. The CPT for a variable in a causal BN

represents the causal mechanism(s) by which the values of its parents affect the probability distribution of its values. If the causal BN represents manipulative causation, then changing the value of a parent changes the probability distribution of the child, as specified by its CPT. In this sense, the causal BN implements the idea that *changing a cause changes the probabilities of the values of its direct effects*.

Example: Modeling Discrimination in College Admissions—Direct, Indirect, and Total Effects of Gender on Admissions Probabilities

Setting: Suppose that a college admissions committee reviewing recent admissions statistics finds that men have a 60% acceptance rate while women with the same academic qualifications have only a 50% acceptance rate. Half of applicants are men and half are women. Their standardized test scores, educational backgrounds, and other academic qualifications are indistinguishable.

Causal prediction question: What would be the effects on acceptance rates for women if each department were required to change its admissions rate for women to be equal to its admissions rate for men with the same academic qualifications? What data are needed to answer this question?

Solution: Accurately predicting what would happen if the policy were adopted requires a causal model for acceptance probabilities. Suppose that the causal DAG model in Fig. 2.4 describes the data-generating process. The data for the node probability tables are listed below the DAG.

$$P(\text{man}) = P(\text{woman}) = 0.50$$

$$P(\text{history} \mid \text{man} = 1), \text{ so } P(\text{math} \mid \text{man}) = 0$$

$$P(\text{history} \mid \text{woman}) = 0, \text{ so } P(\text{math} \mid \text{woman} = 1)$$

$$P(\text{accept} \mid \text{woman}, \text{history}) = 0.70$$

$$P(\text{accept} \mid \text{woman}, \text{math}) = 0.50$$

$$P(\text{accept} \mid \text{man}, \text{history}) = 0.60$$

$$P(\text{accept} \mid \text{man}, \text{math}) = 0.40$$

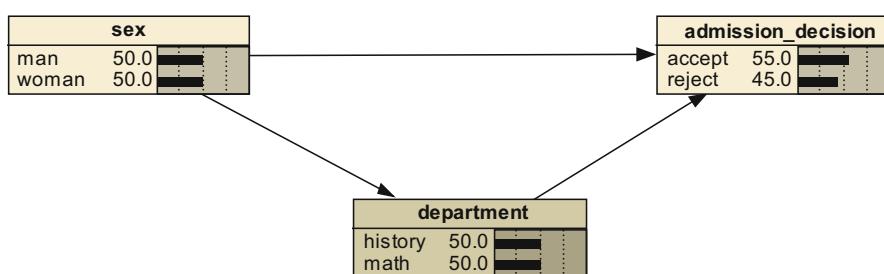


Fig. 2.4 A causal BN for the college admissions decision

In words, half of all applicants are women. Women apply only to the math department and men apply only to the history department; thus, the direct effect of sex on admissions probability, if any, is completely confounded with the effect of department. In the given description of the problem setting, department is an omitted confounder: the description does not mention it. But it must be considered to build a causal model adequate for answering questions about the effects of interventions. Each department currently has an admissions rate that is 0.10 higher for women than for men (given the same qualifications), with the math department's admissions rate being 0.50 for women and 0.40 for men, and the history department's admission rate being 0.70 for women and 0.60 for men. (These numbers would have to come from old data or from policies, since currently only women apply to the math department and only men apply to the history department.) Thus, if each department changed its admissions rates for women to equal its admissions rate for men, then admissions rates for women would fall from 0.50 to 0.40. The *direct effect* of sex on admissions probability is outweighed by its *indirect effect*, mediated by the department applied to, to yield a *total effect* that is opposite in sign to the direct effect. Being a woman increases acceptance probability, other things being equal (namely, department applied to), but decreases acceptance probability overall. This is an example of Simpson's Paradox.

If inquiry revealed a different CPT for the admission decision node, then the predicted outcome would be quite different. If the CPT for the admission decision node were as follows:

$$\begin{aligned} P(\text{accept} \mid \text{woman, history}) &= 0.50 \\ P(\text{accept} \mid \text{woman, math}) &= 0.50 \\ P(\text{accept} \mid \text{man, history}) &= 0.60 \\ P(\text{accept} \mid \text{man, math}) &= 0.60 \end{aligned}$$

then the policy change would increase the admissions rate for women from 0.50 to 0.60. Thus, the effects of the policy change depend on the details of the admission decision CPT. They cannot be inferred solely from available data on current admission rates broken down by sex, or even from current data broken down by sex and department, since these rates only show what is happening now and this information does not reveal what would happen under altered conditions. That requires knowledge of $P(\text{accept} \mid \text{man, math})$, which is not currently observed, as men do not apply to the math department. This illustrates the fact that *the consequences of a policy intervention can be underdetermined by data collected before the policy is enacted*. The Lucas critique mentioned in Chap. 1 expands on this point.

The discrimination example illustrates the differences among direct, indirect, and total effects of sex on admissions probabilities. More generally, epidemiologists distinguish among the following types of effects in causal DAG models (Robins and Greenland 1992; Pearl 2001; Petersen et al. 2006; VanderWeele and Vansteelandt 2009):

- The *direct effect* of a change in a parent on a child is the change in the child's expected value (or, more generally, in its conditional probability distribution)

produced by changing the parent's value, holding the values of all other variables fixed. In a structural equation model (SEM) with $Y = aX + bZ$ and $Z = cX$, for example, the direct effect on Y of a unit increase in X would be a , since Z would be held fixed.

- The *controlled direct effect* is the effect if other variables are held fixed at the same values for all cases described by the causal model. For example, consider an SEM with Y depending (perhaps nonlinearly) on X and Z , and with Z depending on X , where X is exposure, Y is an indicator of health effects found in insurance records of covered treatments, and Z is a covariate such as participation in an employer-provided insurance program. The controlled direct effect of an increase in X on Y would be calculated by holding Z fixed at the same level (all insured or all not) for all individuals in the study. Knowledge of the CPT $P(y | x, z)$ makes this easy to calculate.
- By contrast, the *natural direct effect* of X on Y allows the mediator Z to have different values for different individuals, holding them fixed at the levels they had before X was increased. For data analysis purposes, *partial dependence plots* (PDPs), introduced later in this chapter, provide a useful non-parametric method for estimating natural direct effects.
- The *total effect* of a change in X on Y is the change in the expected value of Y (or its probability distribution) from the change in X when all other variables are free to adjust. In the SEM with $Y = aX + bZ$ and $Z = cX$, an increase in X by one unit increases Y by a total of $a + bc$ units, representing the sum of the direct effect a and the indirect effect mediated by Z , bc .
- The *total indirect effect* is the difference between the total effect and the natural direct effect.

A further assumption often made in interpreting a causal BN in light of cross-sectional data on its variables is that the probability distribution for each variable is *conditionally independent* of the values of its more remote ancestors, and, indeed, of the values of all of its non-descendants, given the values of its parents. (More generally, each variable is conditionally independent of all others, given the values of its parents, children, and spouses, i.e., other parents of its children; these are collectively referred to as its “Markov blanket.”) For the DAG model $X \rightarrow Z \rightarrow Y$, this condition implies that Y is conditionally independent of X given Z . In symbols, $P(y | x, z) = P(y | z)$ for all possible values x , y , and z of random variables X , Y , and Z , respectively. Thus, conditioning on X and Z provides no more information about Y than conditioning on the value of Z alone. Informally, X affects Y only through Z ; more accurately, all of the information that X provides about Z is transmitted via Y . The assumption that each node is conditionally independent of its non-descendants given its parents is termed the *Markov assumption* for BNs, or the *Causal Markov Condition* (CMC) for causal BNs, in analogy to the property that the next state of a Markov process is conditionally independent of its past states, given the present state. Philosophical objections raised about the suitability of CMC for causal analysis has led to refinements of the concept (Hausman and Woodward 1999, 2004) and alternative formulations use more explicitly causal language, such as that *a variable*

is conditionally independent of its non-effects given its direct causes. When causal BNs are interpreted as modeling manipulative causality, CMC implies that changing a variable's indirect causes, i.e., its ancestors, affects its probability distribution only via effects on its direct parents.

In the context of data analysis, CMC is often paired with a second condition, termed *faithfulness*, stating that the set of all conditional independence relations among variables consists of exactly those that are implied by the DAG model. This rules out the logically possible but empirically unlikely situation in which one variable appears to be conditionally independent of another only because multiple effects coincidentally cancel each other out or duplicate each other. For example, in the DAG model $X \rightarrow Z \rightarrow Y$, faithfulness implies that X and Z values do not happen to coincide. If they did, then the conditional independence relations in this DAG would be indistinguishable from those implied by $Z \rightarrow X \rightarrow Y$, making it impossible to tell whether X or Z is the parent of Y (i.e., whether Y is conditionally independent of Z given X or whether instead Y is conditionally independent of Z given X). Likewise, if exercise is a direct cause of food consumption and if both are parents of cardiovascular risk in a DAG model, then faithfulness would require that the direct effects of exercise on cardiovascular risk must not be exactly counterbalanced by the indirect effect via food consumption, creating an appearance in the data that cardiovascular risk is independent of exercise.

An obvious way for CMC to fail is if a common cause not included in the BN affects both a node and one of its ancestors without affecting its parents in the BN. Such a *hidden confounder* (also called an *unobserved confounder* or *latent confounder*) could create a statistical dependency between X and Y in the DAG model $X \rightarrow Z \rightarrow Y$ even after conditioning on the value of Z (Tashiro et al. 2014). The requirement that each variable in a causal BN depends *only* on its parents shown in the DAG, implying that all common causes of any two variables in the DAG are also included in the DAG, is called *causal completeness*. The closely related condition that all such common causes are not only *modeled* (represented in the DAG), but also *measured* or observed, is called *causal sufficiency*. When it holds, the observed values of the parents of a node are sufficient to determine which conditional distribution in its CPT the child's values are drawn from. None of the conditions of causal sufficiency, causal completeness, faithfulness, or CMC necessarily holds in real-world data sets, but they are useful for discussing conditions under which causal discovery algorithms can be guaranteed to produce correct results.

Causal BNs model manipulative causation in a straightforward way. Changing a direct cause in a causal BN changes the conditional probability distribution of its children, as specified by their CPTs. In particular, the probability distribution of a variable can be changed, or *manipulated*, by exogenous acts or interventions that set specific values for some or all of its controllable parents. These exogenously specified values override endogenous drivers, effectively disconnecting any arrows pointing into the manipulated variables, and simply fix their values at specified levels. (Technical discussions often refer to this disconnecting as “graph surgery” on the BN's DAG and denote by “ $do(X = x)$ ” or “ $do(x)$ ” the operation of setting the

value of a variable X to value x , to emphasize that the value of X is exogenously set to value x rather than being passively observed to have that value). Similarly, the CPT for a node in a causal BN depends only on the values of its parents (direct causes), and not on how those values were determined, whether via exogenous manipulation or endogenously.

Example: Calculating Effects of Interventions via the Ideal Gas Law

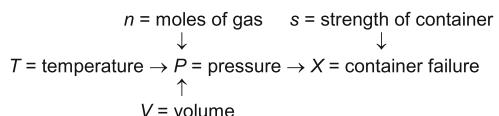
Many BNs contain “function nodes” that calculate the value of a variable a deterministic function of the values of its parents. Netica and other BN software packages allow deterministic functions to be specified in place of CPTs. (A function $y = f(x)$ can be construed as special CPT, with $P(y = f(x) \mid x) = 1$.) Figure 2.5 shows an example involving the continuous variables P = pressure of a gas inside a container, T = temperature of the container, V = volume of the container, n = amount of gas in the container, s = strength of the container. There is also a binary indicator variable X = container failure, having possible values of 1 if the container fails during a certain time interval of observation and $X = 0$ otherwise. The CPT for pressure is given by a formula such as

$$P = nRT/V$$

where R is a constant with value 0.08206 when V is in liters, P in atmospheres, n in moles, and T in degrees Kelvin. This idealization suffices for purposes of illustration, although realistic formulas are more complex. The probability of container failure over the interval of interest depends on the pressure and on the strength of the container, which in turn may depend on container geometry and details of the spatial distribution of material characteristics not shown in the model.

Suppose that the probability of failure over the interval of interest if the container has pressure P and strength s is given by some empirically derived function, $E(X \mid P, s) = F(P, s)$. If a decision-maker controls the temperature at which the container is stored, and if all other parents and ancestors of X have known values, then an intervention that sets the temperature to some specific value T^* (sometimes indicated via the notation “ $do(T^*)$ ”) will thereby cause a risk of failure given by $F(P^*, s)$ where we define $P^* = nRT^*/V$. On the other hand, if an overpressure release valve and a compressed gas supply are rigged up to keep P fixed at some target value P_0 , then the risk of failure will remain at $F(P_0, s)$ no matter how T is varied. This illustrates the fact that X depends on T only through P , and therefore interventions that set the value of P to some specific

Fig. 2.5 A causal BN for container failure due to overpressure



value override the effects of changes in T , effectively cutting it out of the causal network that affects X .

Statistical methods for estimating causal BNs from data, known as *causal discovery algorithms* are discussed more fully later in this chapter. They often simplify the estimation task by assuming that the CMC, faithfulness, and causal sufficiency conditions hold. More general causal discovery algorithms allow for the possibilities of latent confounders and selection bias (Zhang 2008; Ogarrio et al. 2016). When causal sufficiency cannot be assumed, meaning that variables not included in a causal graph model might explain some of the observed statistical dependencies between its variables, the *directed* acyclic graph (DAG) modeling assumption is relaxed to allow undirected (or bidirected) arcs as well as directed ones. This creates *mixed graphs* with both directed and undirected arcs between nodes. Mixed graphs include several specialized data structures (such as maximal ancestral graphs (MAGs) and partial ancestral graphs (PAGs), the latter representing classes of MAGs that are not distinguishable from each other by conditional independence tests) that acknowledge that some of the observed variables may have unobserved common causes, or ancestors, outside the model (Triantafillou and Tsamardinos 2015). An undirected arc joining two variables indicates that their association can be explained by an unobserved confounder. Since BN inference algorithms work equally well with networks having the same arcs oriented in opposite directions as long as the joint probability distribution of variables has been factored correctly (e.g., $P(X, Y)$ can be factored equally well as $P(y)P(x \mid y)$ or as $P(x)P(y \mid x)$, corresponding to $Y \rightarrow X$ or $X \rightarrow Y$, respectively), they can readily be extended to apply to mixed graphs. However, it is then important to keep track of which inferences have causal interpretations and which only reflect statistical dependencies. Current algorithms for causal discovery and inference and for synthesizing causal inferences across multiple studies automatically keep track of causal and statistical inferences and clearly distinguish between them (Triantafillou and Tsamardinos 2015).

Dynamic Bayesian Networks (DBNs)

The DAG models we have looked at so far do not show the passage of time. Yet time is essential to causality, insofar as effects must not precede their causes. Fortunately, it is easy to incorporate time into BNs by using successive time periods, called *time slices*, to distinguish between the values of the same variables at different times or in different time periods. This creates a *dynamic Bayesian network* (DBN) that shows dependencies among the values of different random variables over time. A common notation is to time-stamp variables with subscripts to indicate periods, as in Fig. 2.6. This diagram depicts three time series variables, A , B , and C , which interact with each other over time. Their values at time (i.e., within time slice) t are denoted by A_t , B_t , and C_t , respectively using subscript notation. The value of random variable at time t , denoted by B_t at the right of the network, depends on the concurrent value of

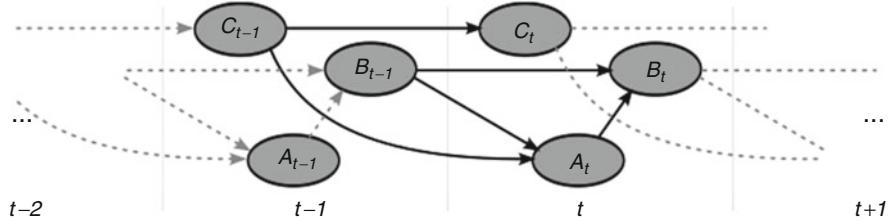


Fig. 2.6 A portion of a dynamic Bayesian network (DBN). Source: https://en.wikipedia.org/wiki/Dynamic_Bayesian_network

variable A , namely A_t , and on the previous value of B , namely B_{t-1} . Dependencies are indicated by the solid dark arrows between nodes. In the special case of a deterministic dynamic model, this dependency among the values of the variables can be described by a difference equation of the form for some appropriate function f :

$$B_t = f(A_t, B_{t-1}).$$

In the more general case of probabilistic dependencies, the corresponding information is given by a CPT specifying the conditional probabilities

$$P(B_t = x | A_t = y, B_{t-1} = z)$$

for all possible of x , y , and z . We might abbreviate this CPT as

$$P(B_t | A_t, B_{t-1}),$$

and use similarly abbreviated notation for the CPTs at other nodes:

$$P(A_t | B_{t-1}, C_{t-1}) \text{ and } P(C_t | C_{t-1})$$

If these CPTs are invariant over time, as might be expected for a description based on stable causal laws or mechanisms, then they can be used to generate an entire multi-period network. The probability distributions for the values of the three variables in each successive period are derived from their values in the preceding period via the CPTs $P(A_t | B_{t-1}, C_{t-1})$, $P(C_t | C_{t-1})$, and $P(B_t | A_t, B_{t-1})$. The grey dashed arrows in Fig. 2.6 suggest how the values for the two time slices shown can be linked to past and future values via these CPTs. Figure 2.7 makes these links explicit by showing a “rolled-out” version of the DBN for three periods. Netica automatically generates such rolled-out networks (which it refers to as time expansions or time-expanded versions) from more concise networks, such as Fig. 2.6, that determine the probabilities of the variable values in each period from their values in previous periods.

A further compression of notation can be accomplished by referring collectively to all of the variables in time slice t (namely, A_t , B_t , and C_t in Figs. 2.6 and 2.7) as simply X_t . Then the DBN can be seen as providing a way to calculate the probabilities of current values of values from their past values, $P(X_t | X_{t-1})$, or, equivalently (since t is just a dummy variable indicating the time stamp), $P(X_{t+1} | X_t)$. This can be

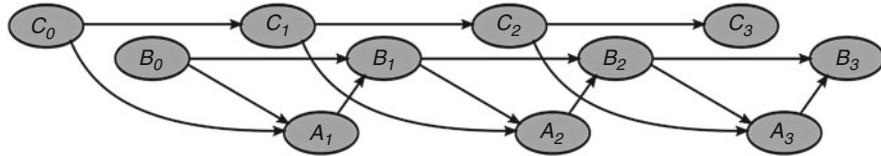


Fig. 2.7 A rolled out version of the dynamic Bayesian network (DBN) in Fig. 2.6. Source: https://en.wikipedia.org/wiki/Dynamic_Bayesian_network

viewed as the state transition probability matrix for a Markov chain. One use of DBNs is to store efficiently prior probabilities for the initial state and CPTs representing state transition probabilities for Markov chains. These can then be used to compute probabilities of future states. (The calculation is given by the famous Chapman-Kolmogorov equation: if P is the one-step state transition probability matrix with the number in its row i and column j giving the probability that the next state will be state j if the current state is state i , then the element in the same position, row i and column j , of P^n is the probability that the system will be in state j after n transitions, starting from state i . Equivalently, the law of exponents $P^{(m+n)} = P^m P^n$ for the transition matrix can be interpreted as showing that the probability of passing from one state to another in $m + n$ steps is found by summing the probabilities of getting to each possible intermediate state in m steps and then from there to the target state in the remaining n steps. If the starting state is uncertain, and uncertainty is described by a prior probability distribution over states, then pre-multiplying this probability vector by the n -th power of P will give the probability distribution for the state after n transitions.) Often, current values depend on lagged values that go back several periods, violating the Markov assumption that probabilities of next states are conditionally independent of the past given the current state. Then a useful trick (called “state augmentation”) is to stack up all of the variable values needed to compute values of variables in the next time slice (time $t + 1$) into one long list, again denoted by X_t even if it includes some values of variables from periods before t . By construction, X_t contains all of the information needed to determine $P(X_{t+1} \mid X_t)$, so the Markov model holds for the augmented state and the large body of well-developed techniques for Markov chains can be applied, including predicting future state probabilities over different forecast horizons via the Chapman-Kolmogorov equation. DBNs can also be used to represent and draw inferences for a wide variety of other uncertain systems, including hidden Markov models (HMMs), multiple time series, and partially observable Markov decision processes (POMDPs).

Many commercial and free software packages are now available for BN and DBN modeling. The Belief and Decision Networks tool at www.aispace.org/mainTools.shtml provides a useful tutorial and software for learning to use BNs. The commercial Bayes Server™ software (www.bayesserver.com) provides a powerful modeling environment for both BNs and DBNs, with support for modeling of multiple time series, latent variables, and both discrete and continuous variables. Free interactive demos (www.bayesserver.com/Live.aspx) allow users to interactively enter

findings into BNs and see how probabilities at other nodes adjust. Other software tools for learning BNs and DBNs from data and using them to draw inferences are listed by the Australasian Bayesian Network Modeling Society at <http://abnms.org/forum/viewtopic.php?f=7&t=11515>. They include BayesiaLab, CaMML for causal Bayesian networks, Genie, Hugin, JavaBayes, Tetrad, and WinBUGs (implementing Gibbs sampling-based inference) as well as the Analytica and Neteica software introduced in Chap. 1 and this chapter.

Causal Risk Models Equivalent to BNs

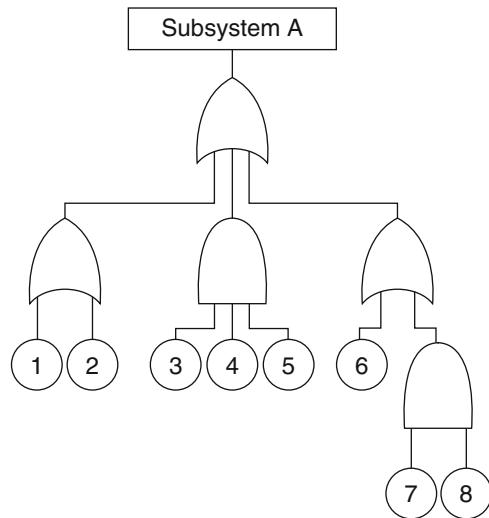
Understanding Bayesian network notation and capabilities yields multiple dividends. Key analytic techniques of system reliability engineering, dynamic systems modeling and prediction, probabilistic risk analysis, and quantitative decision analysis can be viewed as special cases of BNs, thereby gaining in clarity, simplicity, and generality. This section examines how other modeling methods used to clarify causality, uncertainty, and risk can be mapped to the framework of causal BNs.

Fault Tree Analysis

One way to envision, predict, and mitigate risks in complex reliability systems is to imagine that some undesirable event, such as catastrophic failure of a system (e.g., loss of power in an airplane, collision of ships or of trains, core breach in a nuclear power plant, or theft of credit card data) occurs, and then to inquire systematically *how* it might have occurred, how probable or frequent are the conjunctions of conditions that suffice to cause it, and what could have been done to prevent it. The necessary thinking can be organized via a *fault tree*. In a fault tree, nodes of the tree represent events. The undesirable event of interest is called the *top event*. It is conventionally drawn at the top of the tree, which is then expanded downward, as in Fig. 2.8. Below the top event are displayed logical combinations of other events that suffice to cause it. This can often be done conveniently using AND gates to show conjunctions of events that suffice to cause the top event, and OR gates to show alternative ways (i.e., alternative conjunctions of events) that suffice to cause it. For example, if a computer monitor fails to light up when the power switch is turned to on, one possible explanation is that no power is reaching the computer; a different possibility is that the monitor is broken. A fault tree would include both possibilities, depending from an OR gate below the event that the monitor does not light up when switched on. The rest of the tree is developed recursively, by successively developing possible explanations for each sub-event in terms of logical combinations of lower-level events that could cause them. For example, a failure of power to reach the computer could occur if it is not plugged in, or if there has been a power failure, or if there is a fault in the power cord, or if a circuit breaker has been tripped; these

Fig. 2.8 Schematic sketch of a small fault tree. The top event is failure of Subsystem A.” Basic events are numbered 1–8. AND gates have flat bottoms, OR gates have curved bottoms.

Source: https://en.wikipedia.org/wiki/Fault_tree_analysis



would be sub-events below the event of power not reaching the computer when it is switched on.

The process of expanding events into possible explanations (via alternative conjunctions of predecessor events that jointly suffice to cause them) stops when all of the remaining unexplained events in the model (the “leaf nodes” at the bottom of a downward-growing fault tree) are “basic events.” These are events that are not explained or described further. Their occurrence probabilities or rates are estimated directly from data. For example, in Fig. 2.8, there are eight basic events at the bottom of the tree, numbered 1–8. The logic of the tree shows that the top event occurs if basic events 3–5 all occur, or if both 7 and 8 occur, or if any of the other basic events 1, 2, or 6, occurs. The probability or frequency (occurrences per year) of the top event can now be calculated from the probabilities or frequencies of the basic events using rules of probability: the probability of a conjunction of independent events is the product of their probabilities, and the probability of a disjunction is given by a series (the inclusion-exclusion series) that can be approximated by the sum of their probabilities if all probabilities are sufficiently small. Such rules allow the probabilities of the basic events to be pushed up the tree by resolving (i.e., calculating the probability of) each AND and OR gate as soon as the probabilities of all of its children have been calculated, beginning with the parents of the leaf nodes. The probability or frequency of the top event is obtained by resolving that node once all of its children have been resolved.

Fault tree analysis (FTA) can be extended much further by allowing other types of logic gates; optimizing combinatorial algorithms for calculating approximate probabilities; identifying minimal cut sets and path sets that block or imply occurrence of the top event, respectively; speeding computations via Monte Carlo simulation; allowing for (possibly unreliable) repair of failed components; allowing for dynamic

logic gates (e.g., C occurs if A occurs and B occurs after A); showing unavailability of systems over time; calculating various importance metrics for components and subsystems, and so forth. Resources for learning more about FTA include the following:

- For basic concepts and methods of fault tree analysis, see the *Fault Tree Handbook* (NUREG-0492), available for download from the U.S. Nuclear Regulatory Commission (www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0492/).
- A free on-line FTA program at www.fault-tree-analysis-software.com/fault-tree-analysis can be used to experiment interactively with a fault tree solver, e.g., by varying the failure rates for basic events and seeing how the failure rate of the top event changes.
- A dynamic fault tree solver is available at <http://fmt.ewi.utwente.nl/puptol/dftcalc/>. Fault trees are entered via a simple standard syntax (known as the Galileo format after an earlier dynamic fault tree program (Dugan 2000)). Markov chains model component failures over time as the dynamic basic events of the fault tree. A graphic user interface provides easy access to calculations and reports.

Example: A Dynamic Fault Tree Calculator

Figure 2.9 illustrates one output from the DFTCalc Web-Tool dynamic fault tree calculator (<http://fmt.ewi.utwente.nl/puptol/dftcalc/>). The fault tree model is specified via the following Galileo commands:

```
toplevel "A";
"A" 2of3 "B" "C" "D";
"B" lambda=0.1 dorm=0;
"C" lambda=0.2 dorm=0;
"D" lambda=0.3 dorm=0;
```

The first command specifies the top event as A . The second line gives the reliability model, which is that A fails if at least 2 of B , C , and D fail (concisely indicated vs. the “2 of 3” syntax). The remaining lines specify the failure rates for B , C , and D as 0.1, 0.2, and 0.3 expected failures per unit time, respectively. (Lambda is the conventional designator for a failure rate. There is no dormant period before the components come on line and are susceptible to failure.) The output in Fig. 2.9 shows the growth of unreliability (i.e., probability occurrence of the top event) over time and the mean time to failure, which text output identifies as 4.5 time units. The program can be used to study how these outputs change as the failure rates of the basic events B , C , and D are changed, or if their states (working or failed) are specified as assumptions.

Despite its many important successes in a variety of practical applications, FTA has important limitations that can be overcome by re-expressing the fault tree model as a BN. Among them are the following (Bobbio et al. 2001):

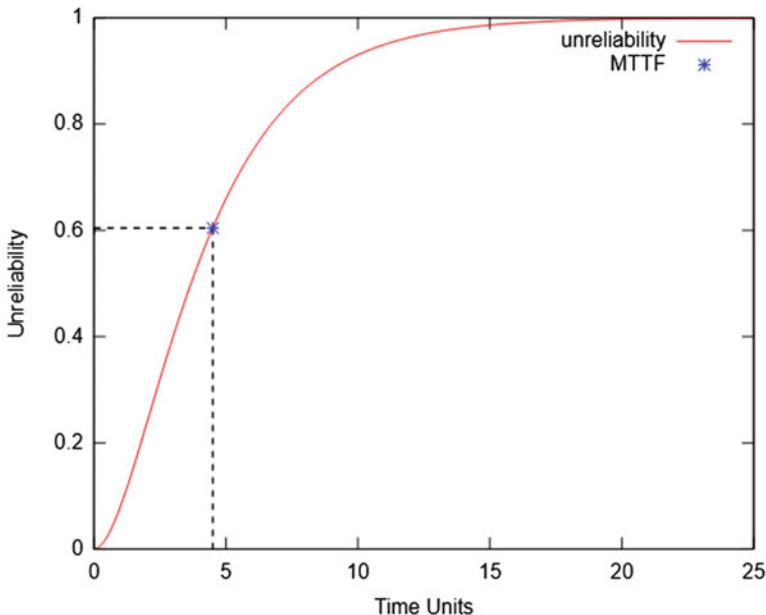


Fig. 2.9 Example output from DFTCalc Web-Tool dynamic fault tree calculator showing increase in probability of top event over time and mean time to failure (MTTF = 4.5 time units, indicated by asterisk). Source: DFTCalc Web-Tool dynamic fault tree calculator, <http://fmt.ewi.utwente.nl/puptol/dftcalc/>

- *Flexibility of event descriptions:* Events in classical FTA are binary: components can switch from fully on (working) to fully off (failed), but intermediate degrees of degraded performance are not allowed. This assumption is relaxed in some generalizations of fault tree analysis that allow for continuous state performance indicators, but most FTA codes are developed for binary logic gates. By contrast, BNs allows binary, unordered or ordered categorical, and continuous random variables (although many BN software implementations discretize arbitrary continuous random variables, e.g., into ten deciles). This flexibility lets BNs handle degrees of performance very easily.
- *Flexibility and generality of failure propagation logic:* FTA is built around logic gates, especially AND and OR gates, although other types of logic gates (including negation and exclusive-or gates) are allowed in some FTA software. Logic gates can easily be modeled as special deterministic conditional probability tables (CPTs). The conjunctive logic gate “C IF A AND B” is equivalent to a CPT with values $P(C | A, B) = A * B$ where A, B, and C are Boolean variables (with possible values 1 = True and 0 = False, indicating occurrence or non-occurrence of the events, respectively). Likewise, “C IF A OR B” is equivalent to $P(C | A, B) = A + B - A * B$. But CPTs can equally easily express many other dependencies, stochastic as well as deterministic. For example, “C occurs in a specified time interval with probability 0.2 if A and not B; C occurs with probability 0.5 if

B and not A ; and C occurs with probability 0.8 if both A and B ” can be represented by a CPT with values $P(C | A, B) = 0.2A + 0.5B + 0.1AB$ for each possible combination of values of the 0–1 input variables A and B . Thus, a CPT is a far more general and flexible representation of dependencies among an event and its parents than is a logic gate.

- *Independence requirements.* Basic events in FTA are often assumed to be statistically independent: the failure probability or rate of one component (represented by a basic event) does not depend on which others have failed. BNs can model dependencies among basic events via arrows linking them and CPTs describing the dependencies.
- *Inference:* Although fault trees are ideally suited for propagating information about failure rates up from the basic events at the leaves of the tree to the top event at its root, they are not designed to support more general inferences. If it is observed that an intermediate event has occurred part way up the tree—e.g., that power has not reached a monitor after a computer is switched on—then FTA does not automatically update the probabilities of basic events or the probabilities of other intermediate events to condition on this observation. A BN solver such as Netica automatically does such updating of probabilities in light of observations.

Fortunately, any fault tree can be converted automatically to a logically equivalent BN by mapping its logic gates to corresponding CPTs and its basic events to input nodes (Bobbio et al. 2001). The BN formulation also supports useful generalizations such as those just mentioned and others, including simultaneous analysis for multiple top events and natural treatments of common cause failures. BN inference algorithms can identify the most likely explanation for occurrence of the top event (or any other subset of events, entered as assumed findings) and can compute the conditional probability that a low-level event has occurred if it is assumed that a higher-level one occurs. These capabilities enable BNs to identify the components that are most likely to be involved in various failure scenarios, helping system designers and operators understand how failures are most likely to occur and what can be done to prevent them or to reduce their risks.

Event Tree Analysis

FTA’s approach of recursively expanding events in terms of combinations of lower-level events explain how they might occur reflects a style of reasoning known in artificial intelligence as *backward chaining*. It reasons backward from a specified target state (occurrence of the top event) to the conditions that could cause it. In the reverse direction, *forward chaining* begins by assuming that some specific initiating event occurs, and then reasons forward about what could happen next and how probable are different sequences of consequences. For example, if off-site power is lost at a chemical manufacturing plant, one possibility might be that on-site generators would activate and supply the power needed for continued operations or for

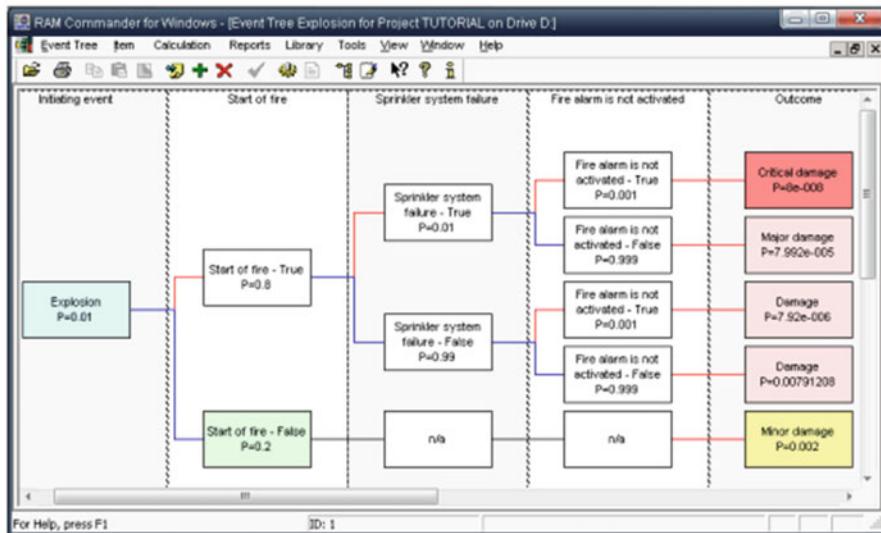
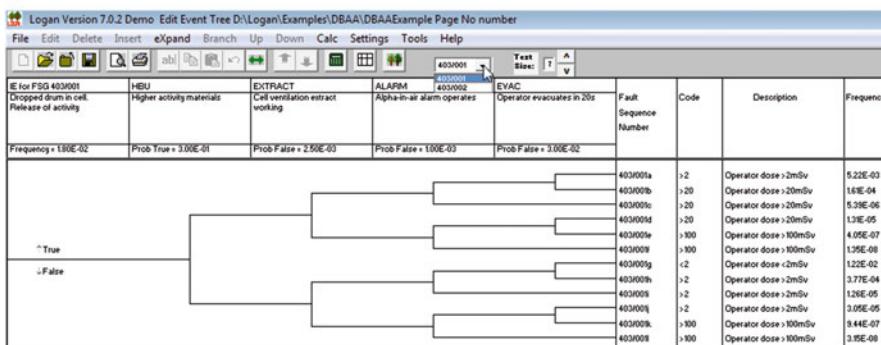
a**b**

Fig. 2.10 (a) An event tree with four events and four possible distinguished outcomes. *Source:* www.reliability-safety-software.com/eta/. (b) An event tree for dose of radiation received by an operator if a dropped drum releases radioactive materials. *Source:* Logan Fault and Event Tree Analysis software, http://logantfa.com/index_files/enlarged_image3.html

safe shut-down. A different possibility might be that one or more of the on-site generators fails to start when needed. An *event tree* shows the various possibilities branching out from the initiating event. *Event tree analysis* (ETA) quantifies the probabilities of different outcomes that might be caused by the initiating event.

Figure 2.10a shows an example with an initiating event consisting of an explosion and three subsequent binary events in which the explosion either does or does not ignite a fire, the sprinkler system either is or is not working, and the fire alarm either

is or is not activated. At the tips or leaves of the tree on the right-hand side are outcomes caused by the different possible event sequences. This example is taken from free demo software for the Event Tree Analysis Module of the RAM Commander suite of reliability and risk analysis software (www.reliability-safety-software.com/eta/). Probabilities or average annual frequencies of each outcome node in any given interval of time can be calculated by multiplying the probability of the initiating event by the conditional probabilities of each event along the path leading to the outcome, where the occurrence probability of each event is conditioned on the events that precede it. (Technically, probabilities and average annual frequencies are slightly different, with probabilities necessarily lying between 0 and 1 and average annual frequencies having units of expected occurrences per year, which in principle has no necessary upper bound. But they are numerically very close when the occurrence probability is small, e.g., 1% or less. Otherwise, the relation between them is that the probability p for occurrence of an event within T years is: $p = 1 - \exp(-\lambda T)$, where λ denotes the average annual frequency. In reliability theory, λ is also known as the failure rate or the intensity for the event.)

Figure 2.10b shows an example of an event tree for the dose of radiation received by a human operator if a drum of radioactive material is dropped and releases radioactive material. The dose received depends on the nature of the material (high activity or not), whether ventilation is working at the time, and on how quickly the operator evacuates, which in turn depends on whether an alarm is working at the time. It is clear that dichotomizing all descriptions is a drastic simplification. In reality, the drum might spill different amounts and might do so at different locations within the cell being considered. The operator's time to evacuate is a continuous random variable rather than a binary one (more or less than 20 s), as in the even tree. But even such coarse-grained descriptions are useful in envisioning what might happen in the event that a drum is dropped, and for understanding how changing the conditional probabilities of some events (e.g., by installing a more reliable alarm) changes the frequency distribution of outcomes.

Like fault trees, event trees can be mapped automatically to equivalent BNs (Bearfield and Marsh 2005). Events and outcomes in the tree become nodes in the BN, with each node having arrows directed into it from the event nodes on which it directly depends. The CPTs for these nodes are obtained from the conditional probabilities in the event tree. Nodes can readily be generalized to allow more than two values, so that ordered-categorical or continuous descriptions of spill sizes, for example, become practical and easy to include in modeling. BN algorithms can then be used to quantify the conditional probabilities of outcomes given observations or assumptions about predecessor events; conversely, given an assumed outcome, standard BN calculations can solve for the most likely explanation in terms of conjunctions of preceding events that could cause it. Thus, the BN formulation combines the advantages of forward chaining reasoning to predict the probabilities of undesired outcomes and backward chaining reasoning to explain them, and hence perhaps help understand how to prevent them.

Bow-Tie Diagrams for Risk Management of Complex Systems

The marriage of a fault tree and an event tree is the *bow tie diagram*, illustrated schematically in Fig. 2.11. The left side is a fault tree turned sideways, so that its top event, called the “Hazardous event” in Fig. 2.11, is shown to the right of the events that cause it. The top event in many applications is an accident, loss of control of a system, or a system failure. Events within the fault tree are often described as failures of different preventive *barriers*, safeguards, or controls that could prevent the hazardous event if they functioned as intended. The basic events are referred to as triggers because they can trigger barrier failures. The right side of the bow-tie diagram is an event tree describing the event sequences that can be caused or triggered by the hazardous event. Again, these are often conceptualized as failures of barriers intended to mitigate the adverse consequences of the hazardous event. Depending on what happens, various consequences can result from the occurrence of the hazardous event.

Bow-tie diagrams can be used qualitatively to identify and review engineering designs and operations and maintenance policies and practices that create barriers to prevent occurrence or to mitigate consequences of a hazardous event. They can also be used quantitatively to calculate probabilities of different outcomes (the “consequences” on the right edge of Fig. 2.11) and to study how they change if different barriers are added or removed. To these useful capabilities, reformulation of the bow-tie model as a BN adds several others (Khakzad et al. 2013), especially the ability to consider the causes and consequences of multiple hazardous events simultaneously. The same barriers may help to prevent or mitigate multiple types of accidents. Understanding the risk-reducing benefits of expensive investments in defensive barriers often requires an event network with a topology more complex than a simple bow-tie, with some trigger events able to cause many hazardous events

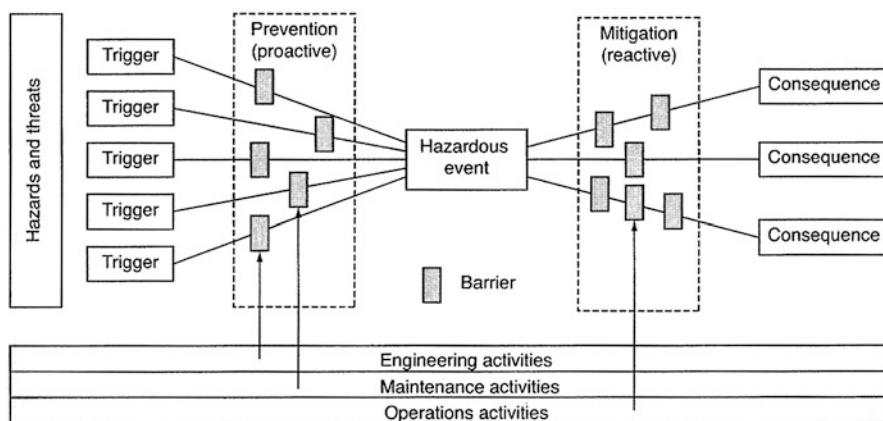


Fig. 2.11 A schematic picture of a bow-tie diagram. Source: https://en.wikipedia.org/wiki/Network_theory_in_risk_assessment

(e.g., an explosion leading to both fire and exposures of workers to chemical or radiological hazards). Bow-tie diagrams can be mapped to BNs, allowing more flexible inferences such as from observed occurrences of precursor events midway through the network to updated conditional probabilities of causes (trigger events) and outcomes or consequences.

Markov Chains and Hidden Markov Models

When repair activities are included in a model of system reliability, fault trees and event trees are replaced by probabilistic dynamic models in which components undergo stochastic transitions from working to failed and back. If the transition intensities (expressed in units of expected transitions per unit time) are constant, then such failure-and-repair processes can be represented as Markov chains. Markov chain models can readily be represented by DBNs. A DBN quantifies CPTs for the conditional probabilities of values of variables in the current time slice, given their values in the previous time slice; this is precisely the information required for a Markov chain model. In notation, a Markov chain specifies $P(X_{t+1} = x_{t+1}, | X_t = x_t)$ where x denotes a vector of values (e.g., 1 = working or 0 = failed) for the random variable vector X , and t indexes time period. A DBN provides this same information, albeit through multiple CPTs instead of one large transition matrix (with on the order 2^{2n} transition rates, or probabilities per unit time, for a system with n binary components and a square transition matrix having 2^n possible configurations of component states for its rows and columns). Fortunately, in most applications, not all variables depend on all other variables. This allows the full joint probability distribution of all variables in one period to be *factored* as a product of the probabilities of each variable conditioned only on the values of variables in its Markov blanket, which are revealed by the DBN DAG structure. The CPTs for each node suffice to compute the state transition probabilities $P(X_{t+1} = x_{t+1}, | X_t = x_t)$ without having to explicitly list the full transition probability matrix.

DBNs can also represent hidden Markov models (HMMs) in which the states are not directly observed but data or observations or signals, y , that depend probabilistically on the underlying states are available, $P(y_t, | x_t)$. DBN inference algorithms can then be used to estimate the current unobserved state from the sequence of observations to date and prior probabilities for states (known as *filtering* in systems dynamics and control engineering) and to predict the probabilities of future states and observations from past and present observations (Gharamani 2001). Applications of HMMs include estimating current disease states from observed patient symptoms and histories and predicting probabilities of future failures or degraded performance of systems from performance measurement logs (Vrignat et al. 2015).

Probabilistic Boolean Networks

A generalization of dynamic fault trees and event trees is the probabilistic Boolean network (PBN), in which each node represents a binary variable with possible values of 1 for on and 0 for off. The probability of each of these two possible states for a node in each period or time slice depends on its own value and on the values of its parents in the previous period. Node values make stochastic transitions between the states over time, creating a Markov chain. PBNs have been used to model gene regulatory networks. They can be represented as special cases of dynamic Bayesian networks (DBNs) (Lähdesmäki et al. 2006).

Time Series Forecasting Models and Predictive Causation

A traditional (discrete-time univariate) time series model in which the probability distribution for a variable in each time period depends on a finite number of its own past values can be represented by a BN with nodes representing the values of the variable at different times and with arrows directed from earlier to later values of the variable. This idea can be extended to multiple time series: the arrows in a DBN show how the currently probability distribution for each variable depends on past (and, if appropriate, current) values of other variables. CPTs quantify these probabilistic dependencies. Popular models for analysis of multiple time series, including vector autoregression (VAR) models, can thus be represented as special cases of DBNs, with CPTs specified via regression models and error distributions. Forecasting future values on the basis of what has been observed so far can then be accomplished by applying BN inference algorithms to compute the conditional probability distributions of unobserved future values given the observed values seen to date. Missing data are handled naturally in this framework: as in other BN inference problems, once simply enters findings for observed values, and posterior probabilities are then computed for values of all unobserved ones, including any unobserved (i.e., missing) past and present values, as well as future values.

In a DBN for multiple time series variables, one time series variable can be defined as a *predictive cause* of another if and only if arrows run from past or present values of the former to present or future values of the latter. That is, X is a predictive cause of Y if and only if the value of Y in a time slice has a probability distribution that depends not only on past values of Y (and perhaps other variables), but also on past values of X (and perhaps the current value of X , if causation within time slices is allowed). If there are no such arrows, meaning that the value of Y at any time is conditionally independent of the past and present values of X , given the past and present values of Y itself (and perhaps other variables), then X is not identified as a predictive cause of Y . Intuitively, X is a predictive cause of Y if and only if the history of X up to the current moment can be used to predict future values of Y better than they can be predicted just using the history of Y (or, more generally, just using the

histories of variables other than X) up to the current moment. This concept of predictive causation is often called *Granger causation* and was originally implemented by using statistical F tests to test whether data allowed rejection of the null hypothesis that mean squared prediction error for one time series, Y , is not significantly reduced by conditioning on the history of X as well as the history of Y itself (Granger 1969). If such Granger causation held in both directions, so that X and Y each help to predict the other, this was usually taken as evidence that some third variable affected both. The basic idea was subsequently greatly generalized and given a non-parametric foundation in information theory by testing whether information flows from one time series variable to another over time, so that conditioning on the history of X reduces the expected conditional entropy (uncertainty) for future values of Y , even after conditioning on other observed variables. This information flow between time series variables is called *transfer entropy* (Schreiber 2000). For the special case of traditional parametric time series with linear dynamics and Gaussian errors originally analyzed by Granger, transfer entropy specialized to Granger causality, i.e., information flows from X to Y if and only if X is a (Granger) predictive cause of Y . More generally, since information theory proves that any two random variables have positive mutual information if and only if they are not statistically independent of each other (Cover and Thomas 2006), the arrows in a DBN representing the data-generating process for multiple time series provide a simple visual way to identify predictive causation: it is indicated by arrows in the DBN directed from past and present values of one time series variable to future values (and possibly present ones, if instantaneous causation within time slices is allowed) of another. Thus, DBNs provide a natural framework not only for time series modeling and forecasting, but also for analysis of predictive causation.

Example: Bivariate Granger Causality Testing Using the Causal Analytics Toolkit (CAT)

Several on-line calculators are available to carry out Granger causality testing between two time series (e.g., www.wessa.net/rwasp_grangercausality.wasp), and it is also implemented in free software such as the granger.test function in the MSBVAR package in R. Rather than delving into each of these software products, we will illustrate Granger causality testing using the cloud-based version of the *Causal Analytics Toolkit* (CAT), a free software package providing algorithms and reports for causal analysis and model building and statistics. CAT integrates many R packages and creates reports that do not require familiarity with R to generate or interpret; hence, it is useful for illustrating what can be done with R packages without requiring readers to learn R. The CAT software can be run from the cloud via the link <http://cox-associates.com/CloudCAT>. The calculations performed using CAT can also be performed with appropriate R packages by readers adept at R programming and the R package ecosystem.

Figure 2.12 shows the Data screen for CAT as of 4Q-2017. (Updates may be made.) A few data sets are bundled with the CAT software, under the “Samples”

The screenshot shows the CAT software interface. On the left, a vertical menu lists various causal analysis commands: Data, Analyze, Bayesian, Causal, Correlations, Describe, Granger, Importance, Plot3D, Predict, Regression, Sensitivity, and Tree. The 'Data' command is currently selected, indicated by a blue background. The main workspace has several sections:

- Data Input:** A 'Upload File' button with a tooltip 'Upload .csv .xlsx .xls file. First row must be column names.' and a message 'No file selected'.
- Samples:** A dropdown menu labeled 'Data' with options: LA, asthma, mutagens, mtcars, iris, and a 'TE key' link.
- Optional Column Selection:** A section with the message 'Optional: Select columns. If no selection, all columns are used in order. Dependent variable must be first, drag and drop to move.' and a note 'Optional: Select/deselect all columns. To delete multiple items in selection box, use Control or Shift key to select.'
- Optional Discretization:** A section with the message 'Optional: Select integer variables to make discrete:' and checkboxes for 'AllCause75', 'tmin', 'tmax', 'month', 'day', and 'year'. The 'month', 'day', and 'year' checkboxes are checked.
- Data Table:** A table titled 'Show 10 entries' containing the following data:

	AllCause75	PM2.5	tmin	tmax
1	151	38.4	36	72
2	158	17.4	36	75
3	139	19.9	44	75
4	164	64.6	37	68
5	136	6.1	40	61
6	152	18.8	39	69
7	160	19.1	41	76

Fig. 2.12 Loading data into the Causal Analytics Toolkit (CAT). *Source:* CAT software at <http://cox-associates.com/CloudCAT>

Data drop-down menu on the upper right. This example will use the LA air pollution, weather variable, and elderly mortality data set introduced in Chap. 1. Recall that, in this data set, *AllCause75* gives daily mortality counts for people 75 years old or older in California's South Coastal Air Quality Management District (SCAQMD), near Los Angeles.

(User data files can be uploaded using the “Upload File” browser bar at the upper left. Subsets of columns can be selected for analysis if desired, but the default is to use all of them, and we will do so in this example.) We specify that the month and year variables, which are additional columns to the right of those shown in Fig. 2.12, should be modeled as discrete variables, rather than as continuous, by checking them in the optional row near the middle of the Data screen. With these data-loading and preparation preliminaries accomplished, Granger causality testing can now be performed by clicking on “Granger” in the list of commands in the left margin. (Clicking on a menu icon causes this list of commands to appear when it is not showing.) Doing so generates the output in Fig. 2.13. Like most CAT reports, this one is formatted as a sequence of increasingly detailed summaries and supporting analyses that can be scrolled through from top to bottom. The top part summarizes lists of significant Granger causes with 1-day lags, including the calculated *p*-values (based on F tests) for rejecting the null hypothesis that one variable does not help to predict another. The bottom table (most of which is not shown in Fig. 2.13) provides supporting details by listing F statistics with *p*-values for each bivariate Granger test performed. In this data set, most pairs of variables help to predict each other (e.g., with a 1 day lag, *AllCause75* is a predictor of *month* as well as *month* being a

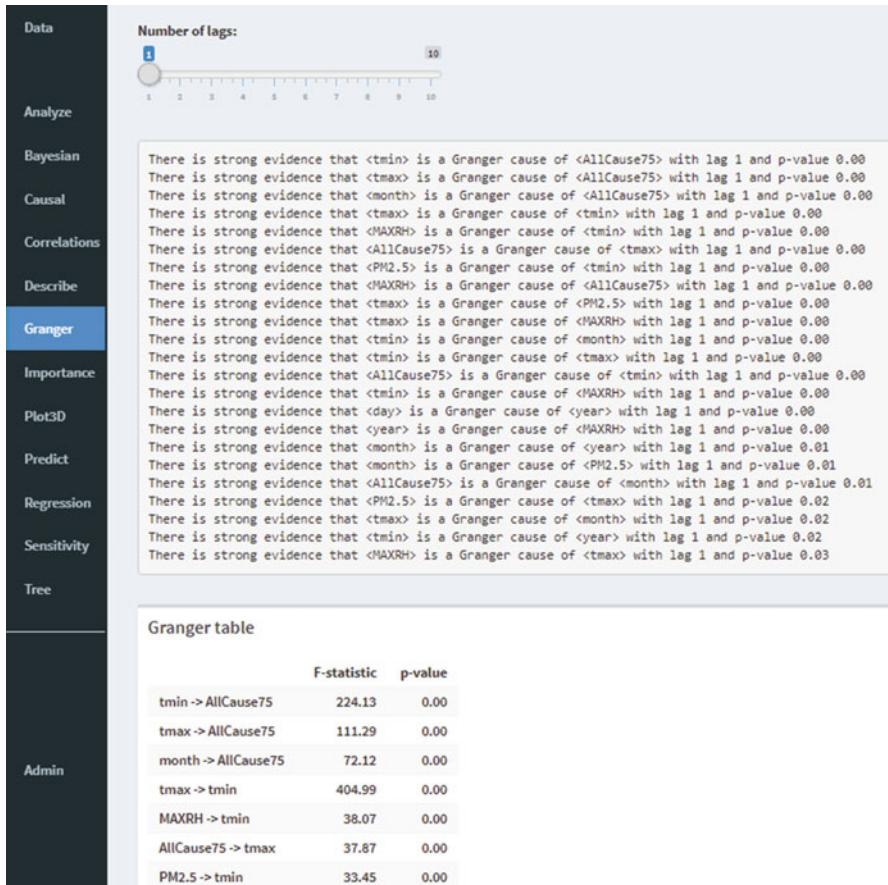


Fig. 2.13 Results of Granger causality testing in CAT

predictor of *AllCause75*), suggesting that most have common causes or other dependencies that are not well revealed by bivariate Granger causality testing. The slider at the top of the output allows similar results to be displayed for longer lags. Overall, these findings indicate a need for multivariate analysis to better distinguish which variables might directly cause which others: the Granger analyses show little more than that most variables are associated with each other over time, and hence are useful for predicting each other over time.

To overcome limitations such as the ones illustrated in this example, where most variables are correlated with time and hence with each other and appear to be Granger causes (significant independent predictors) of each other in bivariate tests, several groups have developed multivariate Granger causality tests and software (e.g., the FIAR package in R for stationary time series variables and the MVGC

toolkit in MATLAB (Barnett and Seth 2014)). However, the framework of Granger causality testing is somewhat fragile, insofar as it assumes specific parametric families of time series models (e.g., vector autoregression (VAR) models) and stationarity of the tested time series. The basic idea can be applied using more robust non-parametric methods, discussed later (e.g., Random Forest ensembles) or dynamic Bayesian networks to determine whether future values of a dependent variable are conditionally independent of the history of a hypothesized cause up to the present, given the histories of itself and other variables. If so, then the hypothesis of predictive causality between them is not supported.

The concept of predictive causation is attractive for several reasons. One is that it includes and refines the intuitive requirement that causes must precede their effects. Predictive causation requires not only this, but also that the history of causes up to a given moment must provide information about the future values of the effect. According to information theory, this implies that the causes help to predict the effect values in future periods, in the sense that conditioning on information about present and past values of causes reduces the expected conditional entropy of the probability distributions for at least some future values of the effect; more colloquially, knowing the causes is expected to reduce uncertainty about their future effects. Another attractive feature of predictive causation is that it is relatively objective, in that statistical tests are available for discovering whether data allow confident rejection, at a specified level of statistical confidence, of the null hypothesis that future values of one variable (the hypothesized effect) are conditionally independent of past and present values of another value (a hypothesized possible direct cause) given the values of other variables. (CART trees, discussed in the following example, provide one way to test this hypothesis, within the limitations imposed by the CART tree algorithm and the sample size and design.) Thus, different statisticians should be able to analyze the same data using the same software and reach the same conclusions about whether one variable can be confidently identified as a predictive cause of another.

Despite these advantages, however, predictive causation does not imply either manipulative causation or explanatory/mechanistic causation. A standard counter-example is that having nicotine-stained fingers might be a *predictive* cause of lung cancer in a data set that records both but that does not include smoking behavior: seeing nicotine-stained fingers might be a reliable indicator of increased future risks of lung cancer, providing predictively useful information not available from other variables in the data set. But it would not necessarily be a *manipulative* cause: keeping one's fingers unstained would not necessarily decrease future risk of lung cancer unless the only way to have unstained fingers is not to smoke. Nor is it a *mechanistic* cause: changes in nicotine staining of fingers do not propagate through a set of mechanisms of to alter risk of lung cancer. In this example, the data set violates the Causal Markov condition (CMC), making it impossible for an algorithm to determine whether nicotine stained fingers are only a predictive cause, or also a manipulative cause or a mechanistic cause, of lung cancer risk.

Structural Equation Models (SEMs), Structural Causation, and Path Analysis Models

Long before Bayesian network models and other probabilistic graphical models were introduced, econometricians and artificial intelligence researchers were already using *structural equation models* (SEMs) to model causal relationships and statistical dependencies among variables (Simon 1953; Simon and Iwasaki 1988). A structural equation shows how a dependent variable depends on other variables—namely, the variables that determine its value, represented by its parents in a directed graph model—via an equation such as

$$\text{output} = f(\text{parent}_1, \text{parent}_2, \dots, \text{parent}_n, \text{error}) \quad (2.18)$$

Equation 2.18 signifies that the value of the dependent variable, here called *output*, depends via some (possibly unknown) function f on the values of other variables, namely its parents, and on a random variable, the *error* term. This error term represents the effects of all other determinants of the value of the dependent variable that are not otherwise represented in the model. It is usual to interpret each such structural equation as representing a causal mechanism, law, or constraint and to assume that its error term is independent of the error terms for other structural equations. It is important to test this assumption in practice, as correlated error terms may betray the presence of an unobserved common cause that should be included in the model as a latent variable, or of a selection bias that must be corrected for, before estimates of the functional relationship between its parents and the dependent variable can be interpreted causally.

The desired causal interpretation of a structural equation is usually the *manipulative causal interpretation*: that exogenously changing the value of a variable on the right side (i.e., a parent of the dependent variable) will cause the value of the dependent variable to change to restore equality between the left and right sides. In this case, the structural equation represents a causal model describing how the value of the dependent variable on the left is determined by the values of its parents on the right (Druzdzel and Simon 1993). A system of such equations is called a *structural equation model* (SEM) if each function determining the value of a variable is invariant to changes in the forms of functions determining the values of other variables (Pearl 2009). The intuition is that each equation describes how the values of its parents (including the random error term) determine the value of a variable, and that this should not depend on how other variables are determined if the equation describes a causal law or mechanism. Non-parents are considered to be ineligible to be direct causes of a variable.

The *structure* of an SEM model is given by the DAG showing the parents of each variable. If each equation of the form (2.18) in an SEM can be expressed by an equivalent CPT, as

$$P(\text{output} = y | \text{parents} = x)$$

where y is a value of the dependent variable (*output*) and x is a vector of values for its parents, and if the resulting DAG is acyclic, then the SEM is simply an alternative notation for a BN. Conversely, any BN with discrete random variables can be expressed as an equivalent SEM in which the value of each node is a deterministic function of the values of its parents and of a single independently uniformly distributed error term, which may be thought of as an unobserved (latent) variable that also affects its value (Druzdzel and Simon 1993). The BN and SEM are equivalent in the sense that they represent the same joint probability distribution of the variables and have the same DAG. In a SEM with a known DAG showing which variables are derived from which others, X can be defined as a *direct structural cause* of Y if X appears in the equation determining Y , i.e., X is a parent of Y in the DAG. Similarly, X can be defined as an *indirect structural cause* of Y if X is an ancestor of Y , but not a direct parent. The intuition behind structural causality is that *effects are derived from their causes*, but causes are not derived from their effects; thus, a parent or ancestor can be a structural cause of it, but a descendent cannot. This interpretation is slightly different from the usual one in causal BN models, which states that *probability distributions for the values of effects depend on the values of their direct causes*. The concept of *deriving* the value for one variable (the effect) from the values of others (its direct causes, including the unobserved value of the error term) via a structural equation is consistent with, but distinct from, the idea that the value of the effect variable *depends on* (i.e., is not statistically independent) of the values of its direct causes.

SEMs with all linear equations and acyclic DAGs are called *path analysis* models. These were the first types of causal DAG models studied, starting with work by geneticist Sewell Wright around 1920, and they occupied much of causal modeling in genetics, biology, and social sciences in the twentieth century (Pearl 2009). Path analysis models generalize multiple linear regression by allowing for linear dependencies among explanatory variables. In a path analysis DAG, the arrows between standardized variables have weights called path coefficients indicating the effect of a unit change in the variable at an arrow's tail on the variable into which it points, i.e., at its head. The total effect of a change in a variable X on a variable Y is calculated by multiplying coefficients along paths and summing over all paths from X to Y . BNs allow much greater flexibility in quantifying causal relationships between variables and their parents than path diagrams by replacing path coefficients with conditional probability tables (CPTs) that can represent arbitrary, possibly very nonlinear, effects and interactions among variables.

Influence Diagrams

The probability of an event occurring in a certain time period can depend not only on occurrences of previous and contemporaneous events, but also on prior or contemporaneous actions by one or more decision-makers. More generally, the conditional probability distribution of a random variable in a period can be changed by actions as

well as by values of random variables in that period or earlier. As mentioned in Chap. 1, influence diagrams (IDs) are generalizations of Bayesian networks that include actions (also called decision nodes) and value or utility nodes that evaluate the consequences of the actions. These additions seem to constitute a clear generalization of BNs. However, with a little ingenuity, the same algorithms used to find the most probable explanations (MPEs) in a BN can also be used to solve for optimal decisions in IDs, meaning decisions that maximize expected utility (Mauá 2016). Every ID problem can be transformed to an equivalent BN inference problem by replacing the decision nodes with chance nodes having equal probabilities for all decisions and replacing utility nodes with binary (0–1) chance nodes with CPTs that assign a conditional probability for 1, given the values of the parents, proportional to the utility of that combination of values (Crowley 2004). The decisions that best explain a utility value of 1 (in the sense of MPE) are then the optimal decisions. Although several algorithms have been developed specifically for inference and decision optimization in IDs (Shachter and Bhattacharjya 2010), the fact that an ID decision optimization problem can be automatically mapped to an equivalent BN inference problem means that BN algorithms can be used to solve both.

Example: Decision-Making with an Influence Diagram—Taking an Umbrella

To illustrate the use of ID software in solving a textbook-type decision analysis problem, consider a decision-maker (d.m.) who must decide whether to take an umbrella to work on a day when the prior probability of rain is 0.40. The d.m.’s von Neumann-Morgenstern utility function, scaled to run from 0 for the least-preferred outcome to 100 for the most-preferred outcome, is as follows: utility = 100 if the weather is sunshine and the d.m. does not take the umbrella; utility = 0 if there is rain and the d.m. does not take the umbrella; utility = 80 otherwise, i.e., if the d.m. takes the umbrella. (Thus, the utility function can be represented as utility = $100 - 20 \cdot \text{TAKE} - 100 \cdot (1 - \text{TAKE}) \cdot \text{RAIN}$ where TAKE is a binary (0–1) decision variable with value 1 if the d.m. takes the umbrella and value 0 otherwise; and RAIN is a binary random variable with value 1 if it rains and 0 otherwise.) The d.m. can also obtain a forecast that has an 80% probability of being correct, i.e., of predicting rain if it will rain and of predicting that it will not rain (“Sunshine”) if it will not rain; hence there is probability 20% that the forecast will falsely predict rain when it will not rain and a 20% probability that it will falsely predict sunshine when it will rain. Entering these numbers into the tables for the *Forecast* and *Consequence_utility* nodes in Fig. 2.14a creates an ID model. Readers interested in following the details of building and using ID models should create this network in Netica. (The free “Limited” use mode suffices for this purpose.)

Note that Netica allows a distinct utility node (indicated by a hexagon) and decision nodes, which do not have probability bars. With these node types, creating an ID is as easy as creating a BN. The utility node, like the chance nodes we have already studied, has a table, but it shows the utility for each combination of values of

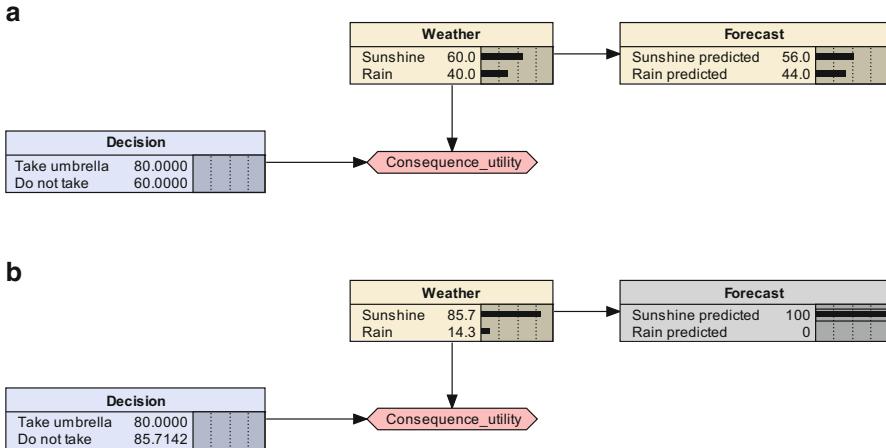


Fig. 2.14 (a) An ID for the umbrella decision. With only prior information ($P(Rain) = 0.40$), it is optimal to take the umbrella. (b) With a forecast of sunshine, the optimal decision changes to not taking the umbrella

its parents, rather than the conditional probability for each combination of values of its parents, as in the CPT of a chance node.

As soon as the ID in Fig. 2.14a is completed and the Netica network is compiled, the decision node starts showing the expected utility for each alternative decision. Without forecast information, the expected utility of taking an umbrella is 80 (since we specified the utility to be 80 whether or not it rains if the d.m. takes an umbrella) and the expected utility of not taking it is 60 (the 0.60 probability of Sunshine times the utility of 100 plus the 0.40 probability of Rain times the utility of 0). (The probability that the forecast will predict rain is $0.8 \cdot 0.4 + 0.2 \cdot 0.6 = 0.32 + 0.12 = 0.44$, as shown in the Forecast node.) Thus, the optimal decision in the absence of forecast information is to take the umbrella. If forecast information is used, and the forecast predicts sunshine, then the optimal decision is to not take the umbrella, with an expected utility of about 85.71, as shown in Fig. 2.14b. The expected utility increases from 80 in the absence of forecast information to $0.56 \cdot 85.7142 + 0.44 \cdot 80 = 83.2$ with the forecast information. This is because there is a probability of 0.56 that the forecast will be for sunshine, in which case the decision changes to not taking the umbrella and expected utility increases from 80 to 85.7142. This increase in expected utility determines the *value of information* (VOI) from the forecast: the d.m. should be willing to pay up to the amount that would reduce expected utility from 83.2 to 80. If the d.m. is risk-neutral and utilities are scaled to correspond to dollars, then the d.m. should be willing to pay up to \$3.20 for the forecast information. If the forecast information costs more than that to acquire, the d.m. should forego it and simply take the umbrella. Such calculations of optimal decisions, expected utilities, and VOI are the staples of decision analysis. BN technology, extended to apply to IDs, makes them easy to carry out.

Decision Trees

Adding decisions and utilities to a BN makes it an ID. Adding decisions and utilities to an event tree makes it a *decision tree*. Utilities are placed at the tips of the tree, i.e., evaluating the outcomes. Decision nodes can be inserted into event sequence paths, with decisions being conditioned on the previous events and decisions in the paths leading to them from the root of the tree. Every decision tree can be mapped automatically to an equivalent ID, and every ID can be automatically re-expressed as a decision tree; thus, IDs and decision trees are logically equivalent classes of models for representing decision problems, although either may have practical advantages over the other in representation, storage, and computational efficiency for specific problems (Shachter 1986). Since every ID can be mapped to an equivalent BN inference problem (Crowley 2004), decision trees can also be represented and solved using BNs.

Markov Decision Processes (MDPs) and Partially Observable MDPs (POMDPs)

Markov decision processes (MDPs) and partially observable MDPs (POMDPs) can also be automatically mapped to influence diagrams or dynamic Bayesian networks in which the CPTs depend on the decisions made (Boutilier et al. 1995; Jonsson and Barto 2007). Indeed, DBNs provide a basis for some of the most computationally efficient algorithms known for solving large-scale MDPs and POMDPs, including hierarchical POMDPs with observed signals and transitions for high-level states consisting of sequences of observations and transitions generated by lower-level POMDPs (Theocharous et al. 2004). A straightforward representation of a POMDP by an ID represents states, observations, and rewards in each period by chance nodes and represents the choice among acts in each period by a decision node. The DAG structure of the ID makes the reward in each period depend on the state and act in that period and, optionally, on the state in the next period; the conditional probability distribution for the next state depends on the current state and act; and the conditional probability distribution for observations in a period depend on the state in that period. The act in each period is a decision node that is informed by observations in that period and earlier and by nothing else (Poole and Mackworth 2017). This representation of a POMDP allows ID algorithms, and hence BN algorithms, to be applied to solve them, although the efficiency of computations can be improved by using specially devised rather than general-purpose algorithms to exploit the structure of the POMDP graph (Theocharous et al. 2004).

Predictive Causality and Predictive Analytics Models

The intuitive idea that *causes help to predict their effects* can be applied to screen for causation outside the context of time series analysis, helping to identify potential causal relationships between variables in a data set on the basis of whether they help to predict each other's values. The discipline of predictive analytics is now well advanced and can be applied to cross-sectional data as well as to time series data. Machine learning software packages such as the Caret package in R are freely available to automate the predictive analytics tasks of splitting data into disjoint training and test sets, optimizing choices of input parameters or settings for each of many parametric and non-parametric predictive models and algorithms, and comparing their predictive performances quantitatively using standard metrics such as sensitivity, specificity, and balanced accuracy. Such software can be used to identify which other variables help to predict the values of a user-specified dependent variable. The logic of predictive causality testing identifies these predictors as its potential direct causes in a data set, and screens out variables that do not help to predict it as not being potential direct causes.

Classification and Regression Tree (CART) Models

Classification and regression tree (CART) algorithms estimate the conditional mean—or, more generally, the conditional mean, standard deviation, and distribution—of a dependent variable based on the values of other variables that help to predict the distribution of its values. They are also called *recursive partitioning* algorithms because they work by partitioning the cases in a data set (its records or rows) into subsets with significantly different distributions of the dependent variable, and then recursively partitioning each subset until no further partitions with significantly different distributions of the dependent variable can be found. Each partition is called a “split” and is represented by a node in a tree, with branches below it showing how the data set is partitioned at that node. A split is formed by conditioning the distribution of the dependent variable on contiguous intervals of values for a continuous predictor variable, or on disjoint subsets of values for a discrete predictor variable. The predictor variable being conditioned on (or split on) is represented by a node in the tree, and the branches emanating from it represent the different sets or ranges of values being conditioned on to get significantly different distributions of the dependent variable. For example, if the probability of an illness depends on age, then consecutive ranges of age (a node variable) could be used as splits to estimate different conditional probabilities of illness given age ranges. Successive splits condition on information about values of additional predictors until no more information can be found that generates significantly different conditional distributions for the dependent variable. The next node to split on is typically selected via a heuristic such as choosing the variable that is estimated to be most informative about

the dependent variable (i.e., maximizing reduction in expected conditional entropy of the dependent variable when split on), given the information (splits) conditioned on so far. Although many refinements have been made in recursive partitioning algorithms since the 1980s, including allowing parametric regression modeling for split selection if desired and avoiding over-fitting by pruning trees to minimize cross-validation error, all variations of CART algorithms still use the key idea of identifying variables that are informative about a dependent variable and then conditioning on values of these informative variables until no further valuable information for improving predictions can be found.

Example: Seeking Possible Direct Causes Using Classification and Regression Tree (CART) Algorithms

In the CAT software, a non-parametric classification tree is generated simply by clicking on the “Tree” command in the command menu at the left side of Fig. 2.12. By default, the dependent variable is taken to be the first column in the data table. The user can select a different dependent variable if desired. Figure 2.15 shows the result of applying the “Tree” command to the LA data set in Fig. 2.12, with daily elderly mortality counts, *AllCause75*, as the dependent variable. (The tree is generated by the party package in R, <https://cran.r-project.org/web/packages/party/party.pdf>, which contains full documentation. Month and year are specified as categorical variables.) The interpretation is as follows. Each “leaf” node at the bottom of the tree shows the conditional expected value for cases (in this case, days, corresponding to the rows in the data table in Fig. 2.12) that the leaf node describes. These are the “y” values for the leaf nodes. The “n” values show how many cases are described by each node. For example, the left-most leaf node (numbered node 5) in Fig. 2.15 describes $n = 29$ cases (days) out of 1461 records (rows) in the data set. The average

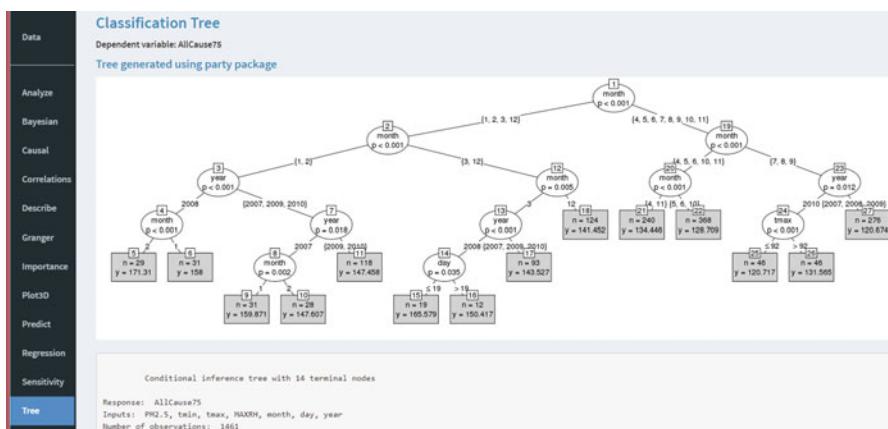


Fig. 2.15 A CART tree for the LA air, weather, and elderly mortality data set in Fig. 2.12

value of the dependent variable, *AllCause75*, is $y = 171.31$ deaths per day for the days described by this node. The description for any leaf node is found by tracing the path from the root node at the top of the tree (node 1) to the leaf node. The conjunction of values (shown along the branches) for the variables (in the nodes) along the path constitute the description of the leaf node to which they lead. For example, the description for node 5, the left-most leaf node in Fig. 2.15, is as follows: February of year 2008. This is found as the conjunction or intersection of months 1, 2, 3, and 12, i.e., December through March, for the left branch below node 1 with months 1 and 2 along the left branch below node 2 and month 2 along the left branch below node 4; together with the year 2008, shown on the left branch below node 3. Similarly, the description for node 27, the right-most leaf node, is July–September of years 2007–2009. The average daily elderly mortality count is only 120.7 deaths per day during these 276 days, compared to 131.6 deaths per day during the hot days (with high temperatures over 92°) in 2010 (node 26) and 171.3 deaths per day in February of 2008 (node 5). Thus, the tree-growing algorithm has discovered, or revealed, that the highest daily death counts occur in the winter months and the lowest occur in the summer months.

In CART trees, as well as in BNs, what is *not* shown in a diagram is often even more interesting and useful than what is shown. In Fig. 2.15, the fact that month appears in the tree as a key predictor of daily elderly mortality counts, but that same-day temperature *per se* does not (with the sole exception of the 2010 summer heat wave, where daily maximum temperatures over 92 were associated with a slight increase in the relatively low same-day daily mortality counts, as can be seen by comparing leaf nodes 25 and 26), suggests that it was not same-day low daily temperatures that directly caused high elderly mortality counts in February of 2008, but something else, for which month is a proxy.

Although Fig. 2.15 illustrates a classification tree treating the data as cross-sectional, lagged values of the different variables can also be used as predictors and included in tree-growing, thus combining time series information and same-period cross-sectional information. Figure 2.16 shows the result of such a dynamic analysis. The dependent variable is *AllCause75.lags.plus7*, which is the default name for daily elderly mortality count 7 days in the future. Values for all variables from 7 days in the future back to the present day were included as potential predictors. The resulting tree in Fig. 2.16 shows that the best predictor (first split at the top of the tree) for these deaths a week in the future is the low temperature today (*tmin*). In fact, lagged values of temperatures and *AllCause75* itself are the only significant predictors of elderly mortality counts in this data set, showing that month and year were acting as surrogates for these values in Fig. 2.15. Neither maximum relative humidity nor fine particulate matter (PM2.5) appears in the tree, indicating that in this data set, elderly mortality counts are conditionally independent of present and lagged values of these variables, given lagged values of elderly mortality counts and temperatures. Thus, daily minimum and maximum temperatures are identified as the only predictive causes of elderly mortality in this data set detected by classification tree analysis. Such analysis provides a non-parametric alternative to multivariate Granger causality testing.

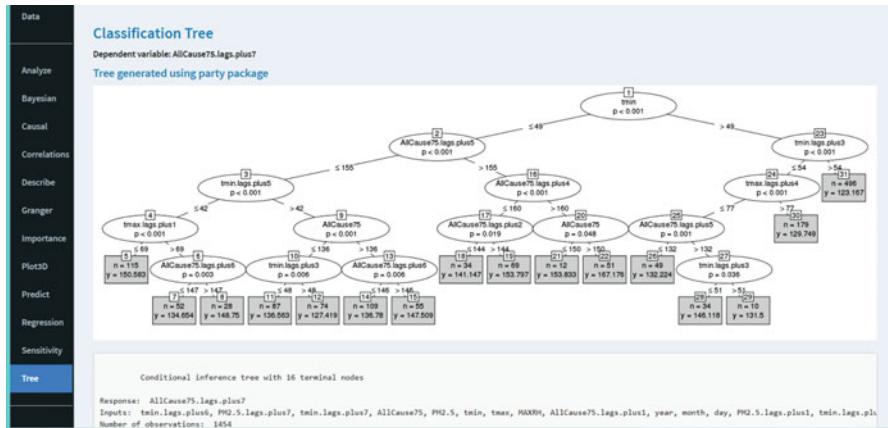


Fig. 2.16 A tree with lagged values reveals that temperatures are predictive causes of elderly mortality

CART trees are closely related to BNs. Indeed, tree-growing algorithms can be used to help identify the DAG structure of a BN from data, as follows. Ideally, in a classification tree, the dependent variable should be conditionally independent of the variables not in the tree, given the variables that are in it. (If this were not so, then the information provided by at least some of the variables omitted from the tree could be used to improve predictions of the dependent variable, since it would depend in part on them.) Thus, the tree can be used as an approximate test for conditional independence. The variables that appear in it should ideally be those in its Markov blanket, i.e., its parents, children, or spouses. The qualifiers “ideally,” “approximately,” and “should” are included because real CART algorithms are not perfect oracles for establishing conditional independence: they are limited by finite sample sizes, sampling variability, and myopic selection of next variables to split on, and hence they may fall short of identifying all and only the variables on which the dependent variable depends. But in practice, they often provide a useful way, albeit only an approximate one, to identify conditional independence among variables in large data sets. Since the DAG structure of a causal Bayesian network represents conditional independence relationships among its variables, algorithms for identifying conditional independence from data can also help to discover these DAG structures.

Example: Conditional Independence Tests Constrain DAG Structures

Problem setting: Suppose that we had an ideal CART algorithm that provided a perfect test for conditional independence by generating trees that contain all and only those variables that the dependent variable depends on, i.e., that share positive mutual information with it, even after conditioning on other variables. Suppose

also that we had a large data set with three columns, or variables, X , Y , and Z , which we will interpret as exposure, health, and lifestyle variables, respectively.

Problem: How could the CART algorithm be applied to the data set to determine which of the following DAGs models 1–5, if any, is consistent with the data?

1. $X \leftarrow Z \rightarrow Y$ (e.g., exposure \leftarrow lifestyle \rightarrow health)
2. $Z \rightarrow X \rightarrow Y$ (e.g., lifestyle \rightarrow exposure \rightarrow health)
3. $X \rightarrow Y \leftarrow Z$ (e.g., exposure \rightarrow health \leftarrow lifestyle)
4. $X \rightarrow Y \rightarrow Z$ (e.g., exposure \rightarrow health \rightarrow lifestyle)
5. $X \rightarrow Z \rightarrow Y$ (e.g., exposure \rightarrow lifestyle \rightarrow health)

Solution: The key is that each model implies certain conditional independence relationships. These implications can be tested by using the CART algorithm to determine which conditional independence relationships hold among the variables in the data set. The conditional independence implications of the different models are as follows:

1. Models 1 and 5 (but none of the others) imply that health response Y is conditionally independent of exposure X given covariate Z . Thus, if either of these models is correct, then a CART tree for Y as the dependent variable should have Z in it but not X , and this pattern suffices to narrow down the choice of models to model 1 or model 5. Two models that have the same conditional independence implications, such as models 1 and 5, are said to be *Markov-equivalent*, or to belong to the same Markov equivalence class. Even a perfect algorithm for determining conditional independence relationships among variables from data cannot distinguish among models in the same Markov equivalence class, and other principles must be used to uniquely identify DAG structures in such cases. For example, in model 1, changes in Z should precede resulting changes in X , but in model 5 this temporal precedence order is reversed.
2. Only in model 2, but not in the other models, are Y and Z are conditionally independent given X . Thus, a CART tree for Y that only has X in it but not Z , uniquely identifies model 2 as the correct one from among models 1–5.
3. Only in model 3 are X and Z unconditionally independent, but conditionally dependent given Y . A CART tree for Y would include both X and Z , which is also true for model 4, but a CART tree for Z that is constrained to split only on X would do so if model 4 generated the data (assuming that information is transmitted from X to Z via Y) and would not do so for model 3, since Z and X are unconditionally independent in model 3.
4. Only in model 4 are X and Z unconditionally dependent, but conditionally independent given Y . Thus, if this model generated the data, then if a tree for Z is split first on X only, then X will enter the tree. But if a tree for Z is grown with both X and Y as allowed variables to split on, then Y will enter the tree and X will not. This combination does not hold for any of the other models.

Thus, an ideal CART algorithm can be used to uniquely identify any of models 2, 3, or 4 or to discern that either model 1 or model 5 generated the data, if it is known that one of models 1–5 generated the data. Similar techniques can be used to

discover or constrain which DAG models might have generated observed data even in the absence of *a priori* knowledge about the data-generating process. This possibility is examined later in the section on causal discovery algorithms.

The second major connection between classification trees and BNs is that CART algorithms can be used to estimate the conditional probability tables (CPTs) for a BN from data if its DAG structure is known. The recipe is simple: grow a tree for each variable in the BN using only its parents as predictors. Each leaf of the resulting tree represents a conjunction of values (or value ranges) for the parents with a corresponding conditional frequency distribution for the dependent variable—precisely what is needed for its CPT. Moreover, the tree algorithm automatically discretizes continuous predictors, partitioning (“splitting”) them into consecutive intervals; and it only creates splits that lead to significantly different conditional probability distributions for the dependent variable. This provides an efficient way to store CPT information—typically far more efficient than listing each possible combination of parent values and child values with a conditional probability for each. For example, the tree in Fig. 2.16 has only 16 leaf nodes, yet was grown for a data set with 64 variables (*year*, *month*, *day*, *AllCause75*, *PM2.5*, *tmin*, *tmax*, *MAXRH*, each lagged by 0–7 days) each with many possible values. Even if each variable had been quantized to only ten deciles, there still could have been up to 100 million combinations to assign probabilities to (allowing for the possibility of a different probability value for each possible combination). A tree with 16 leaf nodes is clearly a much more tractable way to store the information that matters. To use such a tree to find the conditional probability that a dependent variable Y has specific value y , given the values of its parents, one simply “drops” the values of the parents through the tree—that is, applies the successive splits in the tree to the parents’ values to determine which leaf of the tree describes them. Following the path through the tree (the sequence of branches emanating from successive nodes, from the root at the top of the tree to a leaf node at its bottom) determined by the values of the parents leads to a unique leaf node. For example, the tree in Fig. 2.16 would classify a summer day that occurred after a week of days with temperatures always above 60° as belonging to leaf node 31 at the far right, since this case meets its defining conditions, $tmin > 49^{\circ}\text{F}$ and $tmin.lags.plus3 > 54^{\circ}$. Node 31 is the most common leaf (i.e., classification) in this tree, with 496 cases. Whenever the low temperature is over 49° today and is over 54° 3 days from now, the day 7 days from now is classified as belonging to leaf node 31, with a conditional mean value of $E(Y| \text{node 31}) = 123.17$ for elderly mortalities. The estimated conditional probability of any specific value for a case at that node, such as $P(y = 130 \text{ deaths} | tmin > 49, tmin.lags.plus3 > 54)$ is then just the fraction of all 496 cases at that node that have that particular y value. The values of other parents are irrelevant once a leaf node has been reached: all combinations for their values give the same estimated conditional distribution for the dependent variable (since otherwise the tree-growing would have continued). In practice, BN learning programs typically discretize, or bin, the values of continuous variables or variables with many possible values into ten deciles, and then quantify their empirical relative frequencies, which are point estimates of their probabilities. (Some Bayesian approaches adopt a Dirichlet prior and present

posterior distributions rather than point estimates for the probabilities of specific values.) Programs such as Netica display the bar charts for the probabilities of the discretized values of random variables directly on the node. No matter how such final processing and display details are handled, however, the CART tree structure provides a highly parsimonious way to store empirical CPT information.

A CART tree requires specifying a single dependent variable. A BN can be thought of as summarizing the results from multiple interrelated CART trees, one for each variable. As an example, consider the Bayesian network (BN) in Fig. 2.17. This BN was learned from the LA data set in Fig. 2.12 using the “Bayesian” command in the CAT menu, which draws on the *bnlearn* package in R, as described in more detail in the section on causal discovery algorithms.

Each node with inward-pointing arrows can be viewed as having a corresponding classification tree, with the variables that point into it appearing in the tree and with the CPT for the node consisting of the conditional probability distributions for its value specified by the leaf nodes of the tree. To view a node’s tree explicitly, one can use the “Tree” command at the bottom of the menu on the left side of the screen. For example, Fig. 2.18 shows a tree for *AllCause75* with *month* as its parent (and with only the conditional means values, rather than the entire conditional distributions, for the dependent variable *AllCause75* displayed at its leaf nodes). The BN DAG display in Fig. 2.17 suppresses the details of the tree for each node, showing only

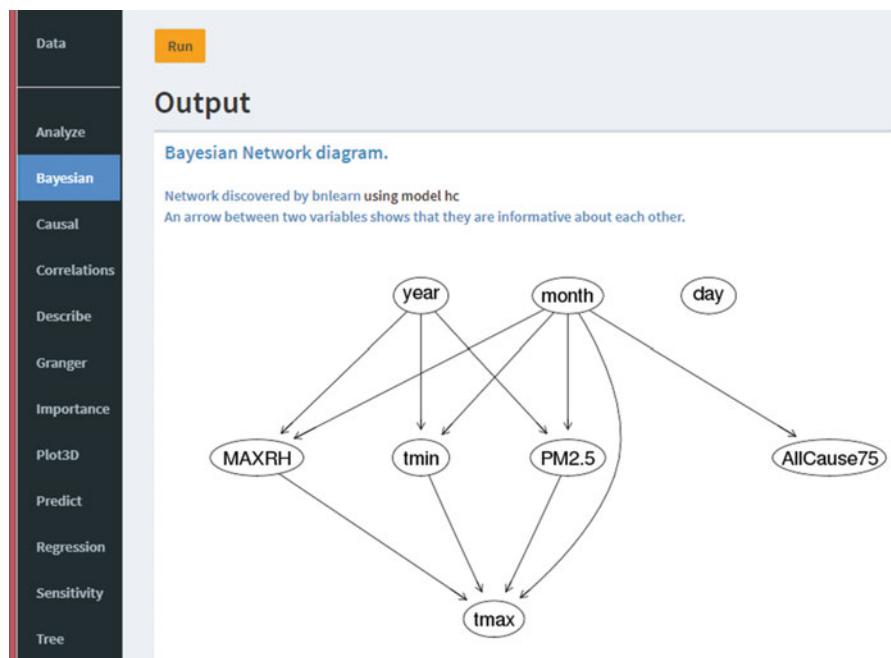


Fig. 2.17 A BN for the LA air, weather, and elderly mortality data set in Fig. 2.12

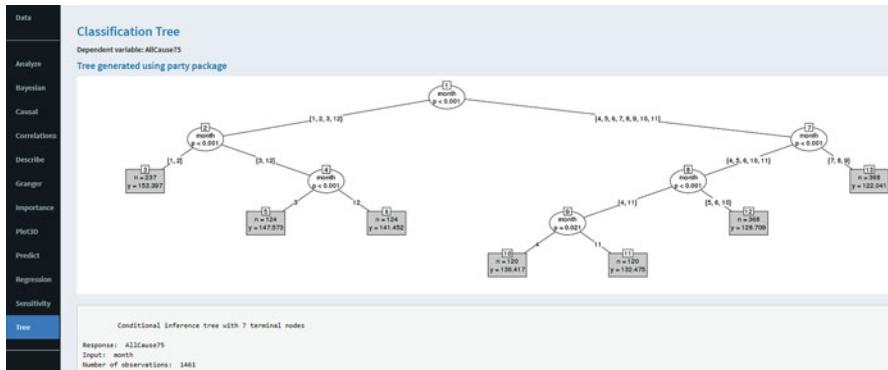


Fig. 2.18 A classification tree for the *AllCause75* node in Fig. 2.17

its parents and not the CPT information on how their values determine the conditional probability distribution for the node. But it has the advantage of showing the parents for all variables simultaneously. The *bnlearn* package uses different detailed algorithms than the “Tree” procedure to generate its CPTs and DAG structures, including binning continuous variables into ten deciles rather than using splits to quantize them. Nonetheless, viewing a BN conceptually as consisting of a collection of classification trees, one for each node, with the parents of a node (if any) appearing in its tree and with the CPT for the node determined by the leaf nodes of its tree (or simply specified as a marginal probability distribution, for an input node having no parents), shows that BNs can be interpreted as multivariate generalizations of CART trees that remove the restriction of considering only one dependent variable at a time.

The Random Forest Algorithm: Importance Plots and Partial Dependence Plots

Suggestive and useful as CART trees are, individual trees are often not robust, meaning that growing trees on several random subsets of the data may lead to quite different trees. This problem is addressed in modern machine learning by using *model ensembles*, i.e., collections of hundreds of individual models (e.g., CART trees). These are generated from samples from the original data set. Ensemble-based predictions are formed by averaging the predicted values for the dependent variable, such as its conditional mean values at the leaf nodes of the trees, given the values of its predictors, over all the models in the ensemble. A popular non-parametric ensemble technique for predictive analytics is the Random Forest algorithm, for which documentation and an R implementation are available via the *randomForest* package in R (<https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>). Python programmers can use `sklearn.ensemble.RandomForestClassifier`.

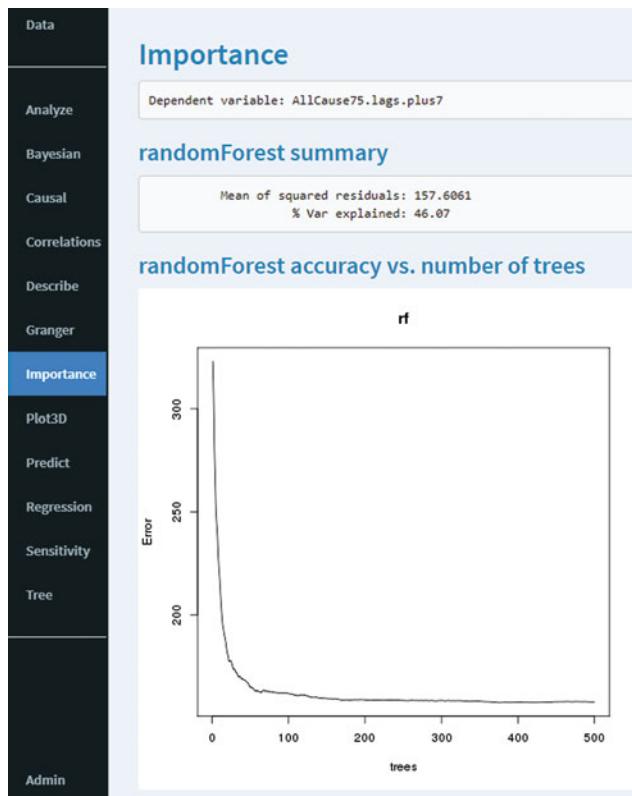


Fig. 2.19 Outputs from the *randomForest* analysis: Percent of variance explained and accuracy vs. number of trees in the ensemble. “rf” is an abbreviation for “random forest”

We will use the Causal Analytics Toolkit (CAT) software (<http://cox-associates.com/CloudCAT>), which runs the *randomForest* R package and presents its results via an internet browser (or, optionally, an Excel add-in). By default, the package averages predictions for the dependent variable over an ensemble of 500 trees grown on different samples from the full data set.

The CAT software provides two commands that generate and display outputs from the *randomForest* package: “Importance” and “Sensitivity.” Each generates multiple outputs. Figures 2.19 and 2.20 show key outputs from the “Importance” command. The plot in Fig. 2.19 shows how the mean squared prediction error (MSE) for the dependent variable *AllCause75.lags.plus7* decreases as the *randomForest* algorithm averages predictions over increasing numbers of trees, falling from more than 300 for a single tree to about 160 when 300 or more trees are included in the ensemble. Conditioning on all other variables allows about 46% of the variance in the dependent variable, daily elderly mortality counts, to be explained, as indicated by the “% Var explained: 46.07” message in the “randomForest summary” section. (In the CAT software, clicking on any such section heading calls up corresponding

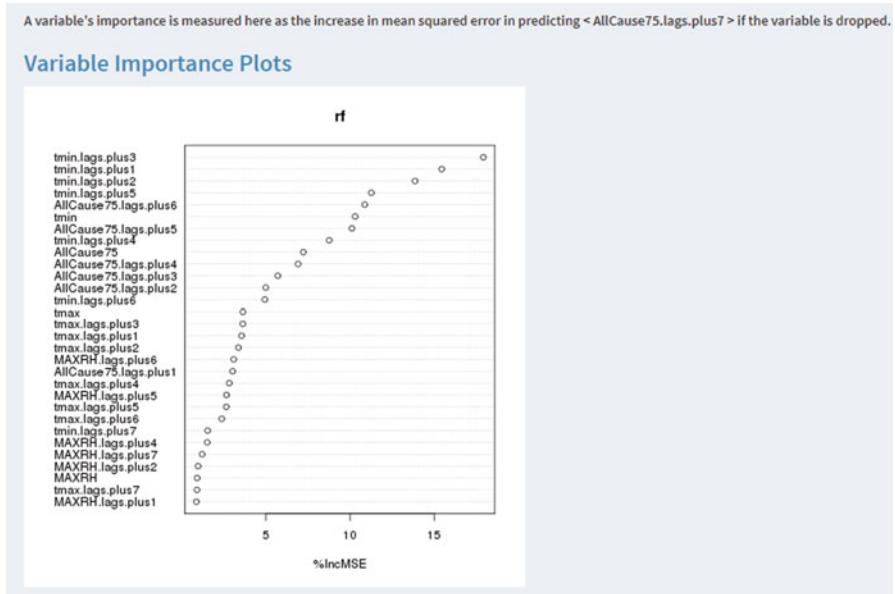


Fig. 2.20 Importance plot from the *randomForest* analysis

documentation, explaining the outputs and how they are calculated in more detail. Since *randomForest* averages over many random samples of data, its outputs vary slightly between repeated runs.)

Figure 2.20 shows a second output from the “Importance” command: estimates of the percentage increases in mean squared prediction error made by the rf ensemble (denoted by “%IncMSE” on the x axis) if each predictor is excluded. This is at best only a rough heuristic for gauging the importance of predictors, as it considers them one at a time and ignores interactions among them. Nonetheless, it can reveal such insights as that all of the top few predictors are lagged values of minimum daily temperature (*tmin*), suggesting that daily elderly mortality rate depends primarily on cold temperatures in recent days. (Only the top 30 predictors are shown; a separate table provides a complete listing of all the variables and their estimated importances.)

The “Sensitivity” command in CAT uses the *randomForest* package to generate a *partial dependence plot* (PDP). This is illustrated in Fig. 2.21 for the partial dependence of elderly mortality in a week (*AllCause75.lags.plus.7*) on low temperature today (*tmin*). Roughly speaking, a PDP shows how the predicted value of the dependent variable (on the y axis) depends on one specific explanatory variable (on the x axis), given the current values of the other variables. It is generated by using the random forest ensemble to predict the average value of the dependent variable as one predictor (here, *tmin*) is varied over its entire range of values in the data set while holding all other variables fixed at their actual values in the data set. In more detail, the process is as follows. First, a random forest ensemble of many (e.g.,

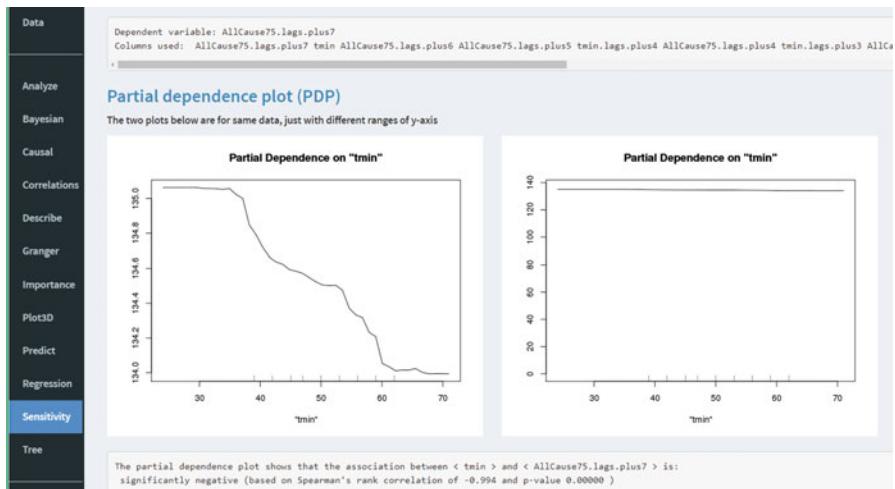


Fig. 2.21 A partial dependence plot (PDP) for elderly mortality count in 7 days from now (*AllCause75.lags.plus.7*) vs. low temperature today (*tmin*)

500) classification trees is created using R’s *randomForest* package. Next, the data set is modified by replacing all of the values in one column—the one for the selected explanatory variable, which is *tmin* in Fig. 2.21—with a constant value, the same for all records (rows) in the data table. For the first iteration, this constant value is the smallest value of the explanatory value in the data set. It is increased in subsequent iterations to the next smallest value, and so on incrementally until it reaches the maximum value of the selected explanatory variable. At each iteration, each modified record in the data set (meaning the original row modified to replace the entry in the column for the selected explanatory variable with the constant value for that iteration) is dropped through each tree in the ensemble, i.e., each tree is used to predict a value of the dependent variable for each modified record. The predicted values of the dependent variable are then averaged over all records for all trees. The result is a single (averaged) predicted value for the dependent variable for the specified constant value of the selected explanatory variable. This constitutes one point on the PDP. As the constant value is incrementally increased from the smallest to the largest value of the explanatory variable, the entire PDP curve is generated.

In CAT, clicking on the “Sensitivity” command uses automatically generates a PDP. The default is to treat the first column in the data table as the dependent variable and the second column as the selected explanatory variable that will go on the x axis, but the user can select any two variables to fill these roles. (Python versions of random forest also allow two explanatory variables to be selected and generate PDP surfaces for the dependent variable as a function of the two selected explanatory variables, but this is not yet supported in CAT.) As shown in Fig. 2.21, CAT generates a curve for the dependent variable as a function of the selected explanatory variable and plots it twice using two different vertical axes: one on the left that emphasizes the changes in the dependent variable over its range of values,

and one on the right that shows the same PDP with a vertical axis that starts at zero. The CAT software (but not the *randomForest* package) also runs a Spearman's rank correlation test to determine whether there is a significant ordinal association between the explanatory variable and the dependent variable in the PDP.

The mechanics of generating a PDP have now been explained. But what is its intended interpretation—what does a PDP mean, and why is it useful? The PDP for a dependent variable Y vs. an explanatory variable X is *not* simply a depiction of the conditional mean value of Y for different assumed values of X predicted by the non-parametric model ensemble of classification trees in the random forest. Rather, it is this with the additional constraint that other variables are held fixed at their current values as Y is predicted for different assumed values of X . In the special case where X is a direct cause (i.e., a parent) of Y in a DAG model, the PDP for Y vs. X is a non-parametric estimate of the *natural direct effect* of X on Y holding the values of other variables fixed. Whether this coincides with the manipulative effect on Y of setting X to different values while holding other variables fixed depends on how well the estimated PDP estimated from the observed joint values of variables among the cases in the data set describes the conditional mean value of Y for different X values that would be calculated via graph surgery (representing direct manipulation of X values) from the CPTs in a valid causal BN model of the same data.

Causal Concentration-Response Curves, Adjustment Sets, and Partial Dependence Plots for Total and Direct Effects in Causal Graphs

One of the most commonly used displays in air pollution health effects research and epidemiology is the exposure *concentration-response curve*. This plots the risk (or incidence rate in a population, in units such as expected cases per person-year) on the y axis against a summary of exposure concentration on the x axis. The exact intended interpretation of such curves is seldom specified. Authors interpret them variously as offering descriptive summaries of observed or expected exposure concentration and response data points (the regression interpretation) and as indicating how responses would change if exposure concentrations were changed (a manipulative causal interpretation), usually without specifying why the concentrations might change or distinguishing among natural direct, controlled direct, total, or other effects. Yet, the same curve cannot represent all of these distinct concepts. To more clearly define what a causal concentration-response curve might mean, consider the example DAG model in Fig. 2.22. This DAG model was drawn using the on-line version of DAGagitty (www.dagitty.net/dags.html), a software package for creating and analyzing the statistical implications of DAG models. The arrows between its nodes are intended to signify that the probability distribution for a health outcome in members of a population (e.g., age at death, lung cancer by age 70, etc.) depends directly on income (which may serve as a proxy for health care and other

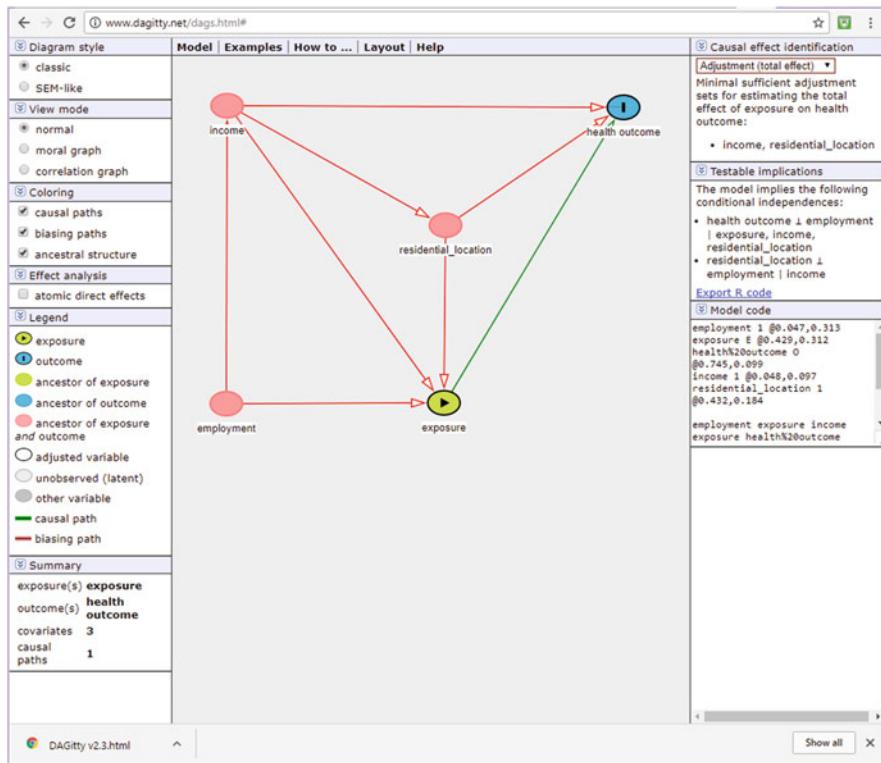


Fig. 2.22 DAGitty diagram for the effect of exposure on health outcome. *Source:* Drawn in browser using DAGitty software (www.dagitty.net/dags.html)

variables), residential location, and exposure. Income, in turn, depends directly on employment (or occupation), and exposure depends both directly on employment and also indirectly on employment via income, as well as depending directly on residential location, which in turn depends on income. A full causal BN model would quantify these dependencies with CPTs for each node (and a marginal distribution for the employment input node). The DAG model implies that the joint distribution of the variables can be factored as follows:

$$P(\text{employment}) * P(\text{income} \mid \text{employment}) * P(\text{residential_location} \mid \text{income}) * P(\text{exposure} \mid \text{employment}, \text{income}, \text{residential_location}) * P(\text{health outcome} \mid \text{income}, \text{residential_location}, \text{exposure}).$$

Given any set of values for the five variables, the joint probability that they will have those values can be calculated by plugging appropriate numbers from the node-specific probability tables into this formula. But even without specifying the quantitative probability tables for the model, the DAG structure alone allows useful

inferences to be drawn about the data analyses needed to quantify total effects and direct effects of one variable on another—in this case, the effect of exposure on health outcome.

DAGitty lets the user select one main cause variable and one effect variable. It then applies well-developed graph-theoretic criteria and algorithms to automatically calculate *adjustment sets* showing what other variables to condition on in order to calculate the total and direct causal effects of the selected cause variable on the selected effect variable (Textor et al. 2016). These are listed in the “Causal effect identification” area at the upper right of the DAGitty screen. In Fig. 2.22, DAGitty lists a single “minimal sufficient adjustment set for estimating the total effect of exposure on health outcome.” It consists of the two confounders *income* and *residential_location*. Both of these must be conditioned in to obtain estimates of the unconfounded causal effect of exposure on health outcome. If all relationships were known to be linear, then a multiple linear regression model for the causal effect of *exposure* on *health outcome* would have to include *income* and *residential_location* on its right-hand side, as well as *exposure*, in order for the resulting regression coefficient for *exposure* to be interpretable as indicating the predicted increase in *health outcome* per unit increase in *exposure*.

Below the adjustment sets, DAGitty lists testable implications of the DAG structure, using the common notation “ $Y \perp X | Z$ ” to denote “*Y* is conditionally independent of *X* given *Z*.” Two testable implication of the DAG structure are shown: that *health outcome* is conditionally independent of *employment* given the values of *exposure*, *income*, and *residential_location*; and that *residential_location* is conditionally independent of *employment* given *income*. The full theory and resulting algorithms for automatically generating all such testable implications and the minimal sufficient adjustment sets for estimating total and direct causal effects are discussed in Textor et al. (2016) and its references. A key principle is to condition on confounders, or common causes, of the exposure and response variables to get unconfounded estimates of total or direct causal effects of exposure on response probabilities; but not to condition on common children (“colliders”) or their descendants, in order to avoid creating Berkson selection biases. A graph-theoretic criterion called “d-separation” that accounts for blocking of information transfer along paths by appropriate conditioning provides the generalization needed to compute minimal sufficient adjustment sets and unbiased estimates of direct and total causal effects, as well as testable conditional independence implications of a DAG model.

To generate partial dependence plots (PDPs) estimating natural direct and total causal effects of one variable on another, the CAT software uses an R package implementation of DAGitty to automatically determine which other variables to condition on, i.e., to compute minimal sufficient adjustment sets. It then conditions on these variables in the random forest ensembles used to compute the PDPs. We illustrate this process for one of the data sets (“asthma”) bundled with CAT. Readers who wish to follow along in detail can browse to the <http://cox-associates.com/CloudCAT> site and load the data set from the “Sample: Data” drop-down menu near the top of the CAT Data screen. This is a random sample of a large data set for adults 50 years old or older from 2008 to 2012 for whom survey data are available from the

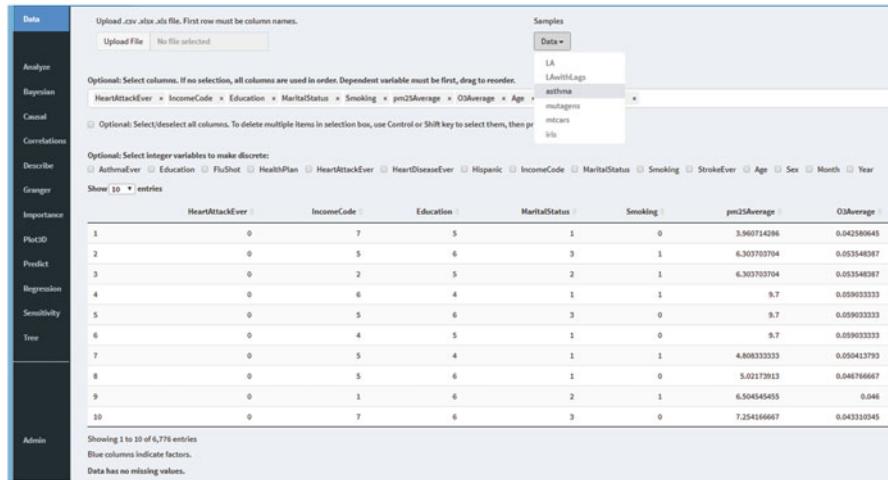


Fig. 2.23 The “asthma” example data set used to illustrate BN learning

Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS), along with county-level average annual ambient concentrations of ozone (O₃) and fine particulate matter (PM2.5) levels recorded by the U.S. Environmental Protection Agency. Figure 2.23 shows the first few records of this data set, with several columns selected for further analysis. More detailed description of the data set and the variables is provided by Cox (2017b). For purposes of illustrating causal analysis using DAGitty and PDPs, all that is needed is the data set and the names of the variables (columns) included in the analysis.

The first step in the analysis is to obtain a BN DAG model of the data. CAT uses the *bnlearn* package in R for this purpose; it is activated by selecting the “Bayesian” command from the command menu. Details of this step, including optional incorporation of knowledge-based constraints such as that heart disease is a possible effect but not a possible cause of exposures, are discussed in the following section on causal discovery algorithms. By default, the “Bayesian” command applies a hill-climbing algorithm (“hc”) that searches for a BN model maximizing a score assessing how well the model explains the data. This algorithm, together with some knowledge-based constraints such as that *month* and *year* can be causes but not effects of other variables, produces the BN DAG in Fig. 2.24.

Once a DAG model has been obtained from the data, possibly with the guidance of knowledge-based constraints on allowed (or required) arrows, the next step is for the user to select an *Exposure* variable (also called a source or cause) and a *target* variable (also called a response or effect) and to specify whether the direct or the total effect of the former on the latter is to be quantified. In CAT, this is done with the help of an interactive BN DAG editor (Fig. 2.25) that redraws the output from the *bnlearn* package (Fig. 2.24) and allows it to be edited. This interactive graph editor allows the user to reposition the nodes, select a source (green node) and a target (pink node) by clicking on them, and specify or revise any knowledge-based constraints about

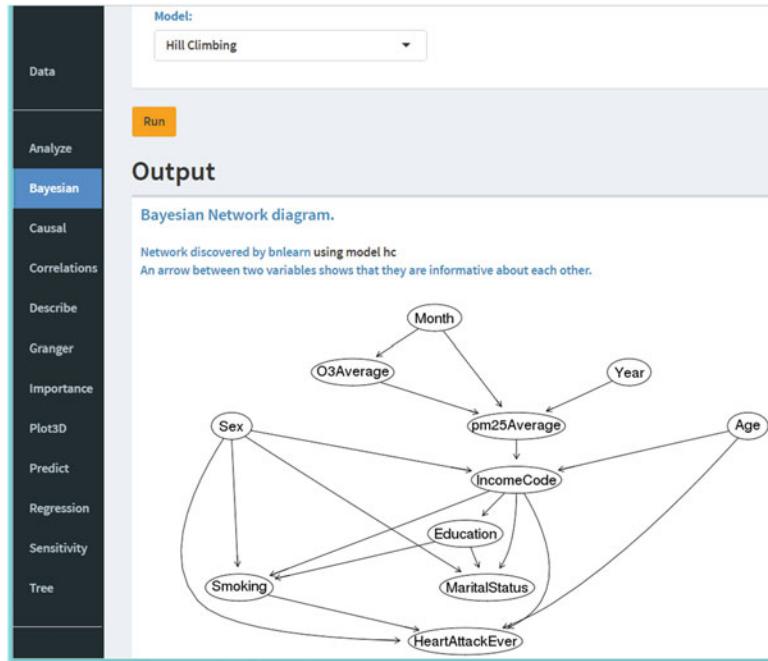


Fig. 2.24 A BN DAG generated for the data set in Fig. 2.23 by the BN learning program

allowed and forbidden arrows to be assumed in creating and using the DAG model. Figure 2.25b shows a close-up of the network diagram in Fig. 2.25a, with *Age* selected as the exposure (cause) and *HeartAttackEver* selected as the target (effect).

Next, the user must specify whether to quantify the direct effect or the total effect of the exposure on the target. This selection is made using the Direct or Total radio buttons near the top of the screen in Fig. 2.26. The CAT software then runs the DAGitty package and lists adjustment sets of variables to condition on to control for potential biases in estimating the selected causal effect (bottom of Fig. 2.25a). In Fig. 2.26, there is only one adjustment set, consisting of *IncomeCode* and *Sex*. CAT completes the analysis by generating a partial dependence plot (PDP) of the target variable vs. the source variable. The PDP conditions on the variables in the selected adjustment set (via the *randomForest* algorithm in R) to control for confounding and selection biases without introducing new ones. Figure 2.26 shows the resulting PDP for *HeartAttackEver* vs. *Age*, controlling for *IncomeCode* and *Sex* by conditioning on them in the trees on the random forest ensemble. The left side has a vertical scale set by the variations in the target variable and the right side shows the same PDP on a vertical scale that includes 0. CAT also tests whether there is a significant ordinal association between the source and target variables using Spearman's rank correlation, which reflects how likely it is that higher values of the effect variable occur for higher values of the exposure variable. As shown at the bottom of Fig. 2.26, there is a highly statistically significant ordinal association between age and self-reported heart disease, as one would expect.

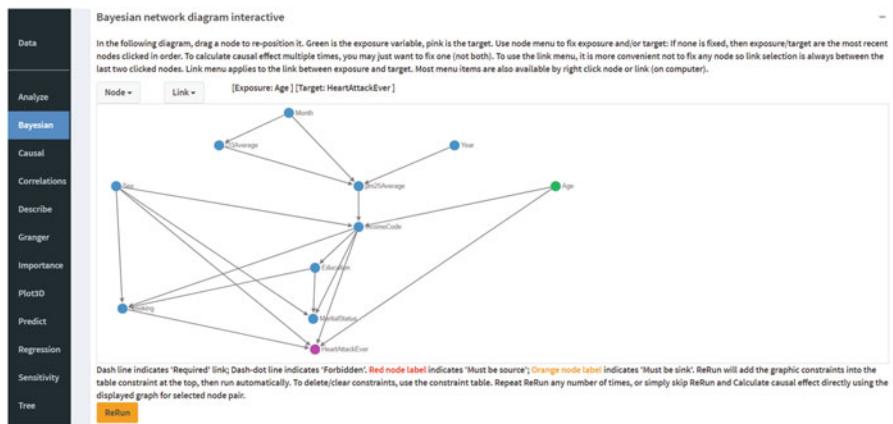
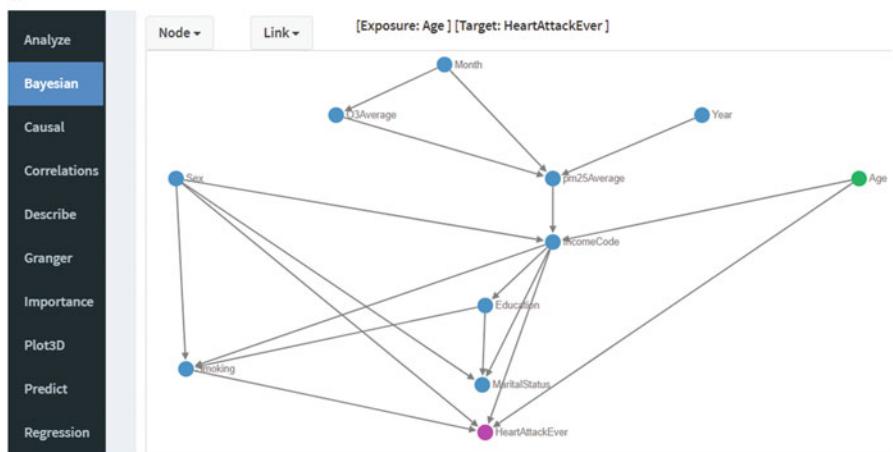
a**b**

Fig. 2.25 (a) The BN from Fig. 2.24 redrawn by CAT to allow interactive editing and selection of variables for causal analysis. (b) Close-up of the network in a. *Age* has been selected as the cause or exposure variable interest and *HeartAttackEver* as the effect or target variable of interest for further analysis

Including Knowledge-Based Constraints and Multiple Adjustment Sets

We have now illustrated the mechanics of estimating the direct or total effect of one variable on another in a DAG model. However, one might (and should) question whether the algorithms used have produced sensible results. Some of the arrow directions in Fig. 2.25b seem counter-intuitive for causal interpretation. The arrow from *PM2.5Average* to *IncomeCode* might possibly be rationalized by considering that locations with dirty air are less likely to attract high-income residents, but the

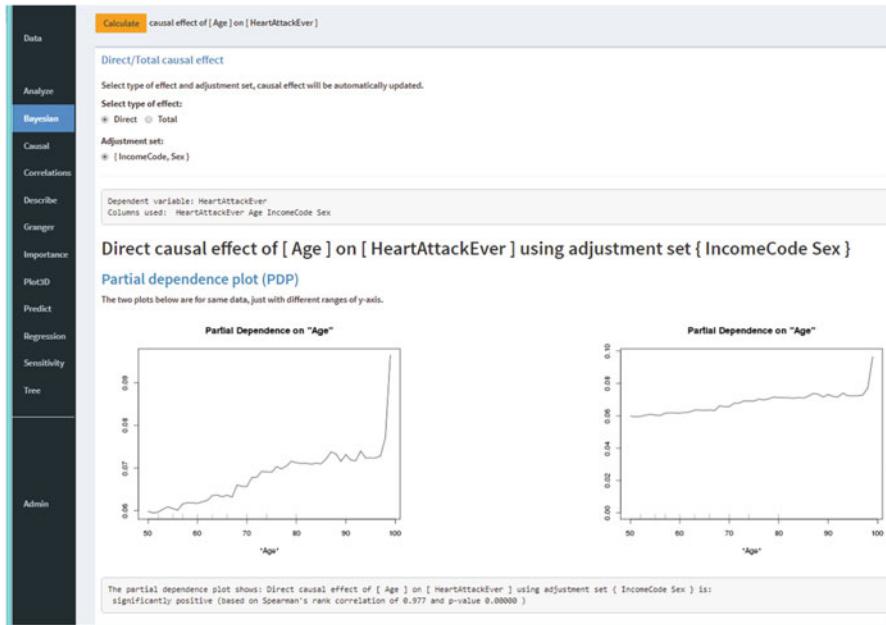


Fig. 2.26 Partial dependence plot for the direct causal effect of *Age* on *HeartAttackEver*, adjusting for *IncomeCode* and *Sex*

arrow from *IncomeCode* to *Education* seems clearly wrong, insofar as educational attainment usually precedes current income. One option is to leave the DAG model in Fig. 2.25b as-is, bearing in mind that the DAG simply represents a possible way of factoring the joint distribution of the variables in terms of marginal distributions for its inputs (*Sex*, *Month*, *Year*, and *Age*) and conditional distributions (CPTs) for its endogenous variables. In this case, the only causal implications of the DAG are that direct causes should be linked by arrows (in either direction) to their effects. But if we want to constrain the arrow directions to better correspond to possible manipulative causation, then the CAT software can be used to enter such constraints. Figure 2.27 shows how to forbid an arrow from *IncomeCode* to *Education* by entering its start and end points on a list of forbidden arrow. Constraints requiring an arrow or specifying a variable as a source (input node with only outward-directed arrow allowed) or as a sink (output node with only inward-directed arrows allowed) can be entered into CAT if desired. With this constraint, clicking on the “Run” button at the bottom of the screen in Fig. 2.27 produces the revised DAG in Fig. 2.28. In the DAG model of Fig. 2.28, the arrow directions seem generally consistent with a possible manipulative causal interpretation.

The use of knowledge-based constraints allows the user to inform the DAG-learning algorithms about such important facts as that month might cause exposure but exposure does not cause month. It is not necessary to impose such constraints to discover which variables are informative about each other, but

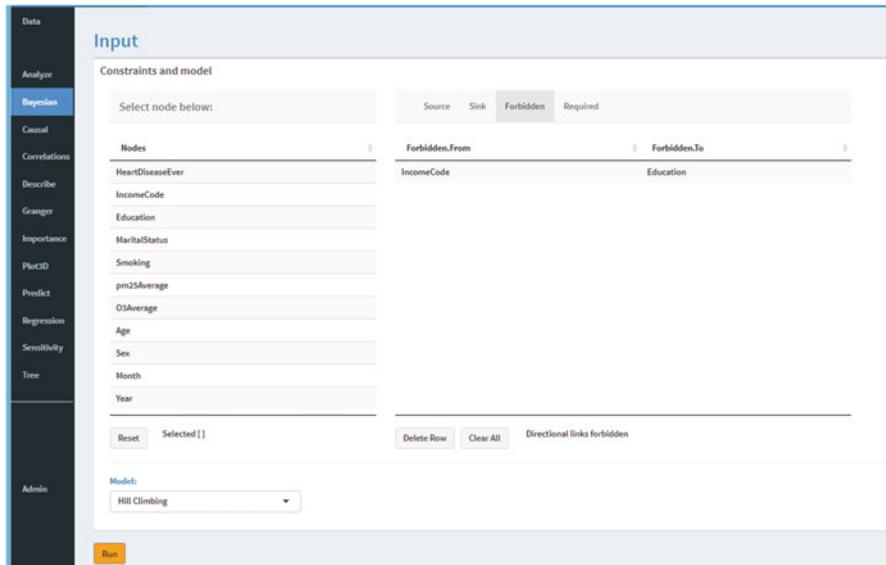


Fig. 2.27 Entering a constraint forbidding an arrow from *IncomeCode* to *Education*

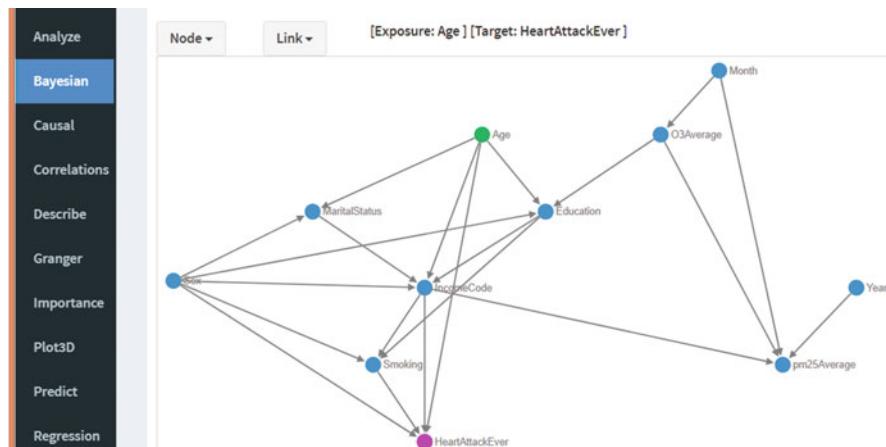


Fig. 2.28 Revised DAG model with knowledge-based constraints from Fig. 2.27

including them allows arrows to be oriented to be more consistent with manipulative-causal interpretations. In some cases, however, the directions of arrows are ambiguous even in principle. For example, if people with higher incomes tend to reside in neighborhoods with less pollution, should an arrow be directed from income to pollution (suggesting that higher incomes allow one to purchase homes in

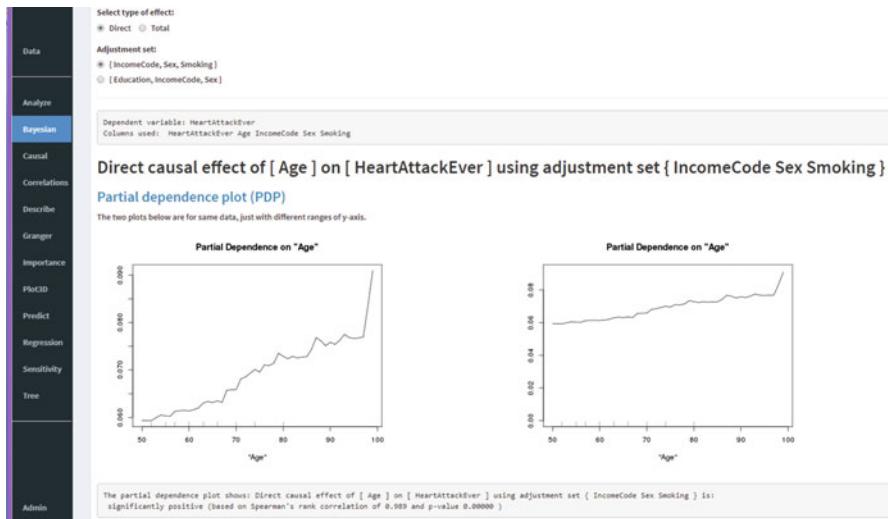


Fig. 2.29 Partial dependence plot for the direct causal effect of *Age* on *HeartAttackEver*, adjusting for *IncomeCode*, *Sex*, and *Smoking*

less polluted neighborhoods) or from pollution to income (suggesting that high pollution drives away high-income residents)? Either resulting DAG provides a valid representation of the same joint distribution of the variables. Some causal graph modeling programs do not require that all arcs between pairs of variables be directed, but DAG models are the most common causal graph modeling framework. We will use the DAG model in Fig. 2.28 with the understanding that some of its arrows could be redirected without greatly changing the results or their interpretation.

In many applications, more than one adjustment set is identified from a DAG via the d-separation and DAG-computation algorithms in DAGitty. Each of them can be used to estimate the effect of the exposure variable on the target variable. Figures 2.29 and 2.30 present an example. The DAG model in Fig. 2.28 has two different adjustment sets that can be conditioned on to obtain unbiased estimates of the direct causal effect of age on self-reported cumulative heart attack risk: $\{IncomeCode, Sex, Smoking\}$ and $\{Education, IncomeCode, Sex\}$. Figures 2.29 and 2.30 show the respective partial dependence plots for heart attack risk vs. age estimated for these two adjustment sets. The two PDPs are similar but not identical, as might be expected due to the random variability in random forest model ensembles.

Once a DAG model has been created for a data set, it can be re-used to quantify direct or total causal relationships between any pair of variables. With CAT's interactive DAG network editor, this is done simply by clicking on a new exposure-target pair and then clicking on the “Calculate” button (top left of Fig. 2.31). Figure 2.31 shows the estimated direct causal effect (PDP) of

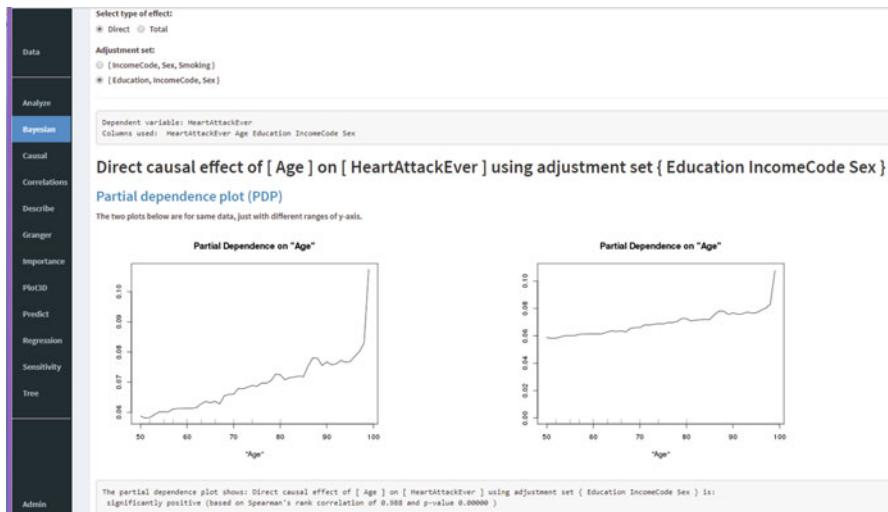


Fig. 2.30 Partial dependence plot for the direct causal effect of *Age* on *HeartAttackEver*, adjusting for *Education*, *IncomeCode*, and *Sex*

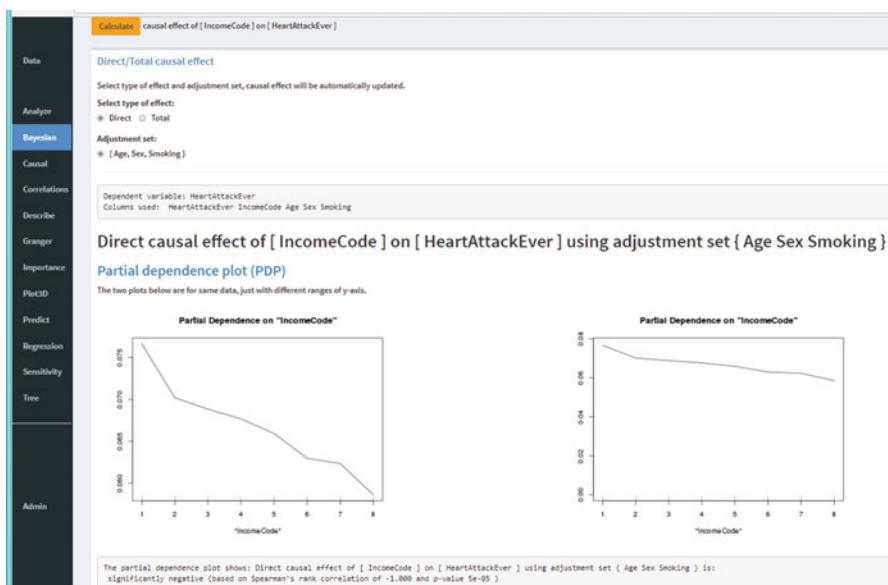


Fig. 2.31 Partial dependence plot (PDP) for the direct causal effect of *IncomeCode* on *HeartAttackEver*, adjusting for *Age*, *Sex*, and *Smoking*

IncomeCode on self-reported cumulative heart attack risk, *HeartAttackEver*, obtained by selecting (clicking on) these variables in order and then clicking on the “Calculate” button. Other causal effects can be estimated equally easily. In this data set, as in many other real-world data sets, health risks decrease significantly at higher income levels.

The *RandomForest* package in R and the PDPs that it generates do not assume a parametric statistical model and therefore do not provide confidence intervals for model parameters and estimates (as some Bayesian extensions of CART methods do that assume parametric prior distributions for tree depth and branching factors). However, CAT provides other ways to test for model validity and to characterize uncertainties in BN results and causal PDP estimates. One option is to run the DAGitty package from CAT. Scrolling down the “Bayesian” screen below the PDP output shown in Fig. 2.31 brings up the portion of this screen shown in Fig. 2.32, and clicking the “Run package dagitty” button then generates testable

Run package dagitty

Results from package dagitty

```
List testable implications of a structural equation model:
Age ||_ Month
Age ||_ O3Average
Age ||_ Sex
Age ||_ Smoking | Education, IncomeCode, Sex
Age ||_ Year
Age ||_ pm25Average | IncomeCode, O3Average
Education ||_ HeartDiseaseEver | Age, IncomeCode, Sex, Smoking
Education ||_ MaritalStatus | Age, Sex
Education ||_ Month | O3Average
Education ||_ Year
Education ||_ pm25Average | IncomeCode, O3Average
HeartDiseaseEver ||_ MaritalStatus | Age, Education, IncomeCode, Sex
HeartDiseaseEver ||_ MaritalStatus | Age, IncomeCode, Sex, Smoking
HeartDiseaseEver ||_ Month | O3Average
HeartDiseaseEver ||_ Month | Age, Education, Sex
HeartDiseaseEver ||_ Month | Age, IncomeCode, Sex, Smoking
HeartDiseaseEver ||_ Year
HeartDiseaseEver ||_ pm25Average | IncomeCode, O3Average
HeartDiseaseEver ||_ pm25Average | Age, Education, IncomeCode, Sex
HeartDiseaseEver ||_ pm25Average | Age, IncomeCode, Sex, Smoking
IncomeCode ||_ Month | O3Average
IncomeCode ||_ Month | Age, Education, Sex
IncomeCode ||_ O3Average | Age, Education, Sex
IncomeCode ||_ Year
MaritalStatus ||_ Month
MaritalStatus ||_ O3Average
MaritalStatus ||_ Smoking | Education, IncomeCode, Sex
MaritalStatus ||_ Year
MaritalStatus ||_ pm25Average | IncomeCode, O3Average
MaritalStatus ||_ pm25Average | Age, Education, IncomeCode, Sex
Month ||_ Sex
Month ||_ Smoking | Education, IncomeCode, Sex
Month ||_ Smoking | Age, Education, Sex
Month ||_ Smoking | O3Average
Month ||_ Year
O3Average ||_ Sex
O3Average ||_ Smoking | Education, IncomeCode, Sex
O3Average ||_ Smoking | Age, Education, Sex
O3Average ||_ Year
Sex ||_ Year
Sex ||_ pm25Average | IncomeCode, O3Average
Smoking ||_ Year
Smoking ||_ pm25Average | IncomeCode, O3Average
Smoking ||_ pm25Average | Education, IncomeCode, Sex
```

Fig. 2.32 Part of the output from the DAGitty package listing testable implications of the DAG model in Fig. 2.28

conditional independence implications (the beginning of the listing of which is shown in Fig. 2.32), followed by a listing of path coefficients and total effects identifiable by regression (with their corresponding adjustment sets) and path coefficients identifiable by instrumental variables, along with corresponding variables to condition on. Testing the testable implications of the DAG model listed by DAGitty, e.g., by using CART trees to check whether the implied conditional independence relations hold in a data set, provide one way to assess its consistency with data.

Figure 2.33 shows another way to evaluate a DAG model: use several different BN learning algorithms to estimate the causal BN DAG from the data and see how well they agree. In CAT, clicking on the “Compare” button (upper left of Fig. 2.33) runs four different BN algorithms (hc = hill climbing, tabu = tabu search, gs = grow-shrink Markov blanket, iamb = incremental association Markov blanket); these algorithms are documented in the *bnlearn* package documentation (<https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf>). The number of these algorithms that agree on each arrow is then tabulated (bottom of Fig. 2.33, table truncated after the first five rows for legibility) and used to generate a composite diagram (top of Fig. 2.33) in which thicker arrows indicate support from more of these DAG-learning algorithms. Checking the “Use short labels” option in the shaded control box to the left of the network causes abbreviated variable names to be used, and this has been done in Fig. 2.33 to increase legibility. A slider (“Min. weight to display”) can be used to display only those arrows that have support from at least a certain number of the DAG-learning algorithms. If this slider is set to show arrows on which all four algorithms agree, then the variables separate into two clusters: *Year*, *Month*, *PM2.5*, and *O3* on the left, and all other variables on the right. *Age*, *Sex*, *Smoking*, and *Income* are identified by all four algorithms as the parents of

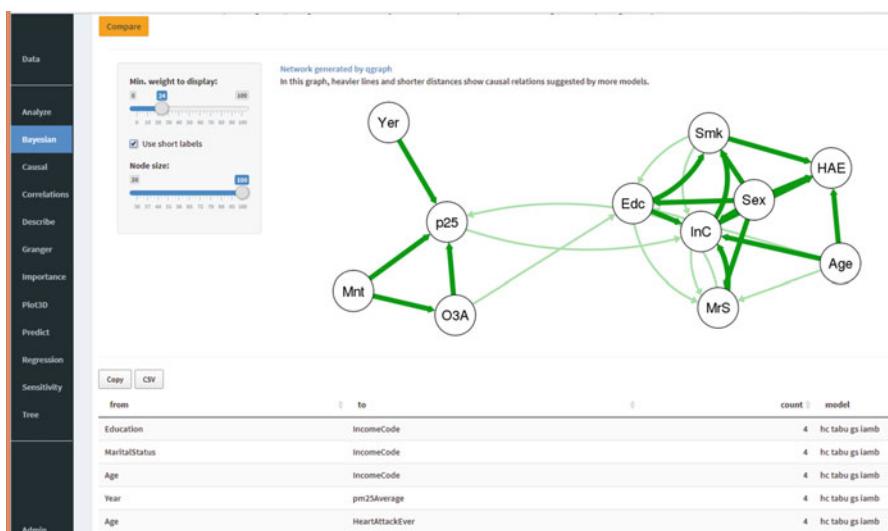


Fig. 2.33 Visualizing consonance of findings from different causal discovery algorithms

heart attack risk, shown as the node *HAE* at the right of the network in Fig. 2.33. (Algorithms gs and iamb can produce undirected arcs when the data do not allow a unique direction to be inferred. In Fig. 2.33, an undirected arc is counted as one arrow in each direction.)

Power Calculations for Causal Graphs

Absence of an arrow between two nodes in a DAG model learned by causal discovery algorithms from available data does not necessarily imply that they are not informative about each other or that one does not cause the other. There might be a dependency between them that is simply too small to be detected. To assess this possibility, it is useful to create simulated data sets from the original one, in which it is assumed that one variable *does* depend on another, with the dependency described by a simple formula such as that each 1% deviation from the mean of one variable causes a corresponding *b*% deviation from the mean of the other variable. Here, *b* reflects the elasticity of the second variable with respect to changes in the first. In many real-world data sets, increasing *b* gradually reveals a fairly sharp, threshold-like transition from values of *b* that are too low to yield arrows between the two variables in the DAG, given the random variations of the variables, and values of *b* that are high enough so that most or all of the DAG-learning algorithms will detect the dependency and create an arrow between them. Such sensitivity analyses reveal the sizes of effects that are too small to detect and those that are large enough to be detected with near certainty.

Figure 2.34 illustrates an alternative way to visually check the validity of missing arrows in a DAG model. If no arrow connects two variables, such as *PM2.5* and *AllCause75* in the DAG in Fig. 2.24, then the partial dependence plot showing how

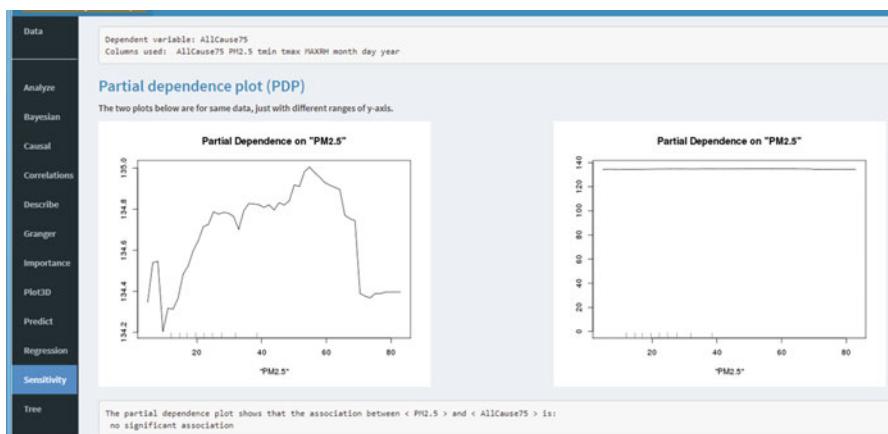


Fig. 2.34 Partial dependence plots for missing arrows (e.g., from *PM2.5* to *AllCause75*) should be approximately horizontal, indicating absence of dependency

one depends on the other (after conditioning on the values of other variables, yielding a natural direct effect estimate) should be approximately horizontal, as shown in Fig. 2.34. In this PDP, *AllCause75* remains flat at about 134.6 ± 0.4 as PM2.5 ranges from about 5 to over $80 \mu\text{g}/\text{m}^3$. The expanded vertical scale on the left shows some noise, or random variation around this level, as should be expected due to the random sampling components of the random forest algorithm used in generating the PDP, but overall the relation is close to flat.

Predictive Analytics for Binary Outcomes: Classification and Pattern Recognition

The classification tree, random forest ensemble, partial dependence plot, and BN-learning and modeling techniques just surveyed are all used as parts of contemporary *predictive analytics*. That direct causes are informative about, and help to predict, their effects (and more specifically, in DBN models, the future of their effects) forges a strong link between predictive and causal analytics, but predictive analytics methods also have many other useful applications. A standard predictive analytics challenge is to learn how to predict the values in one column of a spreadsheet or data table—in other words, the values of some target variable—from the values in other columns. This can be accomplished by partitioning the data into disjoint training and test subsets of records (rows); eliminating redundant columns (duplicates or highly correlated columns) and useless columns (those with the same in all or nearly all rows) training and tuning each of a suite of machine learning algorithms (e.g., regression models, CART, random forest, neural networks, support vector machines, clustering, and other algorithms) on the training set; and evaluating and comparing their performance on the test set. The entire process can be automated using software such as the *caret* (Classification and Regression Training) package in R (<https://topepo.github.io/caret/>; Kuhn 2008). In CAT, selecting a target column to predict (the first column is the default) and a set of columns to use as predictors (the default is all of the other columns) and then clicking on the “Predict” command invokes the automated predictive analytics capabilities of the *caret* package.

We will illustrate the process for the applied problem of learning to predict whether a chemical is a mutagen. The data set used for this example is the “mutagens” data set that comes bundled with CAT; it is excerpted from a quantitative structure-activity relation (QSAR) data set for chemical mutagenicity provided as part of an R package of QSAR data (<https://cran.r-project.org/web/packages/QSARdata/QSARdata.pdf>). Figure 2.35 shows the first ten records in the data set. Each row corresponds to a chemical. Each column is an attribute of the chemical such as its molecular weight, number of halogen atoms, and so forth. The dependent variable, *mutagen*, is in the first column: it has possible values of 1 = mutagen and 0 = not a mutagen, as measured in an Ames *Salmonella* test for mutagenicity.

	mutagen	MW	AMW	Sv	Se	Sp	Ss	Mv	Me	Mp	Ms	nAT	nSK	nBT	nBO	nBM
1	1	326.42	7.59	29.28	42.6	30.56	50.67	0.68	0.99	0.71	2.03	43	25	46	28	23
2	1	174.17	9.17	13.21	19.63	13.44	38	0.7	1.03	0.71	2.92	19	13	19	13	10
3	0	300.33	9.39	20.01	33.65	20.99	61.17	0.63	1.05	0.66	3.06	32	20	34	22	9
4	0	143.21	6.23	12.6	23.06	13.48	26.17	0.55	1	0.59	2.62	23	10	22	9	2
5	0	215.88	17.99	10.6	12.95	11.72	27.11	0.88	1.08	0.98	2.71	12	10	12	10	6
6	1	190.22	7.93	15.4	24.39	15.97	36	0.64	1.02	0.67	2.57	24	14	25	15	11
7	1	328.18	12.62	18.79	27.08	20.04	49.42	0.72	1.04	0.77	2.75	26	18	27	19	13
8	0	324.35	8.11	26.34	40.71	27.36	59.17	0.66	1.02	0.68	2.47	40	24	42	26	15
9	1	136.17	7.56	11.29	18.18	11.75	25.67	0.63	1.01	0.65	2.57	18	10	18	10	7
10	1	323.37	7.89	26.82	41.48	27.91	54.92	0.65	1.01	0.68	2.29	41	24	45	28	13

Fig. 2.35 Mutagen data set

The predictors consist of the large number of chemical properties summarized in subsequent columns. The prediction task is to use a subset of the records to learn prediction rules for classifying chemicals as mutagens or not mutagens based on their chemical properties (attributes). This is typical of a wide class of practical problems in which the goal is to learn to classify cases or patterns accurately based on a training set in such a way that the learned classification rules (i.e., class prediction rules) can be applied successfully to new cases not in the training set.

CAT provides two options for selecting a subset of records to train on: random sampling from the set of all records, or selection of the top part of the data table as the training set, with the rest of the data as the test set. In either case, the user must specify how much of the data set (e.g., 50%, 75%, or some other fraction) to use for training. If the rows correspond to consecutive observations in a time series, then training on the earlier rows and testing them on later rows may correspond to the natural order in which data become available. Figure 2.36 shows the options for creating disjoint test sets and, if desired, for filtering columns, i.e., reducing the number of candidate predictors by using CART trees, linear regression coefficients, or importances in a random forest analysis to select the most promising-looking predictors. Options to automatically detect and drop redundant (highly correlated) columns and to pre-process predictors, e.g., by standardizing them, are included on the right side of the Prediction Options dialogue. The bottom of the screen summarizes the user's choices under "Prediction output." These options are detailed further

The screenshot shows the CAT software interface. On the left is a vertical menu bar with the following items: Data, Analyze, Bayesian, Causal, Correlations, Describe, Granger, Importance, Plot3D, Predict (which is highlighted in blue), Regression, Sensitivity, and Tree. The main area is titled "Prediction Options". It contains a "Train data percentage" input field set to 50, a checked checkbox for "Use top rows as train data only", a "Select filters:" section with checkboxes for Tree, Linear Regression, and Importance (the last one is checked), and a "Pre-process data:" section with checkboxes for Remove highly correlated and Preprocess predictors (both are checked). Below these is a yellow "Run" button. To the right of the Run button is a large section titled "Prediction output" containing the following text:

```

Dependent variable: mutagen
Train data percentage = 50
Use top rows as train data only: TRUE
Filters: importance
Remove highly correlated
Preprocess predictors

Using top 349 rows as training data. Total number of samples is 699

```

Fig. 2.36 Prediction options to be supplied by the user

in the *caret* documentation, which is accessible from CAT via hyperlinks. Once they have been selected, clicking on the Run button causes the rest of the predictive analytics process to run and generate output reports.

Figure 2.37 shows one of the detailed output reports: the observed values (1 = mutagen, 0 = not mutagen) for chemicals in the test set (indicated by InTrain = 0) and the values predicted by each of several machine-learning algorithms (earth, which is the R implementation of multiple adaptive regression splines; rpart and ctree, which are two recursive partitioning CART-type algorithms; random forest (rf); gradient boosted machine (gbm); and boosted generalized linear model (glmboost)). The *caret* documentation and its references provide details of these machine-learning algorithms and many others; for our purposes, however, it suffices to treat them as black-box algorithms for learning predictive rules from training data. The output of each algorithm is the conditional probability that a chemical is a mutagen in the Ames *Salmonella* test, given the values of its other attributes; these probabilities are rounded to the nearer of 0 or 1 in Fig. 2.37. Figure 2.37 shows that there are some chemicals that are easy to classify, in the sense that all of the algorithms correctly predict the value of *mutagen* for them. The chemical in the first row in the table in Fig. 2.37, with SampleID number 525, is an example: all of the predictive algorithms correctly predict the observed value, *mutagen* = 1. By

Predicted results from all models								
You can copy the results to clipboard, or export into CSV file								
<input checked="" type="checkbox"/> Exclude rows in train data								
<input type="button" value="Copy"/> <input type="button" value="CSV"/>		Search: <input type="text"/>						
SampleID	Observed	inTrain	earth	rpart	ctree	rf	gbm	glmboost
525	1	0	1	1	1	1	1	1
526	1	0	1	1	1	1	0	1
527	0	0	0	0	0	0	1	0
528	0	0	0	0	0	0	1	0
529	1	0	1	1	1	1	1	1
530	1	0	1	1	1	1	0	1
531	1	0	1	1	1	1	0	1
532	0	0	1	1	1	1	0	0
533	0	0	1	1	1	1	1	1
534	0	0	1	1	1	1	1	1

Showing 1 to 10 of 175 entries

Previous 1 2 3 4 5 ... 18 Next

Fig. 2.37 Detailed output from “Predict” command showing observed values and predicted values from each of six machine learning algorithms for cases in the test set (*InTrain* = 0)

contrast, all of the algorithms misclassify chemical 534 in the bottom row. The top half of Fig. 2.38 presents a visualization of the predictive performance of the different algorithms on all chemicals in the test set. In this visualization, the top row, “Observed,” shows the correct classification of each chemical (column), with orange used for mutagens and grey used for non-mutagens. A dendrogram clustering tree is used to automatically order the chemicals to allow simple visual interpretation: the solid orange columns at the right indicate chemicals that were correctly classified as mutagens by all algorithms, while solid grey columns, mostly toward the left, represent chemicals that were correctly classified as non-mutagens. The remaining columns, with both grey and orange cells, are for chemicals that are misclassified by at least some of the algorithms—most frequently, by gbm for this data set.

The bottom half of Fig. 2.38 shows the *calibration curve* for each algorithm. This answers the following question: given a prediction of the probability that a chemical is a mutagen, what fraction of chemicals with that predicted probability actually are mutagens? The diagonal line in each plot is the line of perfect calibration for predictions, in which the fractions of chemicals that are mutagens match the predicted probabilities that they are mutagens. The S-shaped curves show the actual empirical relations between predicted probabilities of being a mutagen (horizontal axis) and the fractions of chemicals that are indeed mutagens (vertical axis). These curves show that earth, rpart, and glmboost are fairly well calibrated for these data, although they tend to underestimate low and high probabilities (i.e., predicted low probabilities should be even lower and predicted high probabilities should be even higher). Both ctree and rf tend to systematically underestimate the probability of being a mutagen over most of their ranges of predicted probabilities. The gbm algorithm has by far the worst calibration.

Several standard metrics are commonly used to summarize and compare the performance of different predictive algorithms. Figure 2.39 shows these standard

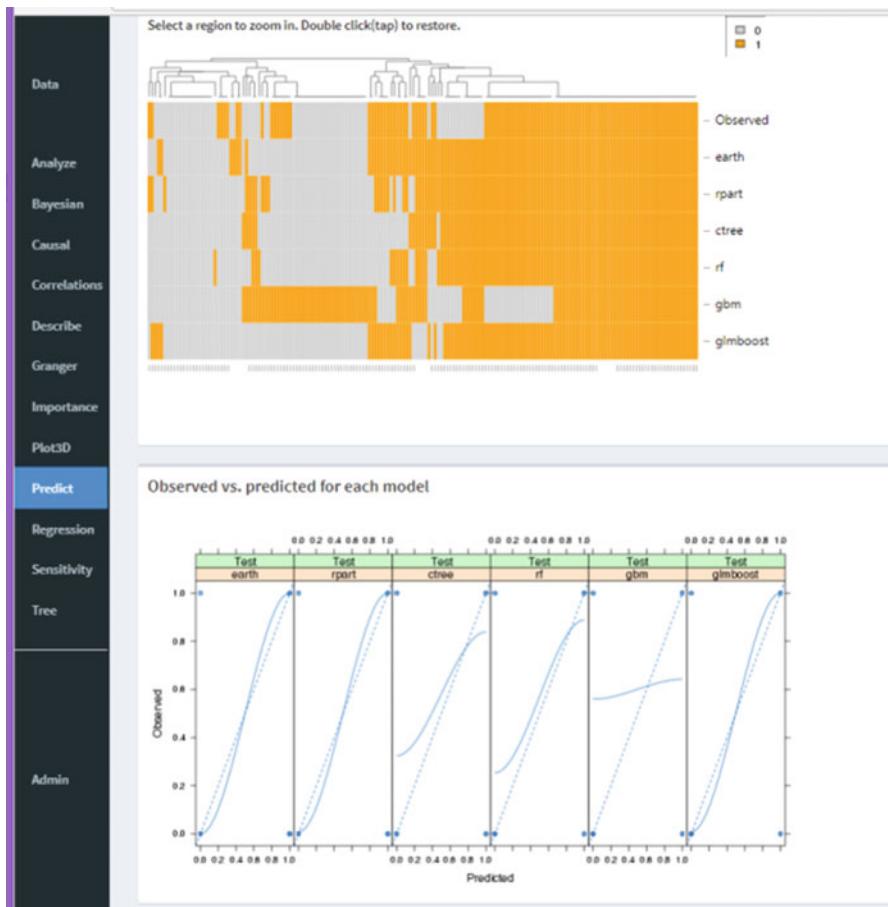


Fig. 2.38 Visualizations of predictive performance of machine learning algorithms (top) and their calibration curves (bottom)

outputs. Clicking on the hyperlink “Simple guide to confusion matrix terminologies” near the top of the screen while using CAT on line pulls up definitions of many of these terms, but the two most important concepts are as follows. The four-quadrant diagrams at the top show the number of false positives (non-mutagens mistakenly classified as mutagens) in the lower left (e.g., 22 for the earth algorithm, 23 for rpart, 21 for ctree, and so on) and the number of false negatives (mutagens mistakenly classified as non-mutagens) in the upper right (14 for earth, 19 for rpart, 29 for ctree, and so on). Better predictive algorithms have smaller lobes and lower frequency counts along this main diagonal. The upper left and lower right give the numbers of true negatives and true positives, respectively: these are the correctly classified chemicals.

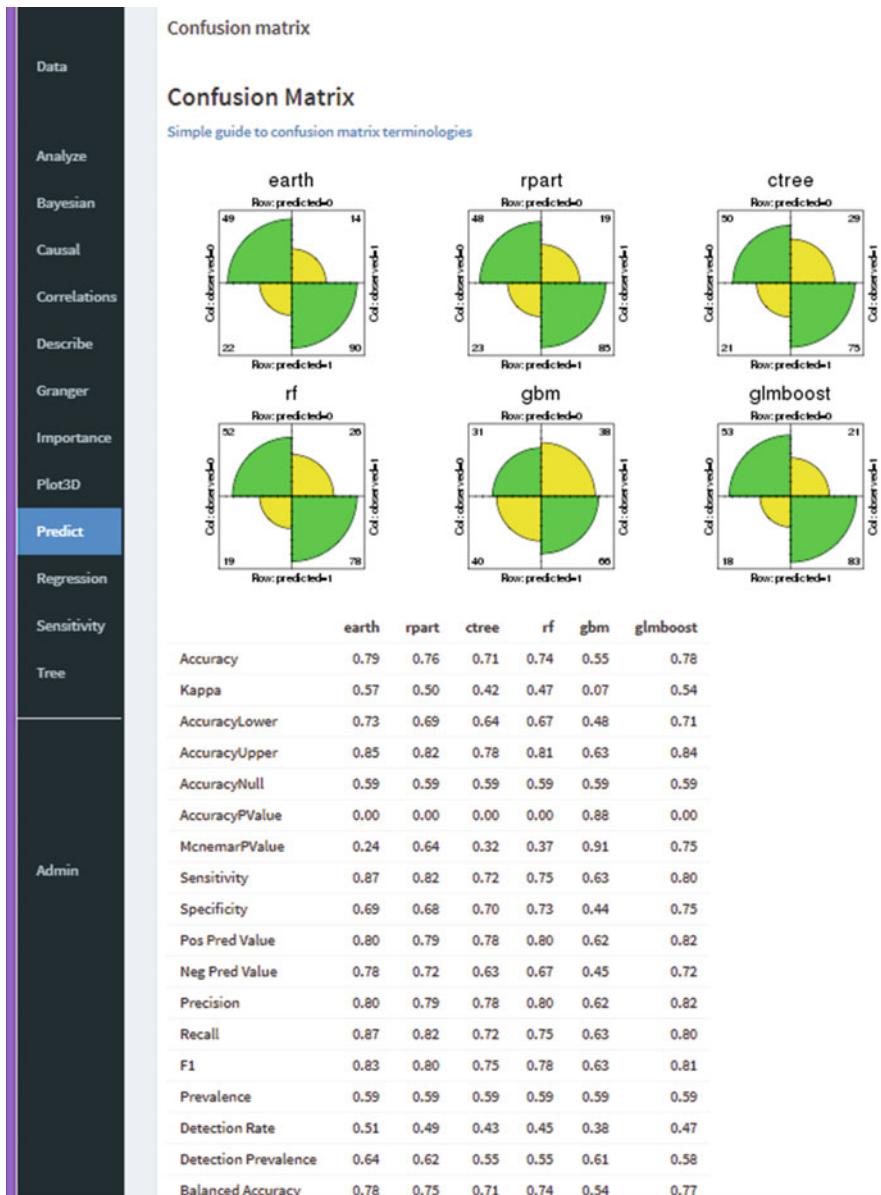


Fig. 2.39 Summary of predictive model performance

Of the many summary performance metrics tabulated below these “confusion matrix” results, one of the most useful is balanced accuracy, shown in the bottom line. This summarizes the probability that each algorithm will predict the classification of a chemical correctly when the prediction task is made as hard as possible by

balancing the samples so that exactly half are mutagens and half are not. This avoids enabling a classifier with poor predictive power to appear to perform well on a highly unbalanced sample (e.g., one with 95% mutagens) simply by guessing the most likely value every time. A balanced accuracy of 0.50 indicates that a predictive algorithm is no more accurate than random guessing, while a balanced accuracy of 1 would indicate a completely accurate classifier. With the exception of gbm, all of the predictive algorithms achieved balanced accuracies of between 0.7 and 0.8 on the test set, indicating substantial predictive power. The earth algorithm performs best by this metric, consistent with its performance in the confusion matrix.

To apply the predictive models to new cases once they have been learned and evaluated, e.g., to predict the mutagenicity of chemicals for which the correct classification is not yet known, the data for these chemicals can be appended to the data table in Fig. 2.35 with blanks left in the *mutagen* column, and the probabilities that *mutagen* = 1 will then be computed for each of these chemicals by each predictive algorithm. The calibration information shown in the S-shaped curves at the bottom Fig. 2.38 can be applied to these probabilities to improve their accuracy by correcting for any systematic distortions (departures from the line of perfect calibration) revealed when the algorithms are evaluated on the test set. Such *probability calibration* adjustments to improve predicted class probabilities are performed automatically by modern machine learning software (e.g., <http://scikit-learn.org/stable/modules/calibration.html>).

The methods for automated predictive analytics surveyed in this section provide a firm computational foundation for investigating questions such as whether, and to what degree, one variable helps to predict another, e.g., as revealed in PDPs and balanced accuracy scores. For time series variables, these methods can also be used to test whether the past values of some variables help to predict future values of other variables, and can do so without committing to the parametric modeling restrictions of classical Granger causality tests. Insofar as direct causes must be informative about, and help to predict, the values of their effects, predictive analytics can serve as a valuable screen for identifying potential direct causes of an effect by determining which variables help to predict its values, even after conditioning on the values of other variables. This is one of the key ideas used to learn causal graph models from data, and hence causal discovery algorithms make heavy use of machine learning methods for predictive analytics.

Learning Causal BN Models from Data: Causal Discovery Algorithms

A key aspect of causal BN modeling is how to learn causal BN models from data. There are two main parts to this task: learning the *structure* of the causal BN model, i.e., the causal DAG; and estimating the conditional probability tables (CPTs) that quantify how each variable depends upon its parents. In sufficiently large and varied

data sets, the CPT estimation task is straightforward. A cross-tabulation table giving the empirical frequency distribution of the value of each variable, given the values of its parents, suffices if variables are few and discrete; these empirical conditional distributions are the maximum-likelihood estimates of the underlying CPT values. Bayesian methods (e.g., conditioning Dirichlet priors on any available data to obtain posterior distributions for the CPT entries) have been developed for smaller data sets, and have recently been extended to allow for mixtures of such priors for drawing inferences about individual cases from heterogeneous populations (e.g., Azzimonti et al. 2017). For larger data sets and for variables with many values, or for continuous variables, CART trees or regression model developed for the value of each node as a function of the values of its parents provide parsimonious representations of the CPT information.

The more difficult task is estimating DAG structures from data. This is often called the *structure learning* problem. State-of-the-art BN learning software typically provides a mix of algorithms for solving it. Some of these were used to obtain the DAG models in Figs. 2.24, 2.25, 2.28, and 2.33. These algorithms incorporate a variety of ideas and principles for detecting information-based causal relationships (e.g., predictive causation) between variables. Among the most useful are the following.

1. *Conditional independence constraints:* As previously discussed in some detail, effects are not conditionally independent of their direct causes in a DAG model, but they are conditionally independent of their more remote ancestors given their parents if the Causal Markov Condition (CMC) holds. Software such as DAGitty details the testable conditional independence constraints implied by a DAG model (see Fig. 2.32); conversely, applying CART trees or other tests to identify conditional independence relations among variables in a data set (to within the limits of accuracy of the test) constrains the set of possible DAGs that are consistent with these relations. Conditional independence constraints restrict the set of possible DAG structures to the Markov equivalence class that is compatible with the constraints.
2. *Composition constraints:* If X determines Y and Y determines Z via suitable smooth, deterministic functions, then the composition of these functions should describe how X determines Z and consistency conditions such as the chain rule (that is, $dZ/dX = (dZ/dY)(dY/dX)$) will hold. If these functions describe manipulative causal relations, rather than only statistical associations, then a small change of size dx in X should cause Y to change by approximately $dy = (dY/dX)dx$, and this in turn should cause Z to change by approximately $dz = (dZ/dY)dy = (dZ/dY)(dY/dX)dx$. Such constraints can be generalized to DAG networks, as in the rules for path analysis for the special case of linear functions and normally distributed error terms. They imply consistency conditions for relations among estimated coefficients, and hence can be used to test whether a proposed DAG structure and set of dependency functions is consistent with data, in the sense that these implied consistency conditions hold. In the probabilistic case, composition relations still hold: thus, in the DAG model $X \rightarrow Y \rightarrow Z$, if Z depends

probabilistically on Y via a CPT $P(z \mid y)$ and Y depends probabilistically on X via a CPT $P(y \mid x)$, then the composition constraint $P(z \mid x) = \sum_y P(z \mid y)P(y \mid x)$ should hold.

3. *Scoring and optimization methods:* Once a DAG structure has been specified, its CPTs can be fit to available data as already described, e.g., by estimating a CART tree for each node as a function of its parents. The completed BN model, in turn, can be used to assess the likelihood of the data given the model, or other related score functions reflecting measures of goodness-of-fit between the observed data and the predictions implied by the model. Variations in the DAG such as adding, deleting, or reversing arrows can then be made to try to find a higher-scoring model. This incremental optimization (or “hill-climbing”) in the space of models can be continued until no further improvements can be found. Heuristics for combinatorial optimization, such as tabu search (which restricts the allowed moves at each stage to prevent cycling or excessive concentration of search in the neighborhood of the current best networks) are often used to improve search efficiency. Scoring methods, including the hill-climbing (hc) algorithm used as the default in the *bnlearn* R package and in CAT, combine computational efficiency with competitive performance compared to constraint-based methods (Nowzohour and Bühlmann 2016). They are among the most successful current techniques for BN learning. Hybrid methods that combine constraints and scoring are also popular for the same reason, although no single BN learning algorithm has proved best for all data sets.
4. *Simplicity in error terms: Nonlinear and non-Gaussian models:* Suppose that a causal discovery algorithm seeks a causal model described by a structural equation of the form $\text{observed effect} = f(\text{cause}) + \text{error}$, where f is a possibly unknown and nonlinear function and error is a measurement error term, not necessarily normally distributed, i.e., Gaussian. If the observed values for the effect variable Y are plotted against corresponding values of the cause variable X and a non-parametric smoothing regression curve (e.g., loess, kernel regression, or iteratively reweighted least squares) is fit to the resulting scatterplot, then the scatter of data points around this curve due to the error term should look roughly the same for all values of X , as shown in Fig. 2.40a. In this figure, the true data-generating process is $Y = X^2 + \text{error}$, where error is uniformly distributed between 0 and 1 (and hence is biased upward by 0.5) for all values of X . On the other hand, plotting X against corresponding observed Y values will typically give vertical error scatters that depend on Y if f is nonlinear or if the error term is non-Gaussian (Shimizu et al. 2006). This is shown in Fig. 2.40b, using the same data as in Fig. 2.40a. Figure 2.40b plots X values against corresponding observed Y values. Clearly, the error variance is smaller at the ends than in the middle, in contrast to Fig. 2.40a. Such heteroscedasticity reveals that the correct causal ordering of X and Y is that X causes Y , rather than Y causing X . More generally, when a causal model implies that error terms (residuals) for a dependent variable, given the values of its causes, have a simple form, the empirical distribution of these error terms can be used to determine which is the dependent variable and which are the explanatory variables that cause it. Linear models with Gaussian

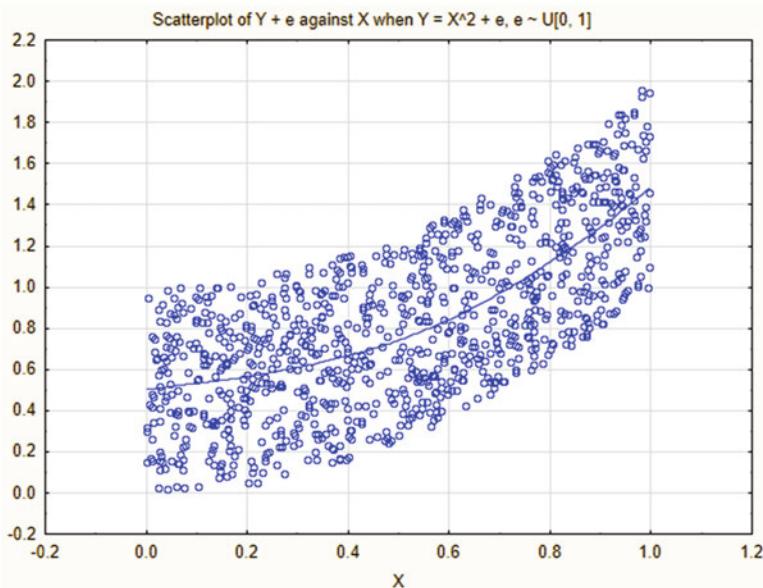
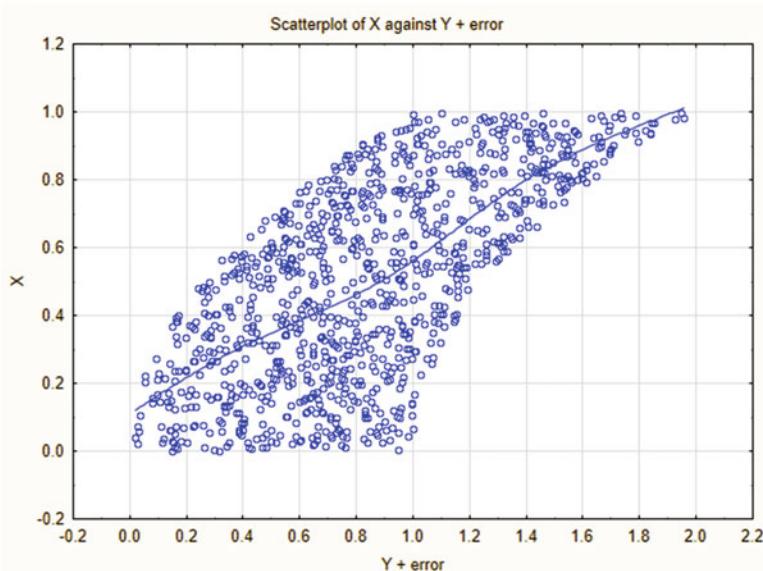
a**b**

Fig. 2.40 (a) If $\text{effect} = f(\text{cause}) + \text{error}$, then plotting effect vs. cause gives a scatterplot with additive errors. The scatterplot shows $Y = X^2 + e$ where $e \sim U[-1, 1]$. (b) Conversely, plotting cause vs. effect gives non-additive errors

errors—the traditional regression setting—are an exception, but nonlinear models or linear models with non-Gaussian errors (“LiNGAM” models) suffice to identify correct causal orderings of the variables in a DAG model under many conditions (Shimizu et al. 2006; Tashiro et al. 2014; Nowzohour and Bühlmann 2016).

5. *Linear Gaussian models using Tetrad*: At the opposite end of the spectrum from nonlinear and non-Gaussian causal models are linear causal models with normally distributed (Gaussian) errors. Programs such as Tetrad (www.phil.cmu.edu/tetrad/) exploit the assumptions of linearity and normal errors to address the important practical problems of estimating and quantifying causal DAG models in the presence of hidden confounders or other latent (unmeasured) variables and estimating models with feedback loops. The presence of latent variables and their connections to observed variables are inferred from otherwise unexplained patterns of correlation among measured variables. Like Netica and other BN programs, Tetrad also works with discrete random variables, in which case the assumptions of linear effects and normal errors are not required.
6. *Invariance and homogeneity principles*. The fact that a causal CPT gives the same conditional probability distribution for a response variable whenever its parents have the same values (used as the information conditioned on) provides a basis for causal decision tree programs that seek to estimate average causal effects as the differences in conditional means between leaf nodes in CART-like trees constructed to have a potential outcomes causal interpretation (Li et al. 2017). It is also the basis for the Invariant Causal Prediction (ICP) package in R (Peters et al. 2016). This package supports causal inference in linear structural equation models (SEMs), including models with latent variables, assuming that data from multiple experimental and/or observational settings reflect additive effects of interventions. Heinze-Deml et al. (2017) discuss non-parametric generalizations.
7. *Structural causal ordering algorithms*: In a system of structural equations, the ones that must be solved first, so that the values of their variables can be used to determine the values of other variables, imply a partial causal ordering of the variables: those that must be solved for first, or that are exogenous to, other variables are possible causes of them. This concept of causal ordering was introduced for systems of linear structural equations by Simon (1953) and was subsequently generalized to include nonlinear structural equations and dynamical systems modeled by ordinary differential equations (ODEs) and algebraic constraints among equilibrium values of variables (Simon and Iwasaki 1988).
8. *Timing considerations*: Time series causal inference algorithms use the fact that information flows from causes to effects over time to constrain the possible causal orderings of variables to be consistent with observed directions of information flow among time series variables. Granger causality and its non-parametric generalization, transfer entropy, provide one set of ordering constraints. Recently, constraint satisfaction algorithms have been applied to infer causal structure from time series of observations subsampled at a rate slower than that of the underlying causal dynamics of the system being observed. In many real-world systems, variables in causal networks are constantly jostled by exogenous shocks,

disturbance, and noise and the effects of these perturbations spread through the network of variables over time. Under certain conditions, such as uncorrelated random shocks and linear effects, the structure of the causal network can be inferred by studying the effects of the shocks on observable variables, even if the underlying causal graph has latent variables and cycles and the shocks are unknown. This possibility has been explored via recent packages and algorithms such as BACKSHIFT (Rothenhausler et al. 2015), although more work needs to be done to extend these developments to non-parametric models, analogous to transfer entropy-based reconstruction of DAG models.

Each of these causal inference principles has led to a substantial technical literature and to software implementations that make them relatively easy to apply to real data sets. For end users, perhaps the most important points are as follows:

1. Multiple causal discovery algorithms are now readily available, as illustrated in Fig. 2.33 for four algorithms. Most are available via free R packages such as *bnlearn* and *CompareCausalNetworks*. Special packages are available for commonly encountered special situations, e.g., the *sparsebn* package (Aragam et al. 2017) is available for learning Bayesian networks from bioinformatics data with many more variables than data records, where some or all of the data records may reflect experimental interventions.
2. Available algorithms incorporate different conceptual principles for inferring causation from data, allowing the possibility of cross-checking each other and enabling robust causal inferences that are supported by multiple principles and contradicted by none.
3. Several algorithms and principles (including conditional independence tests, scoring algorithms for discrete BNs, invariance and homogeneity tests, and transfer entropy among time series) have non-parametric versions. This allows them to be applied to data in the absence of known or specified parametric models.

In addition to these principles and algorithms for discovering potential causal relations and causal graphs and quantifying CPTs from structured data (i.e., data tables or data frames), there have been several research efforts to develop software that can acquire causal Bayesian networks automatically from text (Sanchez-Graillet and Poesio 2004; Trovati 2015) or that can help humans construct them from causal maps elicited from experts (Nadkarni and Shenoy 2004). Such software recognizes that text strings or expert claims such as “Long working hours create stress that can increase heart attack risks” correspond roughly to causal graph models such as the following: *Long_working_hours* → *Stress* → *Heart_attack_risk*. Various heuristics have been proposed for quantifying the corresponding CPTs by counting and taking ratios of different mentions of each condition and pairs of conditions. However, learning about causality directly from texts or by being told by experts and then representing the results by BNs is not yet, to our knowledge, a commercially viable technology. Other knowledge representation and inference techniques for identifying causal information from text, especially machine-learning algorithms for natural

language processing, appear to be very promising (Asghar 2016). It seems plausible that text mining will become an increasingly important contributor to enhanced causal discovery systems in the years ahead.

Taken together, current causal discovery algorithms provide a flexible set of principles and algorithms for learning about possible causal relationships among variables in structured data. These methods emphasize information-based concepts of causation—that is, concepts such as predictive causation, structural causation, manipulative causation, and mechanistic causation that reflect the information principle that causes help to predict their direct effects, even after conditioning on the values of other variables. In this sense causes are informative about their effects; conversely, conditional probability distributions of effects are derived from the values of their direct causes, which therefore help to predict and explain their values.

The empirical performance of different causal discovery algorithms has been assessed in many challenges and competitions (e.g., Hill 2016). Developing, applying, and evaluating algorithms for discovering causal BN models and other causal graph models from data is now a thriving specialty within machine learning, systems biology, artificial intelligence, and related fields. Several professional societies host regular competitions to assess progress in causal discovery algorithms. Advances are reported in sources such as *The Journal of Causal Inference*, *Artificial Intelligence*, *Neural Information Processing Workshops on Causality, Uncertainty in Artificial Intelligence* (UAI) conference proceedings, and documentation of algorithms implemented in R and Python. Applications specifically to inference of gene regulatory networks have been developed through an important series of DREAM challenges (<http://dreamchallenges.org/>), leading to useful benchmark results and to empirical verification that current causal discovery algorithms are indeed useful in a range of systems biology and bioinformatics applications (Schaffter et al. 2011; Hill 2016).

Comparison of Causal Discovery to Associational Causal Concepts and Methods: Updating the Bradford Hill Considerations

Most causal claims that garner headlines today, such as frequent reports about adverse health effects of various substances, are not based on applying the foregoing principles of causal discovery. Instead, they usually reflect subjective judgments about causality. An approach to forming such judgments has been developed and widely applied within epidemiology over the past half century to draw important-seeming, policy-relevant conclusions from epidemiological data. The conclusions are justified as consensus judgments of selected experts and authoritative bodies based on explicit considerations such as whether observed exposure-response associations are judged to be strong, consistent, and biologically plausible. The left column of Table 2.4 provides a fuller list of the considerations about evidence that are most commonly used to structure and support such judgments.

Table 2.4 Comparison of Bradford-Hill and causal discovery principles

Bradford-Hill considerations	Causal discovery algorithms and principles
<i>Strength of association:</i> Stronger associations are more likely to be causal	<ul style="list-style-type: none"> • <i>Information principle:</i> Causes are informative about their direct effects and help to predict them • <i>DAG learning:</i> Effects are not conditionally independent of their direct causes. <i>bnlearn</i> package
<i>Consistency</i> of findings across populations, study designs, times, locations, investigators, etc.	<ul style="list-style-type: none"> • <i>External consistency:</i> Invariance, homogeneity, and transportability of CPTs • <i>Internal consistency:</i> Similar effects estimated via different adjustment sets, principles, and algorithms
<i>Specificity</i> of effects: A specific cause produces a specific effect	LiNGAM for one cause, one effect: $y = f(x) + \text{error}$
<i>Temporality:</i> Causes precede their effects	Information flows from causes to their effects over time. Granger causality tests, transfer entropy
<i>Biological gradient:</i> Data show a dose-response pattern (larger responses at higher exposures)	<ul style="list-style-type: none"> • Partial dependence plots (PDPs) show gradient • Composition principle along DAG paths • LiNGAM: $y = f(x) + \text{error}$
<i>Biological plausibility:</i> Plausible biological mechanism	<ul style="list-style-type: none"> • Likelihood principle: DAG model explains the data • Structural causation; d-connectivity links exposures to responses (e.g., in DAGitty package)
<i>Coherence:</i> Agrees with knowledge of disease biology	<ul style="list-style-type: none"> • d-connectivity of dose and response • Knowledge-based constraints in <i>bnlearn</i> package
<i>Analogy:</i> Similar causes are believed to cause similar effects	Deep learning and automatic abstraction methods for generalizing patterns from data on specific instances
<i>Experiment:</i> Reducing exposure reduces effect	<i>BackShift</i> algorithm for unknown interventions, latent variables; <i>ComparingCausalNetworks</i> package in R

This approach sprang largely from an influential essay by Sir Austin Bradford Hill (1965). The considerations on the left side of Table 2.4 are often referred to as the “Hill criteria,” although Hill wrote that “What I do not believe—and this has been suggested—[is] that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*.” Hill’s approach was later incorporated into various “weight-of-evidence” approaches for systematically documenting considerations and judgments about whether associations are likely to be causal. Table 2.4 matches the original Hill considerations in the left column with roughly corresponding principles of modern causal discovery algorithms in the

right column. These considerations and correspondences are discussed in more detail next and contrasted with causal discovery techniques and BN learning algorithms. Readers who do not care about a detailed comparison can find a briefer discussion in Chap. 14 of the Hill considerations and how to update them using modern ideas and methods.

Strength of Association

The foremost consideration for Hill was strength of association. He wrote, “First upon my list I would put the strength of the association,” and this consideration has subsequently been interpreted by authorities such as the International Agency for Research on Cancer (IARC) to mean that “A strong association (e.g., a large relative risk) is more likely to indicate causality than a weak association.” However, this principle is mistaken. Association is not causation. Evidence of association is not evidence of causation, and strong association does not necessarily (or even, perhaps, usually) indicate likely causation.

A stronger association is not, simply by virtue of being stronger, any more likely to indicate causality than a weaker one. Indeed, a stronger exposure-response association may simply indicate stronger sampling and selection biases, or stronger modeling errors and biases, or stronger coincidental historical trends, or stronger confounding, or stronger model specification errors, or other strong threats to internal validity (Campbell and Stanley 1963). The following conceptual examples make this point; practical examples are discussed later.

- *Example 1: Causation without association.* Suppose that the kinetic energy (KE) of a particle is causally related to its velocity (V) and mass (M) via the structural equation $KE = \frac{1}{2}MV^2$. If the velocities of a collection of particles are uniformly distributed between -1 and $+1$ (or are normally distributed with mean 0), then the association between V and KE , as measured by standard metrics such as Pearson’s correlation or Spearman’s correlation will be approximately zero in a large sample of particles (and is exactly zero on average), even though these variables are causally related as strongly as possible, i.e., deterministically.
- *Example 2: Correlation without causation:* Conversely, suppose that $X(t)$ is an exposure variable expressed as a function of time, t , and that $Y(t)$ is a response variable, also expressed as a function of time. For simplicity, suppose that each of $X(t)$ and $Y(t)$ independently is assigned a random linear trend with mean zero; thus, each has a random average slope and tends to increase or decrease linearly with time unless the slope happens to be exactly 0. Then, with probability 1, their values will be correlated even though each is assigned its slope independently of the other and neither depends on the other. Moreover, the correlation between them will be as strong as possible ($R^2 = 1$) if measurement error and random variation are negligible and sample sizes are large. Yet, this strong association indicates nothing about causality.

In practice, many non-stationary random processes, both time series and spatial, exhibit temporal or spatial trends. Any two variables with trends over the same interval of time or the same region of space will have correlated values, even if their values are determined independently of each other and there is no causal relation between them. Thus, the strength of the associations between them indicates nothing about causality. For example, independent random walks are very likely to have significantly correlated values, illustrating the phenomenon known as *spurious regression* in time series analysis. In real-world applications, associations commonly occur between time series variables (e.g., air pollution levels and mortality rates in Dublin) and between spatial variables (e.g., distance from oil and gas wells, point sources of pollution, or randomly selected locations, and rates of various ailments in people) whether or not there is any causal relationship between them. Many epidemiological journal articles report on such associations and conclude without evidence that they raise concerns about health effects of exposures. Applying appropriate methods of analysis (e.g., Granger causality testing, conditional independence tests) can help to determine whether predictive causation, rather than mere association, holds between variables.

- *Example 3: Association created by model specification error.* Suppose that X is an exposure variable and that Y is a response variable and that each is independently uniformly distributed between 0 and its maximum possible value (or, more generally, has a continuous distribution with non-negative support). Anyone wishing to make a case for banning or reducing exposures within the framework of the Hill considerations can create a statistical association between X and Y , even if they are independent (or, indeed, even if they are significantly negatively associated, as might occur if exposure has a protective effect), by specifying a statistical model of the form $E(Y|X) = KX$, i.e., risk is proportional to exposure, and then estimating the slope parameter K from data and interpreting it as a measure of the potency of X in causing Y . Since the values of both variables are positive, the estimated value of K will also be positive, guaranteeing a positive estimated “link” or association between X and Y no matter what the true relation (if any) between them may be. Of course, a regression diagnostic plot, such as a plot of residuals, would immediately reveal that the assumed model $E(Y|X) = KX$ does not provide an accurate fit to the data if the true data-generating process is quite different from the assumed model, e.g., if it is $E(Y|X) = 0.5$ (i.e., Y is independent of X) or if it is $E(Y|X) = 10 - KX$ (i.e., Y is negatively associated with X). But practitioners who create statistical associations by this method typically do not show diagnostic plots or validate the assumed model, allowing model specification error to be interpreted as evidence for causality because it creates positive associations. This method has been used successfully in the United States to justify banning animal antibiotics and has been recommended by regulators for widespread use (Bartholomew et al. 2005).
- *Example 4: Ambiguous associations.* Suppose that $Y = 10 - X + Z$ and that $Z = 2X$. Then the *direct effect* on Y of an exogenous increase in X by 1, holding other variables (namely, Z) fixed, is to reduce Y by 1; but the *total effect*, allowing

Z to adjust, is to increase Y by 1. “The association” between Y and X depends on what else (Z in this example) is conditioned on in modeling the relation between them. No single measure of association can simultaneously describe all of the possible associations conditioning on different subsets of other variables. Clarifying which subsets of variables should be conditioned on to obtain a causally interpretable association (namely, those in a properly constructed minimal adjustment set, e.g., as produced by DAGitty) is not trivial, but without such clarification, the causal significance of an association, if any, is unknown.

Hill (1965) declared that “I have no wish, nor the skill, to embark upon philosophical discussion of the meaning of ‘causation’” (Hill 1965). As a result, he did not distinguish among distinct varieties of causation, such as associational, attributive, counterfactual, predictive, structural, manipulative, and explanatory causation. Nor did he distinguish among different types of causal effects, such as direct effects, indirect effects, total effects, and mediated effects. However, examples like these make clear that such distinctions are essential for understanding what, if anything, causal assertions imply about how changing some variables (e.g., exposures) affects others (e.g., health effects). Strength of association does not necessarily—or, in practice, usually—shed light on this crucial practical question. Rather, a reported strong association may simply result from particular study design and modeling choices, which we will refer to generically as assumptions. Even a strong association may disappear or be reversed in sign if different assumptions are made. To address the obvious objection that such *assumption-dependent conclusions* do not necessarily reveal truths about the world, it is common practice to present statistical goodness-of-fit measures and sensitivity analyses supporting the thesis that the selected set of modeling assumptions describe or fit the data better than some alternative assumptions. However, goodness-of-fit comparisons are usually quite insensitive to incorrect model specifications and do not establish that the best-fitting model considered provides a usefully accurate description of the data-generating process, let alone that causal interpretations of associations are justified.

Modern causal discovery algorithms overcome these substantial challenges to the usefulness of association as an indicator of causality by replacing *association* with *information*. While stronger associations between variables are not necessarily more likely to indicate causality, it is true that direct causal parents always provide at least as much information about a dependent variable as do its more remote ancestors (and usually strictly more, unless the parents are deterministic, invertible functions of the ancestors). This a corollary of information theory (the data processing inequality) when “information” is interpreted specifically as mutual information between random variables, measured in units such as bits (Cover and Thomas 2006, p. 34). Qualitatively, a variable in a causal DAG is typically not conditionally independent of its direct causes, even after conditioning on other variables (again with certain rare exceptions, such as if a parent and a more remote ancestor have identical values), but it is conditionally independent of its more remote ancestors and non-descendants, given its parents. Conditional independence tests and estimates of CPTs in a causal BN can be based on non-parametric methods (e.g., using CART trees), thus avoiding

the pitfalls of mistaking parametric model specification errors for evidence of causality. In short, the frequently incorrect *association principle*, stating that causes are more likely to be strongly associated with their effects than are non-causes, can be updated to a more useful *information principle* stating that its direct causes provide unique information about an effect variable that helps to predict it (so that conditioning on the value of a parent reduces uncertainty about the effect, including about its future values if it is a time series variable, where uncertainty is measured by the expected conditional entropy of its distribution). A brief, approximate summary of this principle is that *direct causes are informative about (i.e., help to predict) their effects*, even after conditioning on other information; moreover, direct causes are typically more informative about their effects than are more remote indirect causes. Exceptions can occur, e.g., if some variables coincide with or are deterministic functions of each other, but these information-theoretic principles are useful in practical settings and play substantially the role that Hill envisioned for strength of association as a guide to possible causation, but more reliably.

Replacing the association principle with the information principle avoids the difficulties in the foregoing examples, as follows. If $KE = \frac{1}{2}MV^2$ and V is uniformly distributed between -1 and $+1$, then KE and V have positive mutual information even though the average correlation between them is 0. If $X(t)$ and $Y(t)$ are time series variables with random linear trends, then $Y(t)$ cannot be predicted any better from the history of $X(t)$ and $Y(t)$ than from the history of $Y(t)$ alone, even if they are perfectly correlated. If X and Y are two independent random variables, each with positive values, then the mutual information between them is 0 even though fitting the misspecified model $E(Y|X) = KX$ to pairs of values for X and Y (e.g., to exposure levels and corresponding response levels using ordinary least squares) for a sufficiently large sample size would yield a positive estimate of K . Finally, if $Y = 10 - X + Z$ and $Z = 2X$ are structural equations, then a corresponding DAG model will show that Y has X and Z as its parents and that Z has X as its parent; X is the exogenous input to this system, and the reduced-form model $Y = 10 + X$ for the total effect of X on Y is easily distinguished from the partial dependency relation $Y = 10 - X + Z$ via the DAG showing which variables depend on which others.

Consistency of Association

Several of Hill's other proposed principles appeal strongly to intuition. They include consistency (different studies and investigators find the same or similar exposure-response associations or effects estimates in different populations and settings at different times); biological plausibility (the association makes biological sense), coherence (the association is consistent with what is known about biology and causal mechanisms) and analogy (similar causes produce similar effects). It is now understood that such properties, based on qualitative and holistic judgments, appeal powerfully to psychological heuristics and biases such as motivated reasoning (finding what it pays us to find), groupthink (conforming to what others appear to

think), and *confirmation bias*, i.e., the tendency of people (including scientists) to find what they expect to find and to focus on evidence that agrees with and confirms pre-existing hypotheses or beliefs, rather than seeking and using disconfirming or discordant evidence to correct misconceptions and to discover how the world works (Kahneman 2011).

Example: Confirmation Bias and the Wason Selection Task

To experience confirmation bias first-hand, consider the following question about evidence and hypothesis-testing. Suppose that you are presented with four cards. It is specified that each of the cards has a letter on one side and a number on the other. You can see the faces that are turned up of the four cards, and they are as follows:

A		L		2		5
---	--	---	--	---	--	---

Which of these four cards must necessarily be turned over to reveal what is on the other side in order to determine whether the following hypothesis is true?

Hypothesis: Each card with a vowel (A, E, I, O, or U) on one side has an even number (2, 4, 6, or 8) on the other.

In other words, what is the smallest subset of the cards that must be turned over to test the validity of this hypothesis? Most people, including scientists, find it difficult to think clearly and correctly about even such simple but abstract relations between hypotheses and the evidence needed to test them. By far the most common answer is that cards A and 2 must be turned over to confirm whether the A has an even number on its other side and whether the 2 has a vowel on its other side. The correct answer is that cards A and 5 must be turned over. It is indeed necessary to confirm whether the A has an even number on its other side, but it is also necessary to confirm that the 5 does *not* have a vowel on its other side, since finding a vowel there would disconfirm the hypothesis. (Cards L and 2 are irrelevant, since neither one can disconfirm the hypothesis no matter what is on the other side.) This often-repeated experiment, called the Wason selection task, and numerous variations on it, illustrate that people naturally think about confirming evidence more easily than about disconfirming evidence in many settings.

In light of such powerful heuristics and biases in judgments under uncertainty, many of which were elucidated starting in the 1970s (Kahneman 2011), findings of consistency in effects estimates and associations across multiple studies should raise a suspicion of p-hacking and confirmation bias: the possibility that different investigators varied their modeling assumptions until they produced the results they expected or hoped to find, perhaps based on published results from earlier studies. Consistency *per se* is not evidence of causation unless other plausible explanations, such as p-hacking, can be ruled out. Indeed, logically, the proposition that a true causal effect is expected to generate consistent associations across studies is questionable, insofar as different studies involve different conditions and distributions of covariates in the population that should affect estimated associations and statistical

effects. (Moreover, even if the premise were true that “Causality implies consistency,” it would not necessarily justify the conclusion “Consistency implies causality.” That is the formal logical fallacy of affirming the consequent, analogous to turning over card 2 in the Wason section task to find confirming evidence, which is logically irrelevant for testing the stated hypothesis.)

Causal graph models can improve on the traditional consistency consideration by providing much clearer tests of agreement between theory and data. More convincing than agreement with previous findings (which is too often easily accomplished via p-hacking) is to find associations and effects estimates that *differ* across studies, and to show that these differences are successfully predicted and explained by invariant causal CPTs applied to the different joint distributions in the populations of causally relevant covariates (e.g., sex, age, income, health care, etc.) Modern *transport formulas* for DAG models allow such detailed prediction and explanation of empirically observed differences in effects in different populations (Heinze-Deml et al. 2017; Bareinboim and Pearl 2013; Lee and Honavar 2013; <https://cran.r-project.org/web/packages/causaleffect/causaleffect.pdf>). Similar techniques allow the results of multiple disparate studies to be combined, generalized, and applied to new settings using the invariance of causal CPTs, despite the variations of marginal and joint distributions of their inputs in different settings (Triantafillou and Tsamardinos 2015; Schwartz et al. 2011). Even within a single study, DAGitty algorithms will often produce multiple alternative adjustment sets, allowing the same causal effects to be estimated by conditioning on different subsets of other variables, as illustrated in Figs. 2.29 and 2.30. *Internal consistency*, in the sense that estimates of specified total or direct causal effects using different adjustment sets for the same data, and *external consistency*, in the sense that the same invariant causal CPTs are found to hold in settings with very different joint distributions of the values of their direct parents, provide useful refinements to the general consideration of “consistency,” and help to distinguish consistency that represents genuine predictive power (based on discovery of causal invariants that can be applied across multiple settings) from consistency arrived at by retrospective p-hacking to make results agree with expectations.

Plausibility, Coherence, and Analogy of Association

The considerations of biological plausibility, coherence, and analogy also appeal to, and perhaps encourage, confirmation bias if judgments about the plausibility and coherence of observations, study results, and analogies reflect how well they conform to prior expectations. This can inhibit discovery of unexpected truths. It can also encourage use of logically irrelevant or inadequate information to support preconceptions. For example, IARC (2006) suggests as a principle to use in evaluating causal evidence that if a chemical causes cancer in rats or mice, then it is biologically plausible that it will do so in people. But rodents have organs (e.g., Harderian and Zymbal glands) that people do not. They develop cancers via

mechanisms that do not occur in people (e.g., alpha 2 mu globulin protein drop nephropathy in male rats, but not in female rats or in other species). Thus, what is considered plausible may depend on what is known about relevant differences between species. This is consistent with Hill's own caveat that "What is biologically plausible depends upon the biological knowledge of the day" (Hill 1965).

Similarly, many scientific papers that argue that exposures might plausibly pose human health risks use logic similar to the following:

- A. Substance X induces production of reactive oxygen species (ROS), oxidative damage, and proliferation of damaged target cells via an NF- κ B signaling pathway.
- B. It has been found that several carcinogens increase cancer risk because they induce production of reactive oxygen species (ROS), oxidative damage, and proliferation of damaged target cells via the NF- κ B signaling pathway;
- C. Therefore it is plausible that X is also a carcinogen, by analogy to these known chemical carcinogens and supported by mechanistic or mode-of-action evidence about the signaling pathways and responses involved.

Such reasoning may seem quite convincing until we reflect that it is the same basic syllogism as the following obvious fallacy:

- A. X causes responses in living animals;
- B. Known carcinogens cause responses in living animals;
- C. Therefore X is probably a carcinogen.

The fallacy in both cases is that the causal relevance of the described similarities between (A) and (B) is unknown: it is usually unknown whether the described changes caused by X (e.g., ROS production, oxidative damage, etc.) are the specific ones that cause cancer, or whether they are simply similar but normal homeostasis-preserving responses of healthy organisms to stimuli. Responses such as signaling to the nucleus via specific (e.g., NF- κ B) pathways and production of ROS occur in healthy cells and organisms in response to a wide variety of stimuli, as well as in disease processes, and whether they should be interpreted as evidence for a health risk requires far more specific molecular biological detail than is usually provided in arguments with this syllogistic structure. For example, many biological processes have bistable or multistable structures in which stimuli with slightly different intensities or durations can elicit qualitatively very different responses, e.g., healthy vs. pathological. Qualitative descriptions and analogies often do not provide the essential quantitative details required to determine which qualitative response will occur following an exposure.

DAG methods and related causal graph models can improve upon the considerations of plausibility, coherence, and analogy. They replace qualitative judgments about whether belief in causality is coherent, plausible, and analogous to other relevant situations with more definite and independently reproducible criteria. The criterion of *d-connectivity* in DAG models, as calculated via the algorithms in

DAGitty and similar software, establishes whether it is plausible and coherent to believe that exposure is a possible cause of the responses that are attributed to it, in the sense that information can flow from exposure to the response variables. If not, then the data do not indicate that it is plausible that exposure could be a cause of the responses. Knowledge-based constraints for plausibility, such as that daily temperatures might be a cause of death counts, but death is not a possible cause of daily temperatures, can be incorporated into BN learning programs (e.g., using white lists and black lists for permitted and forbidden arrows in *bnlearn*, or using CAT's source and sink constraints). Quantitatively, estimates of possible effect sizes (or, for some algorithms, bounds on possible effect sizes) calculated from the constraints imposed by observations in a causal BN model can help to determine whether estimated effects of exposures on responses are plausible and consistent with what is known. For example, a PDP for the total effect of exposure on the conditional expected value of response can be used to check whether epidemiological estimates that attribute a certain fraction of responses to exposures are consistent with the causal BN model learned from available data. Instead of drawing subjective analogies across chemicals or studies, information can be combined across studies using algorithms that pool dependence and conditional independence constraints from multiple source data sets having overlapping variables, obtained under a possibly wide range of different experimental and observational conditions, and that automatically identify causal graph models (with unobserved confounders allowed) that simultaneously describe the multiple source data sets (Triantafillou and Tsamardinos 2015).

By applying well-supported, publicly available, pre-programmed causal discovery algorithms to data, researchers can minimize effects of confirmation bias, motivated reasoning, p-hacking, and other types of investigator bias. Such algorithm can replace judgments that may be difficult or impossible to verify or independently reproduce with data-driven conclusions that can easily be reproduced by others simply by running the same algorithms on the same data. Applying several different causal discovery algorithms based on different principles, such as those on the right Table 2.4, can reveal conflicting evidence and ambiguities in possible causal interpretations of the data. Some recent causal discovery algorithms pay close attention to resolving conflicting evidence, e.g., by giving priority to better-supported constraints (*ibid*). Such advances in causal discovery algorithms allow sophisticated automated interpretation of causal evidence from multiple sources. Arguably, they are supporting a beneficial shift in scientific method from formulation and testing of specific hypotheses (to which investigators may become attached) to direct discovery of causal networks from data without first formulating hypotheses, as in *bnlearn* and CAT. At a minimum, causal discovery algorithms can provide computer-aided design (CAD) and discovery tools such as *COMBINE* (Triantafillou and Tsamardinos 2015), *DAGitty* (Textor et al. 2016), and CAT to help investigators synthesize causal network models from multiple data sets, understand their testable implications, and compute adjustment sets and effects estimates for those causal effects that can be estimated.

Specificity, Temporality, Biological Gradient

The consideration of *specificity*—that specific causes (such as long, thin amphibole asbestos fibers) cause specific effects (such as chronic inflammation-mediated malignant mesothelioma)—is no longer widely used, since there are now many examples of agents, such as radiation, bacteria, asbestos, and cigarette smoke, that can cause a variety of adverse health effects through common underlying biological mechanisms. Many agents, including these, induce chronic inflammation (via activation of the NLRP3 inflammasome, as discussed in Chap. 9) together with repetitive injury and scarring in target tissues. This conjunction of conditions can lead to multiple diseases such as fibrosis, asbestosis, COPD, and even lung cancer. However, in cases where only a single exposure and a single effect are of interest, the techniques already studied can be applied to determine whether the exposure variable is the sole parent of the effect in a DAG model. Thus, specificity can be included as a special case of the more general techniques now available for learning causal graph models from data.

Temporality, in the form proposed by Hill—that causes should precede their effects—is too weak a criterion to be very useful in cases such as the study of effects on mortality of a Dublin ban on coal burning (Clancy et al. 2002). The proposed cause, “Reduction in particulate air pollution,” did indeed precede the proposed effect, “Reduction in all-cause mortality,” but it also followed it, because all-cause mortality was on a long-term downward trend that both preceded and followed the ban. The ban had no detectable effect on the decline in total mortality rates over time (Dockery et al. 2013). Thus, it would be a logical fallacy (the *post hoc ergo propter hoc* fallacy) to interpret the decline in mortality following the ban as evidence for the hypothesis that the ban, or the large (about 70%) reduction that it caused in particulate air pollution, caused or contributed to the subsequent decline in mortality rates. The much stronger criterion of predictive causality—that the past and present values of causes should help to predict future values of their effects better than they can be predicted without that information—subsumes and strengthens the traditional temporality criterion.

Finally, biological gradient—that larger effect values should be associated with larger values of their causes—is an unnecessarily restrictive requirement that also overlaps substantially with strength of association. Many biological exposure-response relationships are non-monotonic (e.g., U-shaped, J-shaped, or n-shaped) or have thresholds (e.g., because of positive feedback loops that create bistability, or because of discontinuous changes such as rupture of a lysosome as its membrane loses integrity). Thus, a biological gradient should not necessarily be expected even when there are clear causal mechanisms at work. On the other hand, ignored or mis-modeled errors in exposure estimates can make even sharp threshold-like exposure-response relations appear to follow smooth dose-response gradients if the probability that true exposure is above the threshold that elicits a response increases smoothly with the estimated exposure (Rhomberg et al. 2011). In this case, the apparent gradient between estimated exposures and response probabilities

is actually just evidence of exposure estimation error, not evidence that risk increases with exposures below the threshold. However, the more important point is that a positive biological gradient is a very special case of more general hypotheses, such as the LiNGAM hypothesis that response (or response probability) is a possibly non-monotonic function of exposure plus an error term; or the general non-parametric hypothesis that the conditional probability of response (or its conditional probability distribution, if the response variable has more than two values) depends on exposure. Since modern causal discovery methods work with these more general hypotheses, tests for a biological gradient are unnecessary for causal inference. PDPs and other non-parametric methods will reveal exposure-response gradients if they exist, but can equally easily describe non-monotonic exposure-response relations (including U-shaped or n-shaped ones with zero average gradient).

In summary, specificity and biological gradient are needlessly restrictive. Causality is often present even when neither property is satisfied, and current methods can identify causal graph structures without making use of either consideration. Temporality can be replaced by the criterion of predictive causality, which is more stringent but more useful.

Methods and Examples of Associational Causation: Regression and Relative Risks

In addition to the Hill considerations and related weight-of-evidence criteria for judging evidence of causality, determination of associational causality is supported by a variety of statistical and epidemiological methods for identifying exposure-response associations and for testing the null hypothesis that they can be explained by sampling variability if the statistical modeling assumptions used are correct. In the simplest case where each individual in a study population is either exposed or not exposed to some condition, the difference or ratio of response rates in the exposed and unexposed groups can be used as a basis for quantifying measures of exposure-response association. For example, the *relative risk (RR) ratio*, which is the ratio of response rates in the exposed and unexposed populations, provides a frequently used measure of exposure-response association. Variations allow this ratio (or the closely related odds ratio) to be quantified after matching on other variables, such as age and sex. Techniques for matching, stratification, and estimation of relative risks are covered in all standard epidemiology textbooks. If exposure is represented by an ordered-categorical or continuous variable instead of by a dichotomous classification, then regression models are used to quantify exposure-response associations. It is common practice to treat evidence that the regression coefficient for exposure is significantly greater than zero as if it were evidence that exposure increases the risk of response. As we have emphasized, this is a mistake: it conflates the distinct concepts of associational and manipulative causation. Table 2.5 gives examples from the literature on health effects of fine particulate matter (PM2.5) air pollution

Table 2.5 Association and causation conflated in PM2.5 health effects literature health effects (all emphases added)

Claims	Comments
“We observed statistically significant and robust <i>associations</i> between air pollution and mortality... these results suggest that fine-particulate air pollution, or a more complex pollution mixture associated with fine particulate matter, contributes to excess mortality in certain U.S. cities.” Dockery et al. (1993)	Associations do not suggest a contribution to excess mortality unless they are causal
“The magnitude of the association suggests that controlling fine particle pollution would result in thousands of fewer early deaths per year.” Schwartz et al. (2002)	Associations do not suggest results from changes in exposure concentrations unless the associations represent manipulative causal relations
“We examined the <i>association</i> between PM(2.5) and both all-cause and specific-cause mortality... Our findings describe the magnitude of the <i>effect</i> on all-cause and specific-cause mortality; the modifiers of this association, and suggest that PM(2.5) may pose a <i>public health risk</i> even at or below current ambient levels.” Franklin et al. (2006)	An association with mortality is not an effect on mortality. A C-R association does not suggest that exposure poses a public health risk, unless it is causal
“Residential ambient air pollution exposures were <i>associated</i> with mortality... our study is the first to assess the <i>effects</i> of multiple air pollutants on mortality with fine control for occupation within workers from a single industry.” Hart et al. (2011)	Associations with mortality are not effects on mortality (Petitti 1991)
“Each increase in PM2.5 (10 µg/m ³) was <i>associated</i> with an adjusted increased risk of all-cause mortality (PM2.5 average on previous year of 14%... These results suggest that further <i>public policy efforts that reduce fine particulate matter air pollution are likely to have continuing public health benefits.</i> ” Lepenil et al. (2012)	Associations do not suggest that public policy efforts that reduce exposure are likely to create public health benefits unless the associations reflect manipulative causation
“Ground-level ozone (O ₃) and fine particulate matter (PM2.5) are <i>associated</i> with increased risk of mortality. We quantify the <i>burden</i> of modeled 2005 concentrations of O ₃ and PM2.5 on health in the United States... Among populations aged 65–99, we estimate nearly 1.1 million <i>life years lost from PM2.5 exposure</i> ... The percentage of deaths <i>attributable to PM2.5 and ozone</i> ranges from 3.5% in San Jose to 10% in Los Angeles. These results show that despite significant improvements in air quality in recent decades, recent levels of PM2.5 and ozone still pose a <i>nontrivial risk</i> to public health.” Fann et al. (2012)	In the absence of manipulative causation, statistical associations between pollutant levels and mortality risks do not quantify effects caused by exposure on burden of disease or on life-years lost or on deaths, nor do they indicate a risk to public health

<p>“Ambient fine particulate matter (PM2.5) has a large and well-documented <i>global burden of disease</i>. Our analysis uses high-resolution (10 km, global-coverage) concentration data and cause-specific <i>integrated exposure-response (IER) functions developed for the Global Burden of Disease 2010</i> to assess how regional and global improvements in ambient air quality <i>could reduce attributable mortality from PM2.5</i>. Overall, an aggressive global program of PM2.5 mitigation in line with WHO interim guidelines <i>could avoid</i> 750,000 (23%) of the 3.2 million deaths per year currently (ca. 2010) attributable to ambient PM2.5.” Apie et al. (2015)</p>	<p>The Global Burden of Disease IER functions are based on relative risk measures of association. They do not allow prediction or assessment of “how... improvements on ambient air quality could reduce attributable mortality” or avoid deaths unless the underlying relative risks represent manipulative causal relations</p>
<p>“We use a high-resolution global atmospheric chemistry model combined with <i>epidemiological</i> concentration response <i>functions</i> to investigate <i>premature mortality attributable to PM2.5</i> in adults ≥ 30 years and children <5 years...[A]pplying worldwide the EU annual mean standard of 25 $\mu\text{g}/\text{m}^3$ (3) for PM2.5 could <i>reduce global premature mortality</i> due to PM2.5 exposure by 17%...Our results reflect the need to adopt stricter limits for annual mean PM2.5 levels globally...<i>to substantially reduce premature mortality</i> in most of the world.” Giannadaki et al. (2016)</p>	<p>Epidemiological exposure concentration-response associations and estimates of PM2.5-attributable mortalities based on them do not imply that reducing PM2.5 would reduce mortality, or allow such reductions to be predicted, unless the associations represent manipulative causal relations</p>

in which findings about associations are misinterpreted as implying that reducing exposures would reduce health risks. The left column shows claims from various articles suggesting that association implies manipulative causal conclusions. The right column comments on the confusion, in each case, between association and manipulative causation. In general, the policy-relevant conclusions in the left column do not follow from the associational findings presented.

Example of Associative vs. Manipulative Causation in Practice: The CARET Trial

The practical importance of the distinction between associative and manipulative concepts of causation is well illustrated by the results of the CARET trial, a randomized, double-blind 12-year trial initiated in 1983 that sought to reduce risks of lung cancer by administering a combination of beta carotene and retinol to over 18,000 current and former smokers and asbestos-exposed workers (Omenn et al. 1996). This intervention was firmly based on epidemiological studies showing that relative risks of lung cancer were smaller among people with larger levels of beta carotene and retinol. The effect of the intervention was to increase risk of lung cancer. In the words of the investigators, “The results of the trial are troubling. There was no support for a beneficial effect of beta carotene or vitamin A, in spite of the large advantages inferred from observational epidemiologic comparisons of extreme quintiles or quartiles of dietary intake or serum levels of beta carotene or vitamin A. With 73,135 person-years of follow-up, the active-treatment group had a 28% higher incidence of lung cancer than the placebo group, and the overall mortality rate and the rate of death from cardiovascular causes were higher by 17% and 26%, respectively.” That the intervention produced the opposite of its intended and expected effect is a valuable reminder of the key point overlooked in thousands of published epidemiological studies similar to those in Table 2.5: relative risks and other measures of association do not necessarily or usually predict how response probabilities will change if interventions are used to change exposures. Predicting effects of interventions requires methods such as those illustrated in Figs. 2.24, 2.25, 2.26, 2.27, 2.28, 2.29, 2.30, 2.31, 2.32, and 2.33 for quantifying total causal effects of one variable on another, as well as an assumption that the causal graph models learned from data represent manipulative rather than only predictive causality.

Example of Association Created by Regression Model Specification Error

The regression methods used in associational analyses are supported by free, high-quality software for fitting standard regression models to data. Popular choices include Poisson regression models for dependent variables that are counts (e.g., number of deaths per day in a population); logistic regression models for binary dependent variables (e.g., response or no response for each individual); multiple linear regression or generalized linear models for continuous dependent variables;

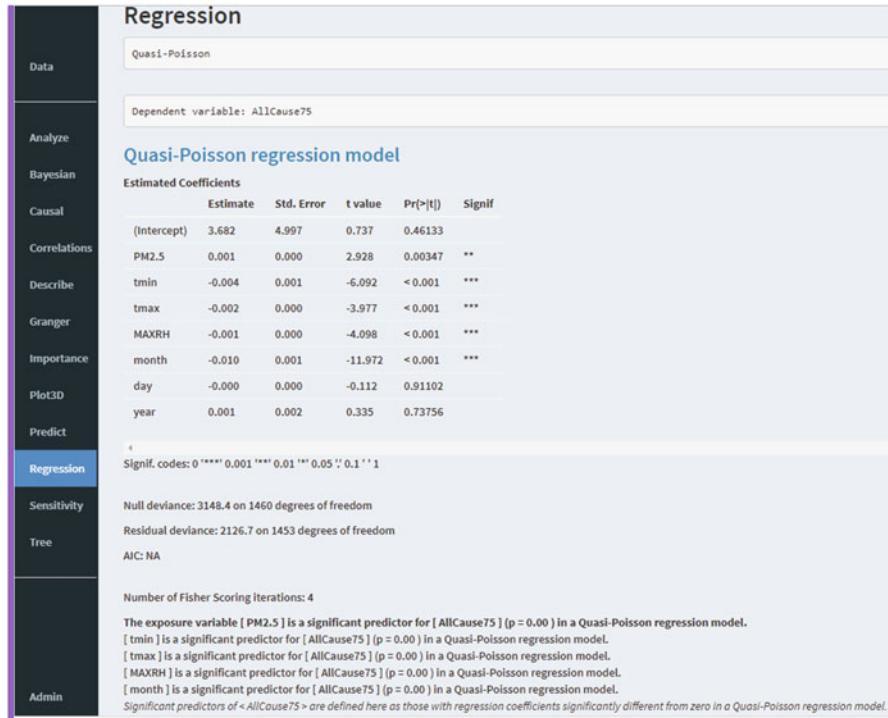


Fig. 2.41 Regression analysis in CAT

and Cox proportional hazards models for survival data. Non-parametric alternatives such as Random Forest and PDP plots can be used instead of parametric regression models, but here we illustrate traditional regression analysis in CAT.

Figure 2.41 shows the top-most results generated by loading the LA data set in Fig. 2.12 and then selecting CAT's *Regression* command. CAT automatically detects that the dependent variable (which by default is the first column, *AllCause75*) is a count variable; fits a Poisson regression model; notes that the Poisson regression modeling assumption of equal means and variances for the dependent variable given the values of the predictors is violated; and therefore fits a more general Quasi-Poisson model. Other appropriate regression models, including linear regression and a Random Forest analysis, are then fit to the same data, and the resulting regression coefficients (for linear regression), confidence intervals for regression coefficients, and diagnostic plots are displayed by the CAT software below the top-most results that are shown in Fig. 2.41. Figure 2.42 shows the results of applying causal discovery algorithms to the same data. In this consensus BN DAG model, all four of the BN learning algorithms used agree that the only two parents of *AllCause75* (at the far right) are *month* and *tmin*: *PM2.5* is not shown as a direct cause, and indeed the DAG structure implies that *AllCause75* is conditionally independent of *PM2.5* given *month*. Why, then, does *PM2.5* appear as a significant predictor of

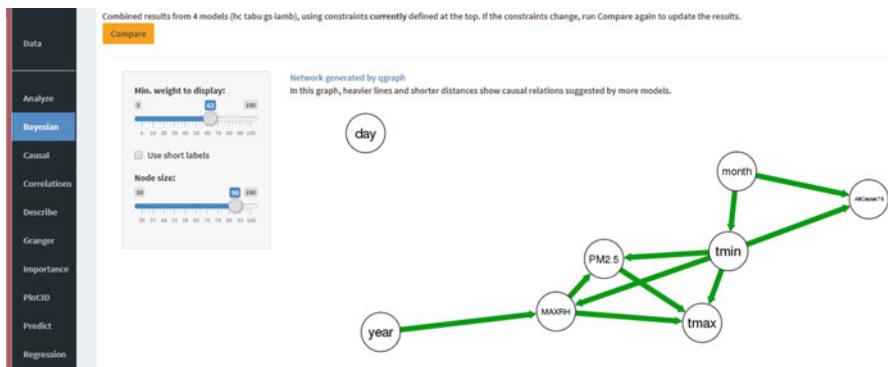


Fig. 2.42 Consensus DAG in CAT

AllCause75 in both quasi-Poisson and linear regression models, even after conditioning on *month* and other variables?

The answer is that *model specification error* makes *PM2.5* informative about *AllCause75* in regression models that treat *month* as a continuous predictor instead of as a discrete categorical predictor. The default in many regression packages is to treat predictors with many values as continuous variables, so that main-effects regression coefficients have interpretations as slope coefficients. But *month* is an unusual variable, in that it cycles through the same 12 values repeatedly. Specifying it as a categorical variable with 12 distinct values (using *as.factor()* in R or by checking the appropriate boxes on the CAT data page) causes *PM2.5* to drop out as a significant predictor of *AllCause75*. But leaving it as a continuous predictor (the default) results in the regression models trying to estimate a single slope coefficient for a variable that cycles. Since *PM2.5* also varies by month, including values of *PM2.5* as predictors for other variables that vary by month, namely *AllCause75*, can help to correct some of the specification error introduced when *month* is misspecified as a continuous predictor. In short, model specification error causes a logically irrelevant variable, *PM2.5*, to become a significant predictor for the dependent variable because it can be used to correct for some of the specification error, thus reducing error variance, i.e., the mean squared error between model-predicted and observed values.

The same phenomenon can be seen more simply in the small DAG model $X \leftarrow Z \rightarrow Y$ with corresponding structural equations $Y = Z^3$ and $X = Z^2$. In this DAG, Y is clearly conditionally independent of X given Z , but if a linear regression model of the form $E(Y | X, Z) = a + b*X + c*Z$ is fit to a large data set (e.g., with Z values uniformly distributed between 0 and 1 and with $Y = Z^3$ and $X = Z^2$), then the regression coefficient for X will be positive. This is not because X contributes any information for predicting Y that was not already available from Z , but simply because the assumed linear model form is misspecified, so that including $X = Z^2$ helps to reduce the mean squared error in predicting $Y = Z^3$ using a linear model. If a non-parametric method such as Random Forest were used, then parametric model

specification error would no longer play this decisive role, and X would no longer appear as a predictor for Y after conditioning on Z . Interpreting X as exposure, Y as response, and Z as lifestyle, this example shows how model specification error can create a significant statistical exposure-response regression coefficient—and hence an apparent associate between X and Y in this example, even after Z has been “controlled for” (or conditioned on by including it on the right side of the misspecified regression model)—even if exposure is not a cause of response. The appearance of $PM2.5$ as a highly significant predictor in the regression model for $AllCause75$ in Fig. 2.41 but not as a parent of $AllCause75$ in Fig. 2.42 reflects a similar instance of association without causation.

A very similar point holds when errors in measured or estimated values of predictors are ignored or are not well modeled. For example, suppose that all three of the variables in the DAG model $X \leftarrow Z \rightarrow Y$ are measured with error, but that the measurement error is much larger for Z than for X . Then in a typical linear regression model that omits error terms for the values of predictors X and Z , the measured values of X may be more strongly associated with the measured values of Y than are the measured values of Z , even though Z and not X is the cause of Y . Again, strength of association does not necessarily indicate likelihood of causation.

Example: Associative Causation in Air Pollution Health Effects Research

Di et al. (2017) note that “In the US Medicare population from 2000 to 2012, short-term exposures to PM2.5 and warm-season ozone were significantly associated with increased risk of mortality. This risk occurred at levels below current national air quality standards, suggesting that these standards may need to be reevaluated.” But, as in other associational studies (Table 2.5), the reported associations lack clear implications for prudent risk management or regulatory actions. Studies of association are (or should be) of limited interest to policy analysts and decision-makers insofar as they fail to address the following key manipulative causal question needed to inform effective decision-making:

- *Q1: How would public health effects change if exposure concentrations were reduced?*

Instead, studies of association address the following easier, but less relevant, question:

- *Q2: What are the estimated ratios (or slope factors or regression coefficients) of health effects to past pollution levels in various researcher-selected models and data sets?*

As indicated by the sample of literature in Table 2.5, hundreds of scientific articles and accompanying press releases and editorials on air pollution health effects research have presented answers to Q2 as if they were answers to the Q1. Ambiguous language, such as that scientific studies “link” mortalities or morbidities to air pollution levels (often meaning little more than that someone divided one by the

other, or regressed one against the other) has obscured the fact that Q2 has been substituted for Q1. Policy-makers need, but lack, trustworthy, data-driven answers to Q1. To recapitulate, answers to Q2 are inadequate substitutes for answers to Q1 for all of the following reasons.

- “*Associations are not effects*” (Petitti 1991). A strong, positive, no-threshold exposure-response association in a population warrants no conclusions about how changing exposure would change response (Pearl 2009). Observing that, historically, $Y = 10X + 50$, where X measures exposure and Y measures response, does not preclude the possibility that increasing X would reduce Y , or leave it unchanged. For example, suppose that the structural (causal) relation is $Y = Z - X$, meaning that exogenous changes in X or Z cause Y to adjust until $Y = Z - X$, where Z is some covariate such as age or poverty. Suppose that, historically, the associative equation $Z = 11X + 50$ has held, perhaps because poor people or older people live disproportionately in high-exposure areas. The reduced-form regression equation describing historical observational data, $Y = 10X + 50$, reveals nothing about how an exogenous reduction in X alone, holding other factors such as Z fixed, would change Y (in this case, increasing Y by one unit per unit reduction in X).
- *Most published associations are assumption-dependent and model-dependent.* That is, they depend on specific modeling assumptions used in producing them. In the example just given, regressing Y against X alone would yield a positive association (regression coefficient of 10) for X : $E(Y | X) = 10X + 50$. Regressing Y against both X and Z would yield a negative association (regression coefficient of -1) for X : $E(Y | X) = Z - X$. Which association, positive or negative, is reported depends on the model selected. Headlines of the form “Study links exposure X to increased risk of Y ” would often be more accurately expressed as “Researchers select a model with a positive association between X and Y .”
- *Historical associations do not predict future effects of interventions.* The Dublin coal-burning ban experience illustrates this point. Associations in the observational data (Clancy et al. 2002) did not correctly predict the lack of effect caused by the large reduction in air pollution (Dockery et al. 2013).
- *Omitted confounders such as lagged daily temperatures create spurious (non-causal) exposure-response associations.* Di et al. omit lagged values of daily temperature for days 2, 3 and more. Yet, Fig. 2.20 shows that lagged daily minimum temperatures out to 7 days are among the most important predictors of daily elderly mortality counts in that data set. Omitting them creates significant positive regression coefficients for PM2.5 as a predictor of daily mortality because the PM2.5 levels are affected by the lagged temperatures and act as a partial surrogate for them if they are omitted.
- *Model specification errors create spurious associations.* Instead of presenting an ensemble of multiple alternative plausible models, Di et al. relied on a conditional logistic regression model. Other models might well have produced different results. In the LA data set, fitting a quasi-Poisson or linear regression model (by clicking on “Regression” in the CAT software) to same-day values of the

variables produces a significant positive regression coefficient for PM2.5 as a predictor of daily mortality, but non-parametric analyses (CART trees, random forest ensembles, and Bayesian networks) show no relation between them. The explanation is that generalized linear models are misspecified for this data set. PM2.5 is informative for predicting mortality in the context of the misspecified parametric regression model because it can be used to partly correct the specification error. It has no predictive value in a correctly specified model.

- *Association is not manipulative causation.* Showing that exposure and mortality rates are associated does not imply that changing exposure would change mortality rates.

Relative Risk and Probability of Causation in the Competing Risks Framework

Despite its inadequacies as a general guide for identifying manipulative causal relationships in data, there are very specific models in which association can successfully play this useful role. One of the best known of these is the *competing risk* framework, in which each of several sources (potential causes of a disease or adverse outcome) is thought of as independently generating “hits” on a target at a random rate with an average intensity expressed in units of expected hits per unit time. The first hit on the target causes the adverse outcome, such as cancer or birth defect. In this specific setting, with the multiple sources “competing” to land the first hit on the target, the probability that each source wins, thereby becoming the cause of the adverse outcome, is the ratio of its intensity to the sum of the intensities from all sources, $\lambda_i/(\lambda_1 + \lambda_2 + \dots + \lambda_N)$ where λ_i denotes the expected hits per unit time (intensity) from source i . This ratio is the *probability of causation* that source i is the cause of the adverse outcome, given that it occurs. It can also be written as $PC_i = \lambda_i/(\lambda_B + \lambda_i)$, where λ_B denotes the background intensity for occurrence of the adverse effect in the absence of exposure to source i , i.e., the sum of the intensities from all other sources. Doubling the intensity of hits from a source approximately doubles its probability of causation if its intensity is small compared to the background intensity. Conversely, for it to be more likely than not that source i is the cause of the adverse outcome, given that the outcome has occurred, it must be the case that $\lambda_i/(\lambda_B + \lambda_i) > \frac{1}{2}$, implying that $\lambda_i > \lambda_B$, and hence that the relative risk ratio $RR = (\lambda_B + \lambda_i)/\lambda_B = 1 + \lambda_i/\lambda_B$ for expected occurrences per person-year among people exposed to source i compared to otherwise similar people not exposed to it, must exceed 2. This criterion is sometimes discussed in the context of legal evidence. For this special case of competing risks, the relative risk ratio $RR = 1 + \lambda_i/\lambda_B$ is a linear function of the hit intensity λ_i and hence it directly reflects the incremental risk caused by exposure to source i . Indeed, since $RR = (\lambda_B + \lambda_i)/\lambda_B$ and $PC_i = \lambda_i/(\lambda_B + \lambda_i)$, their product is $RR * PC_i = \lambda_i/\lambda_B = RR - 1$, from which it follows that $PC_i = (RR - 1)/RR$, or $PC_i = (1 - 1/RR)$ when this is positive, i.e. when $RR > 1$. Probability of

causation is an increasing function of relative risk, ranging from 0 when $RR = 1$ to 1 as RR approaches infinity. Thus, in this case, association as measured by relative risk is an excellent guide to causation, as measured by probability of causation: the greater is the association, the higher is the probability of causation. In the competing risks framework, Hill's original intuition that stronger associations make causation more likely is well justified. However, the competing risks framework makes very restrictive assumptions, especially that causes act independently of each other rather than interacting and that each cause by itself fully suffices to cause the adverse outcome, so that the first-hit metaphor applies. When these assumptions do not hold, there is no longer any necessary relation between association and causation, as illustrated in previous examples, and probability of causation for a single source is no longer well defined.

Example: Calculating Probabilities of Causation

Setting: Mr. Smith, a heavy smoker with a family history of lung cancer, was diagnosed with emphysema in 1990 but continued to smoke until his death in 2015. Between 1995 and 2005, he worked in a rock quarry, where he had an unknown exposure to quartz sand and dust, i.e., respirable crystalline silica (RCS). He wore a respirator at least part of the time, but its efficacy is unknown. In 2015, he was diagnosed with lung cancer (an adenocarcinoma), and died from it later that year. His estate sues the respirator manufacturer, alleging Mr. Smith developed lung cancer because of his RCS exposure and that the respirators he used failed to adequately protect him from the carcinogenic hazard posed by RCS. In support of this claim, they note that the International Agency for Research on Cancer (IARC) has identified RCS as a human carcinogen. Assume that (a) In the absence of risk factors, the age-specific hazard rate for lung cancer for men of Mr. Smith's age in 1990 is about $b = 0.01$ expected new cases per person-year in men who do not yet have lung cancer (Spitz et al. 2007). (For simplicity, we will treat this background rate as approximately constant over the time interval in question. In reality, it varies some with age, but this complication does not change the key ideas for calculating probability of causation or the order of magnitude of the results). (b) A medical history of emphysema increases risk of lung cancer in men who continue to smoke about threefold (Spitz et al. 2007). (c) A family history of lung cancer increases risk of lung cancer about fourfold (Wu et al. 1988). (d) RCS exposure increases risk of lung cancer among heavily exposed men by no more than 1.5-fold (Gamble 2011). (e) These relative risks combine approximately multiplicatively (Spitz et al. 2007). Specifically, if risk factors other than RCS exposure increase Mr. Smith's age-specific hazard rate for lung cancer from b to Rb expected cases per year (where $R > 1$), and if RCS exposure increases his lung cancer risk by a further factor of S , which we estimate as not more than 1.5, then the total risk of lung cancer from these joint exposures is increased to RSb expected cases per person-year. If $R = 3*4 = 12$ for a continuing smoker with emphysema and family history of lung cancer, and if $S = 1.5$ for RCS and $b = 0.01$, then $RSb = 0.18$ cases/year.

Problem: What is the probability that Mr. Smith's RCS exposure caused his lung cancer, meaning that he would not have developed lung cancer by 2015 had it not been for his RCS exposure? Assume that the formula to be used is as follows (www.cdc.gov/niosh/ocas/pccalc.html):

$$PC = \text{exposure risk} / (\text{baseline risk} + \text{exposure risk})$$

This is analogous to the $\lambda_i / (\lambda_B + \lambda_i)$ formula for competing risks, although the resulting PC value may lack a valid interpretation as a probability if competing risk assumptions do not hold. It is then better thought of as a generalized probability of causation, or as an assigned share in causation. For further discussion of foundational issues in the interpretation of such formulas, see Cox (1984, 1987) and Greenland (2015).

Analysis and Model: In the PC formula, the “*baseline risk*” is the absolute probability of lung cancer in the absence of exposure to RCS in a population of similar people (male smokers with Mr. Smith’s causal risk factors, other than RCS). It is calculated as follows. The probability of surviving for T years without lung cancer when the hazard rate is Rb cases per year is given by the standard survival function, $1 - \exp(-RbT)$:

$$\begin{aligned} \text{baseline risk} &= \text{probability of developing lung cancer within} \\ &\quad T \text{ years exposure if no RCS} = 1 - \exp(-RbT) \end{aligned}$$

Similarly, the “*exposure risk*” is the difference between the total probability of lung cancer when RCS exposure is present and when it is absent, holding the levels of all other causal risk factors fixed. That is, *exposure risk* is the increase in probability of lung cancer caused by Mr. Smith’s exposure to RCS, given the other risk factors to which he was exposed:

$$\begin{aligned} \text{exposure risk} &= 1 - \exp(-RSbT) - (1 - \exp(-RbT)) \\ &= \exp(-RbT) - \exp(-SRbT) \end{aligned}$$

The PC ratio is thus

$$\begin{aligned} PC &= \text{exposure risk} / (\text{baseline risk} + \text{exposure risk}) \\ &= (\exp(-RbT) - \exp(-SRbT)) / (1 - \exp(-RbT) + \exp(-RbT) - \exp(-SRbT)) \\ &= (\exp(-RbT) - \exp(-SRbT)) / (1 - \exp(-SRbT)) \end{aligned}$$

Solution: Using $b = 0.01$, $R = 12$, $S = 1.5$, and $T = 25$ years (from 1990, when emphysema was diagnosed, to death in 2015), the probability of causation for RCS is

$$PC = (\exp(-R^*b^*T) - \exp(-S^*R^*b^*T)) / (1 - \exp(-S^*R^*b^*T)) = 0.01.$$

Thus, under these conditions, there is about a 1% probability that Mr. Smith’s lung cancer would not have occurred had he not been exposed to RCS, i.e., “but for” his RCS exposure. If the assumed parameter values are varied, then this probability increases with S to a maximum value of $\exp(-R^*b^*T) = 0.05$ and decreases toward 0 as RbT increases.

Conclusions on Associational Causation

Hill's essay (1965) formulated a question of great practical importance: "But with the aims of occupational, and almost synonymous preventive, medicine in mind the decisive question is where the frequency of the undesirable event B will be influenced by a change in the environmental feature A." This is a question about manipulative causation: how would changing exposure (or "environmental feature") A affect the frequency distribution or probability distribution in the exposed population, of undesirable event B? It overlaps with the questions addressed by modern causal discovery algorithms that quantify total and direct causal effects on a response variable of changes in an exposure variable by using tools such as *DAGitty* to determine what effects can be estimated (and what adjustment sets of other variables must be conditioned on to do so), and algorithms such as Random Forest to estimate them without making parametric modeling assumptions. However, rather than focusing on how to obtain valid scientific answers to this key question, Hill reformulated it as follows: "Disregarding then any such problem in semantics we have this situation. Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?" This is a very different question. It is no longer about how to discover how changing one variable will change another. Rather, it is about what to consider before *deciding* that the *most likely* explanation or interpretation for an observed *association* between two variables is "causation" (without definition or explanation of what that label means, i.e., its semantics). This is a much less interesting question. Even the most likely explanation is quite likely to be wrong when there are many competing plausible ones. Making a decision about how to label "an association between two variables" is less useful for effective decision-making than figuring out how changing one variable would change the other. The crucial idea of manipulative causation has disappeared. Moreover, the new question of what to consider before "deciding that the most likely interpretation of it [the observed association] is causation" imposes a false dichotomy: that an association is either causal or not, rather than some fraction of it is causal and the rest not.

The answers that Hill proposes to the revised question—the nine considerations in the left column of Table 2.4—are not thorough or convincing insofar as they fail to consider a variety of important possible non-causal explanations and interpretations for some or all of an observed association. Table 2.6 lists those we have discussed, with brief descriptions and notes on some of the main techniques for overcoming them. Hill himself did not consider that his considerations solved the scientific challenge of causal discovery of manipulative causal relationships from data, but offered them more as a psychological aid for helping people to make up their minds about what judgments to form: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*. What they can do, with greater or less strength, is to help us to make up

our minds on the fundamental question—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?”

Since Hill deliberately avoided specifying what he means by “cause and effect” in this context (instead “disregarding then any such problem in semantics”), his considerations must serve as an implicit definition: “cause and effect” in this context is a label that some people feel comfortable attaching to observed associations after reflecting on the considerations in Table 2.4. Other possible non-causal explanations for observed associations that might disconfirm the causal interpretation, such as those in Table 2.6, are not included among the considerations. Deciding to label an association as “causal” based on the Hill considerations does not imply that a causal interpretation is likely to be correct or that other non-causal explanations are unlikely to be correct. Indeed, by formulating the problem as “is there any other answer equally, or more, likely than cause and effect?” Hill allows labeling an association as causal even if it almost certainly isn’t. Suppose that each of the eight alternative explanations in Table 2.6 is judged to be the correct explanation for an observed association with probability 11% and that “cause and effect” is judged to be the correct explanation with probability 12%. (For simplicity, let these be treated as mutually exclusive and collectively exhaustive possible explanations, although of course they are neither.) Then “cause and effect” would be the most likely explanation, even though it has only a 12% probability of being correct, and non-causal explanations have an 88% probability of being correct. That would satisfy Hill’s criterion that there is not “any other answer equally, or more, likely than cause and effect,” even though cause and effect is unlikely to be the correct explanation. Deciding to label an association as “causal” in the associational sense used by Hill, IARC (2006), and many others does not require or imply that the associations so labeled have any specific real-world properties, such as that changing one variable would change the probability distribution of another. It carries no implications for consequences of manipulations or for decisions needed to achieve a desired change in outcome probabilities.

Given these limitations, associational methods are usually not suitable for discovering or quantifying manipulative causal relationships. Hence, they are usually not suitable for supporting policy recommendations and decisions that require understanding how alternative actions change outcome probabilities. (An exception, as previously discussed, is that if a competing-risk model applies, then associational methods are justified: effects of interventions that change the intensities of hits from one or more sources change relative risks, cause-specific probabilities of causation, and absolute risk of a hit per unit time in straight-forward ways.) Associations can be useful for identifying non-random patterns that further investigation may explain, with model specification errors, omitted variables, confounding, biases, coincident trends, other threats to internal validity, and manipulative causation being among the *a priori* explanations that might be considered. That associational studies are nonetheless widely interpreted as if they had direct manipulative causal implications for policy (Table 2.5) indicates a need and opportunity to improve current practice.

Table 2.6 Non-causal explanations for observed associations, and methods to overcome them

Source of non-causal association	Methods for overcoming non-causal associations
<i>Unobserved (latent) confounders</i>	These can be tested for and their effects modeled using the <i>Tetrad</i> , <i>Invariant Causal Prediction</i> , and <i>BACKSHIFT</i> algorithms, among others
<i>Spurious regression</i> in time series or spatial observations with trends	Spurious regression arising from coincident trends can be detected and avoided by using conditional independence tests and predictive causation (e.g., Granger causality) instead of regression models
<i>Collider bias; stratification or selection bias</i>	A study that stratifies or matches individuals on certain variables, such as membership in an occupation, or an analysis that conditions on certain variables by including them on the right-hand side of a regression model, can induce exposure-response associations if the variables conditioned, matched, or stratified on are common descendants of the exposure and response variables. The association does not indicate causality between exposure and response, but that they provide alternative explanations of an observed value. Such biases can be avoided by using <i>DAGitty</i> to compute adjustment sets and conditioning only on variables in an adjustment set
<i>Other threats to internal validity</i>	Threats to internal validity (e.g., regression to the mean) were enumerated by Campbell and Stanley (1963), who also discuss ways to refute them as plausible explanations, when possible, using observational data
<i>Model specification errors</i>	Model specification errors arise when an analysis assumes a particular parametric modeling form that does not accurately describe the data-generating process. Assuming a linear regression model when there are nonlinear effects present is one example; omitting high-order interactions terms is another. Model specification errors can be avoided by using non-parametric model ensemble methods such as PDPs
<i>P-hacking</i> , i.e., adjusting modeling assumptions to produce an association (e.g., a statistically significantly positive regression coefficient)	Automated modeling using CAT or packages such as <i>randomForest</i> and <i>bnlearn</i> to automate modeling choices such as which predictors to select, how to code them (i.e., aggregate their values into ranges), and which high-order interactions to include can help to avoid p-hacking biases

(continued)

Table 2.6 (continued)

Source of non-causal association	Methods for overcoming non-causal associations
<i>Omitted errors in explanatory variables</i>	Using job exposure matrices, remote-sensing and satellite imagery for pollutant concentration estimation, or other error-prone techniques for estimating exposures, creates exposure estimates for individuals that can differ substantially from their true exposures. In simple regression models, omitting errors from the estimated values of explanatory variables tends to bias regression coefficients toward the null (i.e., 0), but the bias can be in either direction in multivariate models, and failing to carefully model errors in explanatory variables can create false-positive associations
<i>Omitted interdependencies among explanatory variables</i>	Direct and total effects of exposure on response can have opposite signs. More generally, the DAG model in which variables are embedded can create associations without causation in a regression model that includes on its right-hand side variables not in an adjustment set. This can be avoided by using <i>DAGitty</i> to compute adjustment sets for the total causal effect of exposure on response and then to condition on variables in an adjustment set to estimate that effect

Studies that explicitly acknowledge that statistical analyses of associations are useless for revealing how policy changes affect outcome probabilities are relatively rare. One exception is a National Research Council report on *Deterrence and the Death Penalty* that “assesses whether the available evidence provides a scientific basis for answering questions of if and how the death penalty affects homicide rates.” This report “concludes that research to date on the effect of capital punishment on homicide rates is not useful in determining whether the death penalty increases, decreases, or has no effect on these rates” (National Research Council 2012). Such candid acknowledgements of the limitations of large bodies of existing research and discussion clear the way for more useful future research. In public health risk research, fully accepting and internalizing the familiar warnings that correlation is not causation, that associations are not effects (Petitti 1991), and that observations are not actions (Pearl 2009) may help to shift practice away from relying on associational considerations such as the Hill considerations on the left side of Table 2.4 toward fuller use of causal discovery principles and algorithms such as those on the right side of Table 2.4. Doing so can potentially transform the theory and practice of public health research by giving more trustworthy answers to causal questions such as how changing exposures would change health effects.

Comparison of Causal Discovery to Attributive Causal Methods

In addition to associative causation, two other causal concepts with a major impact on modern epidemiology, social statistics, and policy evaluation studies are *attributive causation* and *counterfactual* (or *potential outcomes*) *causation*. Insofar as these share with associational causation the limitation that they do not provide valid information about how outcome probabilities would change if different actions or interventions were taken, they are not suitable for informing policy decisions about which actions to take. Like associational analyses, they are nonetheless widely used for this purpose. This section extends points already discussed for associational causation to these other two concepts.

Attributive causation is based on the key idea of attributing observed differences in response rates between exposed and unexposed groups, or between more-exposed and less-exposed groups, to the differences in their exposures. Standard measures used in epidemiology that are based on attributing differences in responses to differences in exposures include the following (www.med.uottawa.ca/sim/data/PAR_e.htm):

- The *attributable risk* (AR) of a disease or other adverse outcome among exposed individuals is the difference in incidence rates between exposed and unexposed individuals, measured in units such as cases per person-year, or cases per 1000, per 10,000 or per 100,000 person-years, in the population. AR is commonly interpreted as the excess incidence rate of disease *caused by* exposure among exposed individuals. This causal interpretation is essentially given the status of a definition in many epidemiology textbooks. Likewise, the *attributable number* (AN) of cases per year in a population is the attributable risk multiplied by the number of people exposed. It is commonly interpreted as the number of extra cases per year *caused by* exposure, again without further inquiry into whether exposure actually does cause the cases.
- The *population attributable risk* (PAR) or *population attributable fraction* (PAF) is derived from the relative risk (RR), i.e., the ratio of disease rates in the exposed and unexposed populations, together with the prevalence of exposure in the population, P , i.e., the fraction of the population that is exposed. It can be expressed via the formula

$$PAR = P(RR - 1)/[1 + P(RR - 1)],$$

It is commonly (mis)interpreted in terms of manipulative causation, as the fraction of cases that *would be prevented* if exposure were removed.

- *Burden of disease* (BoD) calculations extend the PAR formula to allow for multiple possible levels of exposure, each with its own relative risk ratio and prevalence in the population. These methods for BoD calculations have been published and applied by the World Health Organization (WHO), which

interprets them as if they indicated manipulative causation, providing illustrative calculations of how “if the risk factor were to be completely removed. . . the BoD reduction can be calculated from a simplified form of the above formula” involving prevalence rates and relative risks (Prüss-Üstün et al. 2003).

An intuitive motivation for these attributive causal concepts and formulas (and closely related or synonymous ones, e.g., etiologic fractions, population attributable fractions, probability of causation) is an assumption that *observed differences in effects (responses) are explained by observed differences in causes (exposures)*. We will call this the *attribution assumption*. But it is not necessarily true. Differences in effects between exposed and unexposed individuals might instead have other explanations, such as some or all of those in Table 2.6. By contrast, the main intuitive motivation for the information-based methods of causal discovery emphasized in this chapter (right side of Table 2.4) is the *information principle* stating that *values of causes (exposures) help to predict values of their direct effects (responses)* in a DAG model; and that in time series data, changes in the values of causes help to predict changes in the values of their effect. In many settings, this is a much better proxy than the attribution principle for manipulative causation—the principle that *changes in causes change the probability distributions of their direct effects*.

Example: Attributive Causation Is Not Manipulative Causation

Suppose that teen pregnancy rates are found to be higher among young women in a certain population who are enrolled in a school lunch program than among young women who are not. Solely for purposes of a simple calculation (i.e., the example is intended to be simple rather than realistic), suppose the following characteristics always occur together (i.e., a young woman who has one has all); and that all teen pregnancies in this population occur among young women with these characteristics and not among young women who do not. The characteristics are: (a) Enrolled in school lunch program; (b) Belong to school chess club; (c) Enrolled in a Latin honors class; (d) Smokes cigarettes; (e) Drinks Coke instead of Pepsi. Then the traditional textbook attributable risk and BoD formulas just described would attribute 100% of teen pregnancies in this population to the school lunch program. They would also attribute 100% of the teen pregnancies in this population to each of the other characteristics that cluster with the enrollment in the school lunch program, i.e., to membership in the chess club, the Latin honors class, cigarette smoking, and drinking Coke instead of Pepsi. Each of these factors would have a probability of causation of 100% for causing teen pregnancy. This illustrates the distinction between the meaning of causation as defined by the World Health Organization and other authorities based on relative risk ratios and attribution formulas, and the usual meaning of causation. Following current practice, these attributable risk calculations could be used to inform policy makers that requiring young women to drink Pepsi instead of Coke would prevent teen pregnancies in this population;

likewise, cancelling the school lunch program or the chess club or the Latin honors class or ending cigarette smoking would each be predicted to prevent 100% of teen pregnancies in this population. This illustrates the difference between policy recommendations based on attributive measures of causation and policy recommendations based on manipulative or mechanistic causation, which would recognize that the attributable risk calculations have no implications for how or whether any of these possible interventions would change teen pregnancy rates in the population. Attributable risk calculations based on relative risks correspond to using only the first of the Hill considerations, strength of association, without considering others such as biological plausibility or coherence with knowledge of how outcomes are caused.

Example: Nine Million Deaths per Year Worldwide Attributed to Air Pollution

A more realistic example of the same methodological points arises from epidemiology. In October of 2017, international headlines announced that “Pollution kills 9 million people each year.” Accompanying articles warned that it “costs trillions of dollars every year” and “threatens the continuing survival of human societies.” Underlying these sensational announcements was a Burden-of-Disease (BoD) study published in the *Lancet* and communicated to the press via carefully designed infographics, jointly published with the advocacy network Global Alliance on Health and Pollution (GAHP), urging more funding for pollution research, advocating clean energy, and linking air pollution to climate change ([Lancet 2017](#)). The technical basis for the study was the uncontroversial observation that mortality rates are higher in some countries and regions than in others. Higher mortality rates tend to cluster with conditions such as lower GDP per capita, higher poverty, poorer nutrition, lower education, higher illness rates, higher pollution levels (e.g., from use of indoor dung-burning stoves), lower savings rates, and poorer healthcare. The BoD study attributed the higher mortalities observed in such circumstances to higher pollution; it could equally well have attributed them to any other factor in the cluster, much as in the teen pregnancy example. Such attributive causation does not imply that reducing pollution without addressing the other factors that cluster with it would have any effect on deaths per year.

Indeed, since the number of deaths per year is the number of live births per year one lifetime ago, exposure to pollution cannot permanently increase average deaths per year by nine million (or any other amount) unless it increases previous birth rates correspondingly—an effect not usually ascribed to pollution. What exposure might do is to change life lengths. But if life length is likened to a pipeline, with the number of people exiting each year (dying) equal to the number entering it (being born) one life length ago, then it is clear that changing the length of the pipeline does not permanently change the number of people exiting from it per year. (If pollution were to cause everyone to die a year earlier than they otherwise would have, for example,

then the excess deaths that occur this year instead of next year would be offset by the fewer deaths that would have occurred this year but that occurred last year instead.) Headlines reflecting attributive causation, such as “Pollution kills 9 million people each year,” should not be misinterpreted as implying that there would be any fewer total deaths per year if pollution were eliminated. Nor should claims that reducing pollution has generated trillions of dollars per year in public health benefits be misunderstood as implying that those same benefits would not have occurred without the reduction in pollution. Attribution-based causal claims are akin to accounting decisions—to which factor or factors do we choose to assign responsibility for outcomes?—rather than facts about the world.

As illustrated in these examples, attributive causation and BoD calculations attribute differences in observed response rates between exposed and unexposed individuals to the observed differences in their exposures (possibly after matching on covariates such as demographic characteristics), whether or not changing exposures would change response rates. The attribution assumption can be justified in cases where the causal agent that produces responses is known, unique, and can be measured, as for many food borne diseases: risk of Salmonellosis depends on how much *Salmonella* is ingested and can be attributed to the source(s) of that exposure. Chapters 5–7 examine this approach in more detail. But in general, the attribution assumption may fail when unique causal agents do not exist, or are unknown or unmeasured.

Comparison of Causal Discovery to Counterfactual Causal Methods

Enough dissatisfaction has accumulated with the attributive approach to inspire development of an alternative based on counterfactual comparisons of observed to unobserved outcomes. The intuitive motivation is the following *counterfactual causation principle*: *Differences between observed response rates under real conditions and model-predicted response rates under hypothetical (counterfactual) conditions are attributed to the differences between real and modeled conditions.* Thus, differences in causes (exposures) are assumed to create differences in their effects (responses) compared to what they otherwise would have been. The most common application of this principle is as follows:

1. Use a regression model to predict what the response rate among exposed people would have been had conditions been different, e.g., had they not been exposed, or had they been exposed to lower levels, of a hazard;
2. Compare this counterfactual response rate to the observed response rate under actual exposure conditions; and
3. Attribute the difference to the difference between modeled exposure conditions (e.g., no exposure) and actual exposure conditions.

Such counterfactual modeling can also be used to address what-if questions that manipulative causation cannot answer, such as “How would my salary today be different if my race, age, or sex were different?” Assuming that the regression model used describes what would have happened under other conditions allows modelers to answer questions about the effects of counterfactual conditions that cannot be achieved (and that may have no clear meaning) in the real world.

A well-known fundamental challenge for such *counterfactual* or *potential outcomes modeling* is that what would have been under counterfactual conditions is never observed. Thus, to apply this approach, what would have happened—e.g., the mortality rate among exposed people under different exposure conditions—must be guessed at or assumed. Without a well-validated causal model, guessing with high accuracy and confidence what would have happened under different conditions may be difficult or impossible. Thus, the counterfactual modeling approach requires having a well-validated causal model before it can give trustworthy estimates of the counterfactual responses on which its estimates of causal impacts of exposure depend. In essence, it simply begs the question of how to obtain such a model.

A recent technical advance, *doubly robust estimation* of average causal effects in populations, combines two potential outcomes methods, outcome regression modeling and propensity score modeling (Funk et al. 2011). It requires only one of the two models to be correctly specified in order to yield unbiased effects estimates. However, the problem of correctly specifying even one model remains. Even state-of-the-art potential outcomes algorithms such as targeted maximum likelihood estimation (TMLE), which have the double robustness property, still have substantial error rates and problematic performance in practice in many real data sets (Pang et al. 2016).

The example of the Dublin coal-burning ban discussed earlier illustrates some of the limitations of the counterfactual approach. The original study (Clancy et al. 2002) compared observed mortality rates following the ban to mortality rates in the same population before the ban. It attributed the quite significant difference between them to the effects of the ban. The authors concluded that the ban had caused a sizable reduction in mortality rates, based on the counterfactual modeling assumption that, in the absence of the ban, mortality rates would have remained at their original levels. By contrast, the follow-up investigation a decade later (Dockery et al. 2013) compared observed changes in mortality rates from before to after the ban in the affected population to changes in the mortality rates over the same period among people living outside the affected area. This time, the authors concluded that the ban had caused no detectable effect in all-cause mortality, based on the alternative counterfactual modeling assumption that, in the absence of the ban, mortality rates would have fallen by as much in the population inside the area affected by the ban as they fell in the population outside it. Changing counterfactual assumptions about what would have happened to mortality rates in the absence of a ban completely changed the conclusions about the effects caused by the ban.

Many technical developments over the past four decades have been introduced to try to make the counterfactual approach logically sound and useful to practitioners, as discussed further in Chap. 14. Innovations include variations on the theme of

stratifying or matching individuals in the exposed and unexposed groups on observed covariate values in hopes of making it more likely that differences in responses between the matched groups are caused by differences in exposure; and various types of conditioning on observed covariate values using regression models to play a role similar to matching or stratification in “adjusting” or “controlling” for non-exposure differences between the groups. However, none of these efforts overcomes the basic problem that counterfactual responses are speculative. Unless a valid causal model is obtained by other means, or randomized experiments are carried out, as in the case of the CARET trial, what the correct counterfactual values for responses are remains unknown.

Unfortunately, counterfactual modeling efforts have introduced many confusions and mistakes in attempted causal analyses. For example, it has gradually been understood that attempting to “control” for the effects of covariates by matching or stratifying on their values or conditioning on them in a regression model may create an artificial exposure-response association due to collider bias. Model specification errors in potential outcome regression models are almost certain to lead traditional potential outcomes methods to false-positive results, mistakenly rejecting the null hypothesis of no effect even when it is true, if sample sizes are large enough (the “g-null” paradox). Even if these problems could be resolved, the question that counterfactual causal analysis methods usually attempts to answer based on data and assumptions is how outcomes *would have been* different if exposures had been different (or if intervention or other conditions had been different) in specified ways. But this is not the same as the question that decision makers and policy analysts need answered, which is how future outcome probabilities *will be* different if different interventions are undertaken now. What responses would have been if different interventions had been made in the past is in general not the same as what they will be if those interventions are undertaken now, unless the answers are calculated using invariant and stationary causal laws (e.g., causal CPTs). But this again requires a valid causal model.

Counterfactual and potential outcome methods do not solve the challenge of developing valid causal models from data, but they can use valid causal models, if they are available, to estimate causal impacts of interventions and to answer what-if questions about how average responses would have differed if exposures had been different. Unfortunately, misspecified models can also be used to answer the same questions, but the answers may be wrong, and the errors they contain may be unknown because they are based in part on conjectures about what would have been instead of being based entirely on observed data. For practical purposes, therefore, we recommend using methods and principles of causal discovery such as those on the right side of Table 2.4 to learn valid causal models from data. Counterfactual analyses and interpretations can then be undertaken with these models if desired to address retrospective evaluation questions, such as how interventions (e.g., a coal burning ban) probably changed health outcomes, and to address prospective questions such as how a ban now would affect probabilities of future health outcomes.

Example: Attribution of Rainfall to Climate Change, and the Indeterminacy of Counterfactuals

In August of 2017, Hurricane Harvey dumped more than 40 inches of rain on parts of eastern Texas in 4 days, leading to over \$200 billion dollars of flooding damage in 2017 U.S. dollars. Speculation soon began about how much less rain might have fallen in the absence of man-made climate change. Simulation models were used to suggest numerical answers attributing a fraction of the unusual rainfall, such as 15%, to warming over the previous century, meaning that 15% less rain would have been expected had the warming not occurred. However, the correct answer depends on *why*, in the counterfactual scenario, warming would not have occurred. Positing the onset of a civilization-destroying ice age by 1990 that prevented further warming might have different implications for what would have happened to the rainfall in Texas in August of 2017 “had the warming not occurred” than if the explanation were instead earlier and more widespread adoption of nuclear power or other energy sources that reduced greenhouse gas emissions. This illustrates the *indeterminacy of counterfactuals*: assuming that some aspect of the world (temperature rise) were different does not by itself explain *why* it would have been different, and yet this may be crucial for determining what else would have happened (rainfall over Texas in August of 2017), or the probabilities of what else might have happened (probability distribution for rainfalls of different sizes).

Even if well-validated causal models are available to provide credible answers to what would have happened—or, more accurately, what the probability distribution of outcomes would have been—for different input assumptions or scenarios, which specific counterfactual input assumptions or scenarios should be used is typically under-determined by conditions that stakeholders ask about. Asking how much less rain would probably have fallen during Harvey had the warming in the previous century not occurred does not specify a causally coherent scenario about how Harvey would have occurred, and how it would have differed (e.g., producing more or less rainfall than that observed) without that warming. Indeed, it is highly nontrivial, and well beyond current weather and climate simulation capabilities (and perhaps mathematically impossible, if the assumed equations and conditions are not consistent), to identify initial conditions a century ago and plausible changes in decisions since then that would have produced both no warming in the previous century and also a Harvey-like storm over Texas in August, but with significantly less rainfall. Applying simulation models to hypothetical what-if input assumptions to produce corresponding probability distributions for outputs may yield conclusions that represent no more than user-selected inputs and modeling assumptions presented in the guise of discoveries.

To obtain valid counterfactual findings about how the real world would have been different if past inputs had been different, it is important to use validated causal models and causally coherent input assumptions—that is, initial conditions and input scenarios that *explain* how hypothesized conditions (such as an August 2017 storm over Texas) are created by the application of causal laws or validated models to the

stated initial conditions and input scenarios. If a suitable Bayesian network model were available, for example, one could enter assumed findings, such as “Warming over past century = 0” and “Storm occurs over eastern Texas in late August of 2017 = TRUE” and then let the model compute the conditional probability distribution for “Amount of rainfall from storm” (or report that the assumed findings are mutually inconsistent, if that is the case). To avoid inconsistencies and ambiguities, hypothetical counterfactual conditions or constraints such as “A storm occurs in the same time and location, but without the previous century’s warming” should not be *assumed*, but should be *derived* (if possible) from causal models of the (random) consequences of explicitly stated initial conditions and input scenarios. Such specificity about explanations for assumed counterfactual conditions is not provided in statistical counterfactual models. It is seldom provided in simulation modeling. Without such specificity, attributions of excess rainfall or other quantities to partially specified causes, such as presence or absence of prior warming (without further explanation and detailed modeling), may have no clear meaning.

Comparison of Causal Discovery to Structural and Mechanistic Causal Modeling

Long before the advent of causal Bayesian networks and other causal graph models, scientists and engineers successfully described causality in a variety of dynamic physical, socioeconomic, and biological systems using mathematical models, especially systems of ordinary differential equations (ODEs) or partial differential equations (PDEs) with algebraic constraints, to describe how the rates of change of some variables depended on the values of other variables. The equations represented invariant causal laws, such as that the rate of flow of a material across a boundary was proportional to the difference in its concentrations on the two sides of the boundary, or that the rate of change of material in a compartment was the difference between its rate of flow into the compartment (influx) and its rate of flow out of the compartment (efflux). The basic insight into causality was that initial conditions, subsequent exogenous inputs, and invariant causal laws—that is, structural equations that can be applied to the initial conditions and exogenous inputs, but whose form does not depend on them—determine both how the state of a system evolves over time, and also the changes in outputs caused by an exogenous change in the initial conditions or by changes in inputs. Equations 1.7 and 1.8 in Chap. 1 express this idea in mathematical notation and extend it to include stochastic systems in which probability distributions over states evolve over time based on initial conditions and inputs.

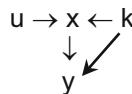
Example: Dynamic Causal Analysis of the Level of a Single Variable in a Compartment

Setting: One of the simplest examples of dynamic causal analysis involves a single compartment with an inflow of u units of material per unit time; an outflow of y units per unit time; and a state x indicating the number of units of material in the compartment. The compartment could be a bathtub filling with water, a cell in a tissue exposed to a chemical, or a population with births plus immigrations as inflows and deaths plus emigrations as outflows. Suppose that the outflow is proportional to the current contents, $y = kx$, and that the inflow is constant. The inflow rate u and the outflow rate per unit of content, k , are exogenously determined inputs to the model.

Problem: Find the causal impacts on the steady-state equilibrium level in the compartment of the following two exogenous changes in inputs: (a) Cut u in half. (b) Cut k in half.

Solution: The ODE describing this one-compartment system is $dx/dt = u - kx$. In steady state equilibrium, the inflow and outflow are equal: $u = kx$. Hence the steady-state level in the compartment is $x^* = u/k$, and the answers to the questions are that (a) Cutting u in half cuts the steady-state level in half; and (b) Cutting k in half doubles the steady-state level.

Discussion. In this explicit dynamic model, it is clear what is determined by what: the exogenous inputs k and u determine the rate of change in x (via $dx/dt = u - kx$), and hence the value of x over time given an initial value of x ; and x and k at any moment determine the outflow y (via $y = kx$). The structure of the model can be diagrammed as follows, where arrows here indicate that the value of each variable is derived from the values of the variables that point into it.



Such a structure implies a partial causal ordering of its variables. The direct causes (parents) of a variable are the variables from which its value is derived, i.e., those that point into it. Nodes with no inward-pointing arrows represent exogenous inputs. This concept extends to even very large dynamic simulation models consisting of ODEs and algebraic formulas, allowing a causal ordering of variables based on the structure of derivations of their values from the exogenous inputs (Simon and Iwasaki 1988).

Today, *system dynamics* modeling offers software and guidelines for representing dynamic systems by ODEs and mathematical formulas. Modern system dynamics modeling software lets users easily draw diagrams with compartments (boxes) representing dynamic quantities (state variables), thick arrows representing dynamic flows (influxes and effluxes) into and out of the compartments, and thin information

arrows showing how the values of some variables, such as inflow and outflow rates, depend on the values of others.

Figure 2.43 shows a portion of a system dynamics modeling tutorial using the free in-browser software Insight Maker. In this simple susceptible-infected-recovered (SIR) model of infectious disease dynamics in a population, the three possible states (compartments) for individuals at any moment are called *Healthy* (i.e., susceptible), *Infected*, and *Immune* (i.e., recovered and no longer susceptible). Initial conditions specify the number of people or the fraction of the population in each state. In this model, 100% of the population is initially in the *Healthy* state. Exogenous inputs are the *Infection Rate* (expressed in units of expected infections per healthy person per year) and the *Recovery Rate* (expressed in units of expected recoveries per infected person per year. More intuitively, this can be viewed as the reciprocal of average recovery time, so that a recovery rate of 2 would correspond to an average recovery time of half a year.) The equations describing causality specify that the flow from *Infected* to *Immune* is equal to the size of the infected population times the recovery rate; likewise, the flow from *Healthy* to *Infected* is the product of *Healthy* and *Infection Rate*. These laws could be written explicitly (and in earlier continuous simulation modeling languages, they had to be written explicitly), e.g., as the system of ODEs

$$d(\text{Healthy})/dt = -\text{Infection_Rate} * \text{Healthy}$$

$$d(\text{Infected})/dt = \text{Infection_Rate} * \text{Healthy} - \text{Recovery_Rate} * \text{Infected}$$

$$d(\text{Immune})/dt = \text{Recovery_Rate} * \text{Infected}.$$

However, it is more parsimonious, and is supported by current system dynamics software packages, for the user to simply specify the equations for each of the two flows:

$$\text{Infection} = \text{Infection_Rate} * \text{Healthy}$$

$$\text{Recovery} = \text{Recovery_Rate} * \text{Infected}.$$

The software can then complete the ODE specification using the conservation law that the rate of change for any compartment at any time is the difference between its influx and efflux rates then.

Once a system dynamics model has been fully specified by specifying its initial conditions (i.e., initial contents in each compartment), exogenous inputs, and equations for flows, it can be run to calculate the time courses of all variables. The software uses numerical integration algorithms to automatically solve for, or simulate, the values of all variables over any user-specified time interval. For the SIR model in Fig. 2.43, the percents of the total population in each of the three states are shown over a time interval of 20 years. These trajectories are generated and displayed automatically by the Insight Maker solver. Special techniques of numerical integration for “stiff” systems can be applied if different variables have transients on very different time scale, but such details of the underlying solvers are typically hidden from users, who need only to specify models and run them to get outputs.

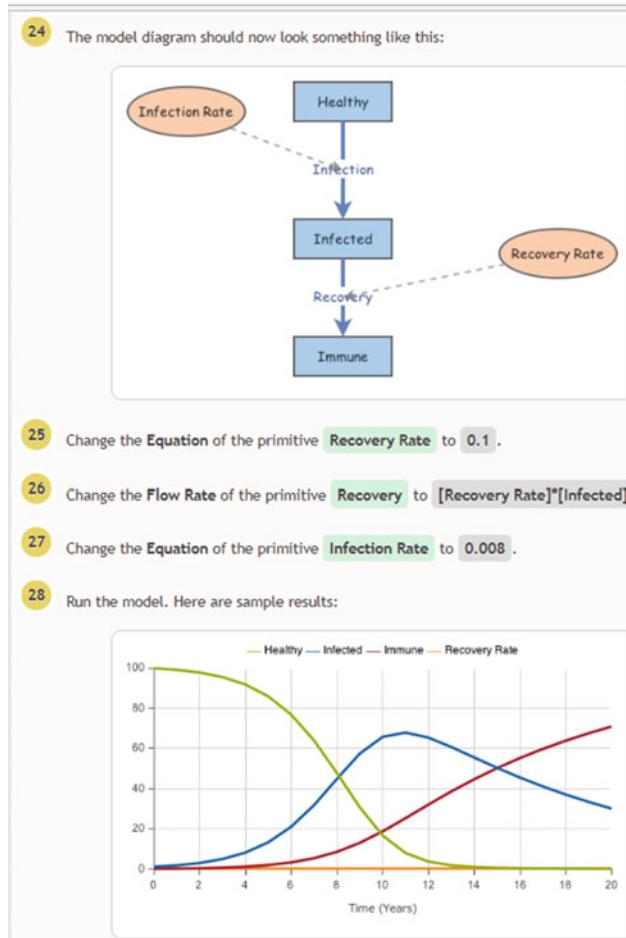


Fig. 2.43 System dynamics modeling using the free Insight Maker software (<https://insightmaker.com/node/3778>)

Dynamic models often contain cycles. Indeed, standard practice in system dynamics simulation modeling is to begin the formulation of a model by using qualitative *causal loop diagrams* showing signed arrows linking variables into networks of overlapping positive or negative feedback loops to understand which changes in variables tend to reinforce each other via positive loops, and which tend to counter-balance or stabilize each other via negative feedback loops. For example, the infection rate in Fig. 2.43 might depend on the fraction of the population infected, *Infected*, creating a reinforcing causal loop between the number of people already infected and the rate of infection of susceptible people. Thus, dynamic *structural models*, consisting of ODEs and algebraic equations organized to allow

simulation of the values of all quantities over time, given initial conditions and exogenous inputs, are not always equivalent to causal DAG models (Simon and Iwasaki 1988). They provide time-dependent outputs and detailed descriptions of transients that are typically abstracted away in BNs, DBNs and other DAG descriptions.

Detailed dynamic causal simulation modeling methods include the following:

- System dynamics modeling, such as the SIR model in Fig. 2.43;
- Agent-based models (ABMs, also supported by Insight Maker). In an ABM, local interactions among agents are described by rules or equations and the dynamic evolution of aggregate population variables emerges from these interactions.
- Networks of ODEs. These are networks in which each variable changes its own level at a rate that depends on the levels of its neighbors. Important special cases include chemical reaction networks, “S-systems” (Savageau and Voit 1987) of parametric ODEs used to describe metabolic networks and other biological networks, and related dynamic networks used in systems biology (Machado et al. 2011). Networks of ODEs are especially useful for studying how steady-state equilibrium levels of variables (if they exist) adjust in response to changes in initial conditions or exogenous inputs; how many different equilibria there are; the stability and basins of attraction for different equilibria; the sizes and durations of exogenous disturbances needed to move the network from one equilibrium to another; and whether there are stable periodic solutions or chaotic solutions.
- Other network simulation models. There are many other network models in which the state of each element evolves based on the states of its neighbors, although networks of ODEs are among the most flexible and useful classes of such models. For example, Boolean logic networks model each variable as having only two values, conventionally represented by 0 (“off”) or 1 (“on”). Each variable’s value in each period depends on the values of its parents in the previous period; the dependence can be deterministic (as in classical “switching networks” of Boolean elements) or probabilistic. Other network formalisms, such as Petri nets (in which a finite number of tokens at each node move among a finite number of places according to specified rules) and networks of finite-state machines (i.e., finite automata) have been subjects of much theoretical analysis and some applications in systems biology, e.g., in modeling gene regulatory networks (Lähdesmäki et al. 2006).
- Discrete-event simulation (DES) models for stochastic systems, as discussed in Chap. 1.

When enough knowledge about a system is available to create a well-validated dynamic simulation model, perhaps with some of its input values or initial conditions sampled from probability distributions representing uncertainty about their values, it can be used to answer questions about how changes in inputs have affected outcome probabilities (for retrospective evaluation and attributional studies) or will affect them (for probabilistic forecasting, decision support, and policy optimization studies). Such mechanistic modeling provides a gold standard for causal modeling of

well understood systems. DAG models can simplify and summarize the input-output relations calculated from these more detailed dynamic models of causal processes, e.g., by using conditional probability tables (CPTs) or other conditional probability models (e.g., CART trees, random forest ensembles, Monte Carlo simulation steps) in a BN to summarize the conditional probabilities of different steady-state output levels for different combinations of exogenous inputs and initial conditions, as determined by more detailed dynamic simulation modeling. Perhaps more importantly for many applications, DAG structures and BNs can also be generated directly from appropriate cross-sectional or time series data, as previously illustrated, even if a more detailed dynamic simulation model is not available.

The analogous task of learning ODE-algebraic or ODE network simulation models directly from input and output observations can be very challenging. A vast engineering literature *on dynamic system identification* algorithms deals largely with special cases such as linear time-invariant (LTI) systems or single-input, single-outputs (SISO) systems (Iserman and Münchhof 2011). BNs, SEMs, and other DAG models and causal graphs provide less detailed representations of causal processes than dynamic simulation models, but are relatively easy to fit to data using non-parametric methods such as those described previously (e.g., using the bnlearn and CAT software). Although they usually cannot answer questions about the dynamic trajectories followed by variables as they adjust to changes in inputs, these models can provide very useful information about how outcomes (or their probabilities, in stochastic models) change as inputs are varied.

Example: A CPT for a One-Compartment Model with Uncertain Inputs

Returning to the one-compartment model with input rate u and output $y = kx$ when the compartment contains x units, recall that for any given initial value of x and for any specified values (or histories, if they are time-varying) of the exogenous inputs u and k , integrating the ODE $dx/dt = u(t) - k(t)x(t)$ provides the time course for $x(t)$. This eventually approaches the steady-state equilibrium value $x^* = u/k$ if both u and k are held fixed. Now, suppose that the future value of k is uncertain, being equally likely to be 0.2, 0.3, or 0.5. Then any choice of value for the fixed input u will induce a conditional probability distribution for the resulting steady-state value of x : it is equally likely to be $5u$, $3.33u$, or $2u$. If u has only a few discrete possible values, or if it is discretized to a small grid of possible values, then a CPT displaying probabilities of 1/3 for each of the values $5u$, $3.33u$, or $2u$ for each value of u can be assembled. If u is continuous, then a simulation step that samples from these three values with equal probabilities, given any input value u , will represent this conditional probability relation without needing to store an explicit CPT.

Example: Causal Reasoning about Equilibria and Comparative Statics

Economists, medical doctors, chemists, and other scientists and engineers make extensive use of causal models that study how the equilibrium in a system changes as exogenous inputs or constraints vary. A demand curve plotting quantity demanded against price shows the equilibrium quantity demanded at each price, without describing the adjustment process or detailing the transients needed to move to the new level when price changes. If the demand curve as a whole shifts up for a certain good (perhaps due to a supply failure for a competing good), while its supply curve remains fixed, then an economist can calculate the new, higher price and the quantity that will be produced and consumed at that price without modeling the transients that accomplish the change. Because it ignores transients, such comparison of equilibria before and after an exogenous change is called comparative statics: it is highly useful for predicting and quantifying the changes that will be caused by exogenous changes in policies, although dynamic models are needed in order to study their timing. Similarly in chemistry and other fields, including economics, Le Chatelier's Principle states that exogenously varying one constraint or factor among many that jointly determine equilibrium (typically in a negative feedback loop) will shift the equilibrium to oppose, or partly offset, the change. In biology, mechanisms that maintain homeostasis play a similar role, and medicine and physiology make heavy use of comparative statics in diagnosing what underlying changes might have led to observed symptoms of disrupted homeostasis. In general, reasoning about structure—which variables are derived from which others—and function in dynamic systems can draw on a rich mix of well-developed applied mathematical modeling tools for predicting equilibria and, if needed, adjustment transients in both deterministic and stochastic dynamic systems. The causal analytics enabled by Bayesian networks and other causal graph models abstracts the structures from such more detailed mathematical models and allows relatively simple but useful modeling of probability relations among levels of variables using much less detailed information.

Historical Milestones in Development of Computationally Useful Causal Concepts

Our review and application of causal inference algorithms has deliberately emphasized principles and algorithms that have succeeded in competitive benchmarking tests, while skipping over centuries of previous work. As noted by Pearl (2014), “Traditional statisticians fear that, without extensive reading of Aristotle, Kant and Hume, they are not well equipped to tackle the subject of causation, especially when it involves claims based on untested assumptions.” Even the relatively short history of computational approaches to causal analysis of data, which is only about a century old, can be intimidating. Some of its key milestones are as follows:

- 1920s: *Path analysis* was introduced and developed by geneticist Sewell Wright ([1921](#)). This was the first approach to use directed acyclic graph (DAG) models in conjunction with quantitative analysis of statistical dependencies and independencies to clarify the distinction between correlation and causality. They have been so used ever since. Although Wright's path analysis was restricted to linear models, it can be seen as a forerunner of the Bayesian networks introduced some 70 years later, which generalize path coefficients to *conditional probability tables* (CPTs). These allow for non-parametric estimation of arbitrary (possibly non-linear) probabilistic dependencies among variables by specifying the conditional probabilities for the possible values of a variable, given each combination of values for the variables that point into it in a DAG model. In practice, this conditional probability distribution or table at a node of a DAG model can be represented relatively efficiently as a classification tree for the node's value, given the values of its parents (inputs) in the DAG, rather than by explicitly listing all possible combinations of input values ([Frey et al. 2003](#)). Path analysis and closely related linear structural equations models (SEMs) were extensively developed by social scientists and statisticians in the 1960s and 1970s and became a primary tools of causal analysis in the social sciences in those decades ([Blalock 1964](#); [Kenny 1979](#)).
- 1950s: *Structural equation models (SEMs)* were developed as tools for causal analysis. For example, polymath and Nobel Laureate Herbert Simon defined causal ordering of variables in systems of structural equations ([Simon 1953](#)) and applied conditional independence and exogeneity criteria for distinguishing between direct and indirect effects and between causal and spurious correlations in econometrics and other fields ([Simon 1954](#)).
- 1960s: *Quasi-experiments* were introduced, standard threats to valid causal inference in observational studies were identified and listed, and statistical designs and tests for overcoming them in observational studies were devised, most notably by social statisticians Campbell and Stanley ([1963](#)). These methods were extended and applied to evaluation of the success or failure of many social and educational interventions in the 1960s and 1970s, leading to a large body of techniques for program evaluation. The methods of data analysis and causal analysis developed for quasi-experiments, which consist largely of enumerating and refuting potential non-causal explanations for observed associations, have subsequently been extensively applied to "natural experiments" in which changes affect a subset of a population, allowing a quasi-experimental comparison of changes in responses in the affected subpopulation to contemporaneous changes in responses in the unaffected (control) subpopulation.
- 1965: *Hill considerations for causality introduced*. In 1965, Sir Austin Bradford Hill, doubting that any valid algorithmic approach for causal discovery could exist, introduced his "considerations" to help humans make judgments about causality based on associations ([Hill 1965](#)). These considerations stand apart from much of the rest of the history of causal analysis methods, being neither greatly influenced by nor greatly influencing the technical developments that have led to successful current algorithms for causal discovery and inference.

They have been enormously influential in encouraging efforts to use judgment to interpret associations causally in epidemiology and public health risk assessment, however. Some attempts have been made to link Hill's considerations to counterfactual causality (Höfler 2005), but they play no role in current causal analysis algorithms, and the rates of false positives and false negative causal conclusions reached with their help have not been quantified. As a psychological aid to help epidemiologists, risk assessors and regulators to make up their minds, Hill's considerations have proved effective, but their performance as a guide for drawing factually correct conclusions about causality—especially manipulative causality—from observational data is less clear.

- 1970s: *Conditional independence tests and predictive causality tests for time series* were developed to identify predictive causal relationships between time series, most notably by Nobel Laureate econometrician Clive Granger and colleague Christopher Sims, building on earlier ideas by mathematician and electrical engineer Norbert Wiener (1956). Granger (or Granger-Sims) tests for predictive causality have been extended to multiple time series and applied and generalized by neuroscientists analyzing observations of neural firing patterns in the brain (Friston et al. 2013; Furgan and Sival 2016; Wibral et al. 2013).
- 1980s: *Counterfactual and potential outcomes techniques* were proposed for estimating average causal effects of treatments in populations, largely by statistician Donald B. Rubin and colleagues, building on work by statistician Jerzy Neyman in 1923. Over the course of four decades, specific computational methods put forward in this framework to quantify average causal effects in populations, usually by trying to use observations and assumptions to estimate what would have happened if treatments or exposures had been randomly assigned, have included matching on observed covariates (Rubin 1974), Bayesian inference (Rubin 1978), matching with propensity scores (Rosenbaum and Rubin 1983), potential outcomes models with instrumental variables (Angrist et al. 1996), principal stratification (Zhang and Rubin 2003), and mediation analysis (Rubin 2004). These methods have been influential in epidemiology, where they have been presented as suitable for estimating average effects caused by treatments or interventions. But they have also been criticized within the causal analysis community as being needlessly obscure, reliant on untestable assumptions, and prone to give biased, misleading, and paradoxical results in practice, in part because they do not necessarily estimate genuine (manipulative) causal effect (e.g., Pearl 2009). From this perspective, the useful contributions of the potential outcomes framework can be subsumed into and clarified by methods of structural equations modeling (*ibid*).

The 1980s also saw the introduction of classification and regression trees (CART) methods (Breiman et al. 1984). These would eventually provide nonparametric tests for conditional independence, useful for learning Bayesian network structures from data (Frey et al. 2003). They also provided the base nonparametric models for Random Forest ensembles and related non-parametric ensemble algorithms now widely used in machine learning (Furqan and Siyal 2016).

- 1990s: *Probabilistic graphical models* were developed in great detail and given clear mathematical and conceptual foundations (Pearl 1993). These included Bayesian networks and causal graph models, together with inference algorithms for learning them from data and for using them to draw causal inferences and to estimate the sizes of effects caused by interventions. These methods are most prominently associated with the Turing Award-winning work of computer scientist Judea Pearl and his coauthors. They grew out of the intersection of artificial intelligence and statistics. They provide a synthesis and generalization of many earlier methods, including structural equations modeling (both linear and nonlinear), probabilistic causation, manipulative causation, predictive (e.g., Granger) causation, counterfactual and potential outcomes models, and directed acyclic graph (DAG) models, including path analysis. Conditional independence tests and quantification of conditional probabilistic dependencies play key roles in this synthesis, as set forth in landmark books by Pearl (2000) and Koller and Friedman (2009). The full, careful development of probabilistic graphical models and algorithms created what appears to be a lasting revolution in representing, understanding, and reasoning about causality in a realistically uncertain world.
- 2000-Present: *Causal discovery and inference algorithms* for learning causal DAG models from data and for using them to draw causal inferences and to quantify or place bounds on the sizes of impacts caused by different interventions have been extensively developed, refined, tested, and compared over the past two decades. Important advances included clarifying which variables in a DAG model must and must not be conditioned on to obtain unbiased estimates of causal impacts in known DAG models (Textor et al. 2016; Shpitser and Pearl 2008), as well as transport formulas for applying causal relationships discovered and quantified in one or more learning settings to a different target setting (Hernan and VanderWeele 2011; Lee and Honavar 2013; Bareinboim and Pearl 2013). Recent years have also seen substantial generalizations of earlier methods. For example, transfer entropy, a nonparametric generalization of Granger causality, quantifies the rates of directed information flows among time series variables. Introduced by physicist Thomas Schreiber (2000) and subsequently refined and extended by workers in computational finance and neuroscience (Wibral et al. 2013), transfer entropy and closely related methods appear to be promising for creating algorithms to discover causal DAG structures and quantitative dependency relationships and time lag characteristics from observations of multiple time series.

Even such an abridged list of milestones makes clear that causal analytics is now a large and deep field with a host of interrelated technical concepts and algorithms supported by a confluence of insights and methods from statistics, social statistics and program evaluation, electrical engineering, economics and econometrics, physics, computer science, computational finance, neuroscience, and other fields. Any brief survey must therefore be relatively superficial; full treatments run into thousands of pages (e.g., Koller and Friedman 2009), and even documentation for R packages implementing the key ideas can be hundreds of pages.

This deep grounding of current information-based causal analytics methods and algorithms in nearly a century of computational methods backed by centuries of philosophizing about causality might well inspire a prudent humility (Pearl 2014). Yet, for the practitioner with limited time and a need to draw sound causal inferences from data, two relatively recent developments make even superficial understanding of key ideas and software packages highly useful. The first is that many formerly distinct causal analysis methods have now been synthesized and unified within the framework of information-theoretic methods and directed acyclic graphs. This framework brings together ideas from potential outcomes and counterfactual causation, predictive causality, DAG modeling, and manipulative causality (Pearl 2000, 2010). The second is the success of the object-oriented software paradigm in platforms such as R and Python. Modern software enables and encourages encapsulation of technical implementation details so that only key ideas and behaviors of software objects need be understood to use them correctly. This allows users with only a superficial understanding of exactly what a software package does and how it works to use it appropriately to do valuable tasks. For example, a user who understands only that causes must be informative about their effects, and that this can be indicated graphically by arrows between variables showing which ones are identified as being informative about each other and which are conditionally independent of each other, can use this limited understanding to interpret correctly the results of sophisticated algorithms such as those in the CAT package. As a practical matter, making tools such as Bayesian network learning algorithms, classification trees, and partial dependency plots widely available and easy to apply can complement insights from regression-based and other associational and counterfactual methods to reveal and quantify potential causal relationships in observational data.

Conclusions

This chapter has introduced several different concepts of causation and has discussed limitations and challenges for applying them in practice to describe how different factors affect risks; predict how alternative actions or changes in the controllable inputs to a system would affect outcome probabilities; optimize decisions to increase the probabilities of desired outcomes; and evaluate how well past actions or policies have succeeded in bringing about their intended goals. Table 2.7 summarizes the major concepts of causation discussed, and challenges for applying each one. Table 2.8 identifies some of the main communities using each concept (middle column) and techniques for implementing each concept using data analysis and modeling methods (right column). Major themes of the chapter are as follows:

- Decision-makers need to understand *manipulative causation* to make well-informed decisions about how the choices they make affect probabilities of outcomes.

Table 2.7 Summary of causal concepts and challenges

Causal concepts	Main ideas	Limitations and challenges
Probabilistic	A cause makes its direct effects more probable	<ul style="list-style-type: none"> Direction of causation unclear: If $P(X Y) > P(X)$ then $P(Y X) > P(Y)$ (since $P(X Y)P(Y) = P(Y X)P(X)$) Observing vs. doing: Seeing a high value of X can make seeing a high value of Y more likely even if increasing X reduces Y
Associational	Stronger associations are more likely to be causal	<ul style="list-style-type: none"> Association is often model-dependent Reducing an associational cause may not reduce its associated effects <ul style="list-style-type: none"> Confounding $X \rightarrow Z \rightarrow Y$ Collider bias $X \rightarrow Z \leftarrow Y$ “Causal” is not dichotomous: Many paths
Attributive	Some fraction of effect can be attributed to each cause based on relative risk ratios (associations)	<ul style="list-style-type: none"> Sums of attributed risks often exceed 100% Reducing a cause may not reduce effects attributed to it
Counterfactual	Causes make probability distributions of their effects different from what they otherwise would have been. Effect size = estimated average difference in responses between real and counterfactual exposed populations	<ul style="list-style-type: none"> What would have been is unobserved Assumption-dependent estimates based on modeling assumptions are often wrong (e.g., Dublin coal burning ban example) What effects would have been if exposure had been different depends on <i>why</i> it would have been different, which is seldom specified
Predictive	Causes help to predict their effects. Effects are not conditionally independent of their direct causes	Confounding: Nicotine-stained fingers can be Granger causes of lung cancer (if smoking is not conditioned on), but cleaning fingers would not necessarily reduce risk of lung cancer
Manipulative	Changing a cause changes its effect (or its probability distribution)	How changing X would change Y cannot necessarily be predicted uniquely from observational data unless a valid causal model is available
Mechanistic/explanatory Structural	Changing causes changes their effects via networks of law-like mechanisms Values of effects (or their probability distributions) are derived from values of their direct causes	<ul style="list-style-type: none"> Mechanisms may be unknown Pathways linking mechanism may be unknown Direct causal parents may be unknown Formulas, models, or CPTs for

(continued)

Table 2.7 (continued)

Causal concepts	Main ideas	Limitations and challenges
		deriving the probabilities of effect variable values from the values of their direct causal parents may be unknown
But-for; Producing	Effects would have been reduced or absent if causes has been reduced or absent	<ul style="list-style-type: none"> What would have been is not directly observed, but must be estimated using assumptions or models (which may be wrong) or mechanistic knowledge

- Manipulative causation is not implied by associational, attributional, or counterfactual causation. This creates a need and an opportunity for other methods to inform decision-makers about the probable consequences of alternative choices.
- Manipulative causation is implied by mechanistic/explanatory causation and by structural causal models that show how the values of some variables are determined from the values of others via structural equations or simulation formulas representing causal mechanisms.
- However, causal structures (e.g., causal graph or BN network topologies) and mechanisms for deriving the value or probability distribution of a variable from the values of the factors or variables on which it depends (e.g., via structural equations, CPTs or conditional probability models in a BN, or simulation steps in a system dynamics model or a discrete-event simulation model) are often initially unknown or uncertain for many systems and risks of interest. Algorithms and principles for discovering them from data (Table 2.4, right column) and for designing studies to produce such data are therefore of great practical interest.
- Predictive causation can often be inferred from data using Granger causality tests and similar statistical methods and using BN learning tools and other machine-learning methods for causal graphs.
- Predictive causation does not necessarily imply manipulative causation, as illustrated by the counter-example of nicotine-stained fingers being a predictive cause but not a manipulative cause of lung cancer.
- Knowledge-based constraints, e.g., specifying that sex and age are sources in a causal graph and that death is a sink, can help orient arrows in a causal graph or BN so that they have valid manipulative interpretations. No fully automated procedure exists that is guaranteed to produce valid manipulative causal graph models from observational data. However, multiple causal discovery algorithms assisted by knowledge-based constraints provide a useful practical approximation to this ideal. Fully automated methods may produce some arrows (or, for some algorithms, undirected arcs) between variables that indicate only that they are informative about each other in a data set, and not necessarily that changing one would change the other.

Table 2.8 Summary of key users and techniques for different causal concepts

Causal concept	Key users	Techniques
Probabilistic	<ul style="list-style-type: none"> • Philosophers • Statisticians 	<ul style="list-style-type: none"> • Conditional probability calculations • Statistical modeling
Associational	<ul style="list-style-type: none"> • Regulators (e.g., EPA, OSHA, FDA, etc.) • World Health Organization, IARC, other public health authorities 	<ul style="list-style-type: none"> • Relative risk (RR) ratios • Epidemiological association metrics • Regression modeling
Attributive	<ul style="list-style-type: none"> • Lawyers • Activists, regulators, litigators • World Health Organization, IARC, other public health authorities 	<ul style="list-style-type: none"> • Burden-of-disease calculations • Other epidemiological measures: Population attributable fraction, probability of causation, etiologic fraction, etc.
Counterfactual	<ul style="list-style-type: none"> • Epidemiologists • Social statisticians • Policy analysts, especially those working on program evaluation 	<ul style="list-style-type: none"> • Potential outcomes models and methods <ul style="list-style-type: none"> – Matching, propensity scores, marginal structural models – Principal stratification – Conditioning via regression models – Quasi-experimental designs
Predictive	<ul style="list-style-type: none"> • Economists and econometricians • Statisticians • Neuroscientists • Physicists • Machine learning and AI researchers 	<ul style="list-style-type: none"> • Granger causality testing • Transfer entropy • Bayesian Networks (BNs) • Dynamic Bayesian Networks (DBNs) • Other time series techniques for conditional independence testing and causal graph modeling • Predictive analytics
Manipulative	<ul style="list-style-type: none"> • Decision makers • Policy makers • Policy evaluators • Engineers 	<ul style="list-style-type: none"> • Causal graph models with knowledge-based constraints • BN learning and inference: conditional independence tests, partial dependence plots (PDPs), <i>DAGitty</i>, <i>bnlearn</i>, <i>CAT</i> • Response surface modeling, adaptive learning and optimization
Mechanistic/explanatory causation Structural causation	<ul style="list-style-type: none"> • Scientists • Scientific modelers • Engineers • Economists and econometricians • Social science researchers • Geneticists • Systems biologists 	<ul style="list-style-type: none"> • Simulation modeling <ul style="list-style-type: none"> – System dynamics modeling, continuous simulation (ODEs, <i>Insight Maker</i>) – Discrete-event simulation – System identification methods • Structural equation modeling (SEMs) • Path analysis • Simon-Iwasaki causal ordering

- To handle model uncertainty, i.e., uncertainty about the correct description of the data-generating process or system underlying the observed data, machine learning algorithms such as Random Forest combine non-parametric estimation of conditional probability relations with the use of model ensembles that allow for the possibility that any of many models might provide the best description. Averaging predictions from many models in such an ensemble typically leads to better predictions (e.g., with lower false positive and false negative rates for classification tasks and smaller mean squared prediction errors for continuous predicted quantities) than any single model.
- Bayesian networks (BNs) provide a useful unifying framework for many aspects of descriptive, predictive, prescriptive, and evaluation analytics. They also support learning from data and collaboration by experts in different parts of the BN.
 - *Description:* The network topology of a BN reveals multivariate patterns of dependencies and conditional independence among variables that are more informative than descriptive methods such as exploratory data analysis and visualization, clustering, or regression alone.
 - *Prediction:* A quantified BN model with all of its CPTs or other conditional probability models specified can be used to predict the values of some variables from observed or assumed values of others (“findings”) via conditional probability calculations, while handling missing data gracefully by only conditioning on what is observed. With stronger assumptions (e.g., linear models, Gaussian errors), BNs and related techniques such as SEM modeling and path analysis can be extended to allow hidden or latent variables; these in turn provide a way to deal with measurement or estimation errors in variables, since the true values can be regarded as latent variables for which the measured or estimated values are observable indicators. Dynamic Bayesian networks (DBNs) provide a way to use multiple observed interdependent time series to help forecast each other’s future values. A variety of other predictive models, such as Hidden Markov Models (HMMs) and Kalman filters for dynamic systems, can be expressed as DBNs.
 - *Prescription and decision optimization:* BN inference algorithms can be used to optimize decisions in influence diagrams (IDs).
 - *Evaluation:* If knowledge-based constraints are incorporated that allow the arrows in a BN to be interpreted as representing manipulative causation, then the causal BN can be used to answer evaluation questions about how much difference past policies or interventions have made in changing outcome probability distributions from what they otherwise would have been.
 - *Learning:* BN learning principles and algorithms such as those on the right side of Table 2.4 can be used to help learn BNs directly from data, although what can be learned is often only predictive causation. Knowledge-based constraints are typically needed to obtain arrows that have valid manipulative-causal interpretations.
- Causal graph methods provide transport formulas for generalizing causal relationships discovered in one or more source data sets (e.g., by identifying invariant

laws or CPTs that hold across settings) and applying them under novel conditions and to target environments not used to produce the training data. Related techniques allow combination and synthesis of causal modeling information from multiple observational and experimental data sets with overlapping variables (Triantafillou and Tsamardinos 2015).

Although much remains to be done, and causal discovery and modeling algorithms are being actively developed by vibrant research communities, the very substantial accomplishments to date provide powerful methods that have proved their empirical value in neuroscience, financial economics and econometrics, control engineering, machine learning, and many applied areas.

The different concepts and methods of causal analytics summarized in Table 2.7 have attracted different followings, as suggested in the middle column of Table 2.8. Most modern concepts of causality are probabilistic; all of those in Table 2.8 agree that, in general, causes change the probability distributions of their effects. Deterministic relationships between changes in causes and changes in effects, as in ODE models, are special cases of more general probabilistic formulations in which conditional probabilities are 1 for one response and 0 for others. (However, attempts to understand causality entirely in terms of probability have been largely confined to philosophers (e.g., Suppes 1970) and are today widely viewed as unsuccessful (Pearl 2009)).

Associational, attributive, and counterfactual causation are widely used in epidemiology and public health. They provide numbers that can often be computed relatively easily from available data, e.g., using observed exposure prevalence numbers and relative risk ratios, or by using observed differences in response rates between a group with a defined exposure or intervention and a control group. The resulting numbers are commonly offered as answers to causal questions by epidemiologists, regulators, activists, litigants, and public health authorities, although they typically do not address manipulative causation. These numbers underlie many sensational headlines about “links” (usually meaning associations) between various exposures and adverse health effects; about pollution killing millions of people per year; about substances being determined by various authorities to be human carcinogens; or about bans of coal burning in Dublin saving thousands of lives. Such reports are widely used to support calls for action and policy recommendations. They are sometimes cited in court battles as evidence for or against probability of a plaintiff’s injury and in calculating probabilities of causation, and they have been used in worker compensation programs to attribute harm or shares in causation to specific causes.

However, associational, attributive, and counterfactual causation usually have no necessary real-world implications for how or whether taking different actions would affect (or has affected) outcome probabilities. Theoretical exceptions occur for associational and attributive causal methods if health effects can be shown to be related to exposures via a competing risk model; and for counterfactual causal methods if the counterfactual modeling assumptions can be shown to be correct. Such exceptions are rare in practice. It is usually the case that even the most sensational headlines based on associational, attributive, or counterfactual causation

linking an exposure to adverse health effect do not imply that reducing or eliminating the exposure would reduce the adverse health effect—its frequency, probability, prevalence, incidence rate, or severity—in an exposed population. This is well understood by many specialists in epidemiology and statistics, but deserves to be much more widely understood by scientists, reporters, and the public. Communicating the limitations of associational, attributive, and counterfactual causal calculations is made more challenging by the understandable tendency of expert practitioners to emphasize their desirable features, such as clarifying the meaning of the causal questions being asked and allowing calculations of numbers that can be readily independently reproduced and verified by others starting from the same data and using the same methods. The advantages of rigor and objectivity are often touted without simultaneously emphasizing that the resulting numbers, causal conclusions, and risk and burden estimates do not mean what most non-specialists think they do: that changing the claimed causes of effects would change the claimed effects.

Toward the other end of the spectrum of causal concepts in Table 2.8 is mechanistic or explanatory causation. Mechanistic causal models describe and explain how initial changes in some variables propagate through network structures of causal mechanisms to bring about subsequent changes in other variables. This is the domain of the scientist and engineer: understanding how causes and effects are connected (structure) and how changes are transduced through complex systems. Quantitative descriptions and dynamic simulations based on networks or systems of equations expressing (theoretical or empirical) causal laws determining how some variables change when others change provide a tremendously powerful paradigm for description, explanation and diagnosis, prediction, what-if counterfactual analysis, and design optimization of systems. Structural causal models consisting of ODEs and algebraic equations (as well as PDEs and stochastic differential equations or stochastic process simulations for some systems) are included with explanatory causation in Table 2.8 because showing how the values of some variables are derived from the values of others—the central concept of structural equations—provides a way to describe the causal mechanisms linking them (Simon and Iwasaki 1988). However, understanding how a complex system works in enough detail to create a valid simulation model or structural equation model describing its dynamic behavior in response to changes in exogenous inputs may require a great deal of knowledge, without which mechanistic causal modeling becomes impossible.

Two other causal concepts in Tables 2.7 and 2.8 are predictive causality and manipulative causality. As already discussed, manipulative causality is, or should be, of primary interest to decision makers and policy analysts. In practice, predictive causation is often a highly useful screen for manipulative causation, insofar as manipulative causation usually implies predictive causation, so that predictive causation is close to being a necessary, although not a sufficient, condition for manipulative causation. Moreover, as we have seen (e.g., Fig. 2.28), incorporating mild knowledge-based constraints, such as that cold temperatures might affect mortality and morbidity but mortality and morbidity do not affect daily temperatures, into predictive causal discovery algorithms such as those in *bnlearn* often suffices to allow them to discover causal graph structures with arrows having manipulative as

well as descriptive and predictive interpretations. Both predictive and manipulative causation are less demanding of detailed knowledge about the structure and functioning of a system and its components than mechanistic causation: knowing *that* changing one variable changes another (or its probability distribution), and by how much, requires less information than understanding *how* changes propagate from one to the other. However, both predictive and manipulative causation are usually more demanding than associational and attributive causation based on observed prevalences and relative risk ratios or on regression coefficients, and than counterfactual causation based on unvalidated modeling assumptions. Both predictive and manipulative causality require knowledge of the dependence and conditional independence relations among variables (e.g., as revealed by a causal graph structure) and CPTs or some other way to specify conditional probability relations, such as regression models or simulation steps, to quantify dependencies. This intermediate level of detail is where most practical work falls. Knowing the probability distribution for changes in outcomes caused by changing controllable inputs is all that is needed to support well informed decisions—but this information *is* needed. Attempts to bypass or simplify it by using more readily available information, such as statistical measures of association or attributable risk, do not provide decision makers with the essential information needed to identify decisions that make preferred outcomes more likely. Nor do they provide judges or litigants with the information needed to identify but-for causation.

This chapter has presented causal concepts and analytics methods with an eye toward practical applications. It has emphasized important distinctions among alternative concepts of causation; surveyed their uses and limitations, especially in risk analysis applications; presented the main principles and ideas of algorithms that have proved useful for causal discovery, inference, and modeling; illustrated modern software implementing them; and discussed how they can be applied to support descriptive, predictive, prescriptive, and evaluation analytics tasks. The following chapters present a variety of applications and extensions of these ideas.

References

- Andreassen S, Hovorka R, Benn J, Olesen KG, Carson ER (1991) A model-based approach to insulin adjustment. In: Proceedings of AIME'91, pp 239–248
- Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *J Am Stat Assoc* 91(434):444–455
- Apte JS, Marshall JD, Cohen AJ, Brauer M (2015) Addressing global mortality from Ambient PM2.5. *Environ Sci Technol* 49(13):8057–8066
- Aragam B, Gu J, Zhou Q (2017) Learning large-scale Bayesian networks with the sparsebn package. arXiv: 1703.04025. <https://arxiv.org/abs/1703.04025>. Accessed 19 Dec 2017
- Asghar N (2016) Automatic extraction of causal relations from natural language texts: a comprehensive survey. <https://arxiv.org/pdf/1605.07895.pdf>. Accessed 19 Dec 2017
- Azzimonti L, Corani G, Zaffalon M (2017) Hierarchical Multinomial-Dirichlet model for the estimation of conditional probability tables. <https://arxiv.org/abs/1708.06935>. Accessed 18 November 2017

- Bareinboim E, Pearl J (2013) Causal transportability with limited experiments. In: Proceedings of the 27th AAAI conference on artificial intelligence, pp 95–101. ftp://ftp.cs.ucla.edu/pub/stat_ser/r408.pdf
- Barnett L, Seth AK (2014) The MVGC multivariate granger causality toolbox: a new approach to granger-causal inference. *J Neurosci Methods* 223:50–68
- Bartholomew MJ, Vose DJ, Tollefson LR, Travis CC (2005) A linear model for managing the risk of antimicrobial resistance originating in food animals. *Risk Anal* 25(1):99–108
- Bearfield G, Marsh W (2005) Generalising event trees using bayesian networks with a case study of train derailment. In: Winther R, Gran BA, Dahll G (eds) Computer safety, reliability, and security. SAFECOMP 2005, Lecture notes in computer science, vol 3688. Springer, Berlin, Heidelberg
- Blalock HM (1964) Causal inferences in nonexperimental research. The University of North Carolina Press, Chapel Hill, NC
- Bobbio A, Portinale L, Minichino M, Ciancamerla E (2001) Improving the analysis of dependable systems by mapping fault trees into Bayesian networks. *Reliab Eng Syst Saf* 71:249–260
- Bontempi G, Flauder M (2015) From dependency to causality: a machine learning approach. *J Mach Learn Res* 16:2437–2457
- Boutilier C, Dearden R, Goldszmidt M (1995) Exploiting structure in policy construction. In: Proceedings of the 14th international joint conference on artificial intelligence, Montreal, QC, Canada, pp 1104–1113
- Brewer LE, Wright JM, Rice G, Neas L, Teuschler L (2017) Causal inference in cumulative risk assessment: the roles of directed acyclic graphs. *Environ Int* 102:30–41. <https://doi.org/10.1016/j.envint.2016.12.005>
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman and Hall/CRC, Boca Raton
- Campbell DT, Stanley JC (1963) Experimental and quasi-experimental designs for research. Houghton Mifflin Company, Boston, MA
- Charniak E (1991) Bayesian networks without tears. *AI Mag* 12(1):50–63. <https://www.aaai.org/ojs/index.php/aimagazine/article/download/918/836>
- Clancy L, Goodman P, Sinclair H, Dockery DW (2002) Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 360(9341):1210–1214
- Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C (2010) Illustrating bias due to conditioning on a collider. *Int J Epidemiol* 39(2):417–420
- Cossalter M, Mengshoel O, Selker T (2011) Visualizing and understanding large-scale Bayesian networks. In: Proceedings of the 17th AAAI conference on scalable integration of analytics and visualization, AAAI Press, pp 12–21
- Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, Hoboken, NJ. - ISBN-13 978-0-471-24195-9. ISBN-10 0-471-24195-4. <https://archive.org/details/ElementsOfInformationTheory2ndEd>. Accessed 9 Jan 2018
- Cox LA Jr (2017a) Do causal concentration-response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality. *Crit Rev Toxicol* 47 (7):603–631. <https://doi.org/10.1080/10408444.2017.1311838>
- Cox LA Jr (2017b) Socioeconomic and air pollution correlates of adult asthma, heart attack, and stroke risks in the United States, 2010–2013. *Environ Res* 155:92–107. <https://doi.org/10.1016/j.envres.2017.01.003>
- Cox LA Jr (1984) Probability of causation and the attributable proportion of risk. *Risk Anal* 4:221–230. <http://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1984.tb00142.x/full>
- Cox LA Jr (1987) Statistical issues in the estimation of assigned shares for carcinogenesis liability. *Risk Anal* 7(1):71–80
- Crowley M (2004) Evaluating influence diagrams. www.cs.ubc.ca/~crowley/papers/aiproj.pdf
- Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Dominici F, Schwartz JD (2017) Association of short-term exposure to air pollution with mortality in older adults. *J Am Med Assoc* 318 (24):2446–2456

- Ding P (2017) A paradox from randomization-based causal inference. *Statist Sci* 32(3):331–345. <https://arxiv.org/pdf/1402.0142.pdf>
- Dockery DW, Rich DQ, Goodman PG, Clancy L, Ohman-Strickland P, George P, Kotlov T, HEI Health Review Committee (2013) Effect of air pollution control on mortality and hospital admissions in Ireland. *Res Rep Health Eff Inst* 176:3–109
- Dominici F, Zigler C (2017) Best practices for gauging evidence of causality in air pollution epidemiology. *Am J Epidemiol*
- Dockery D, Pope C, Xu X et al (1993) An association between air pollution and mortality in six US cities. *N Engl J Med* 329:1753–1759
- Druzdzel MJ, Simon H (1993) Causality in bayesian belief networks. In: UAI'93 proceedings of the ninth international conference on uncertainty in artificial intelligence, Washington, DC, 9–11 July 1993. Morgan Kaufmann Publishers Inc., San Francisco, CA, pp 3–11
- Dugan JB (2000) Galileo: a tool for dynamic fault tree analysis. In: Haverkort BR, Bohnenkamp HC, Smith CU (eds) Computer performance evaluation. Modelling techniques and tools. TOOLS 2000. Lecture Notes in Computer Science, vol 1786. Springer, Berlin, Heidelberg
- Fann N, Lamson AD, Anenberg SC, Wesson K, Risley D, Hubbell BJ (2012) Estimating the national public health burden associated with exposure to ambient PM_{2.5} and ozone. *Risk Anal* 32(1):81–95
- Franklin M, Zeka A, Schwartz J (2006) Association between PM_{2.5} and all-cause and specific-cause mortality in 27 US communities. *J Expo Sci Environ Epidemiol* 17:279–287
- Frey L, Fisher D, Tsamardinos I, Aliferis CF, Statnikov A (2003) Identifying Markov blankets with decision tree induction. In: Proceedings of the third IEEE international conference on data mining, Melbourne, FL, 19–22 Nov 2003. pp 59–66
- Friston K, Moran R, Seth AK (2013) Analyzing connectivity with Granger causality and dynamic causal modelling. *Curr Opin Neurobiol* 23(2):172–178
- Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M (2011) Doubly robust estimation of causal effects. *Am J Epidemiol* 173(7):761–767. <https://doi.org/10.1093/aje/kwq439>
- Furgan MS, Sival MY (2016) Inference of biological networks using Bi-directional Random Forest Granger causality. Springerplus 5(514). <https://doi.org/10.1186/s40064-016-2156-y>
- Gamble JF (2011) Crystalline silica and lung cancer: a critical review of the occupational epidemiology literature of exposure-response studies testing this hypothesis. *Crit Rev Toxicol* 41(5):404–465. <https://doi.org/10.3109/10408444.2010.541223>
- Gharamani Z (2001) An introduction to Hidden Markov models and Bayesian networks. *Int J Pattern Recognit Artif Intell* 15(1):9–42. <http://mlg.eng.cam.ac.uk/zoubin/papers/ijprai.pdf>
- Giannadaki D, Lelieveld J, Pozzer A (2016) Implementing the US air quality standard for PM_{2.5} worldwide can prevent millions of premature deaths per year. *Environ Health* 15(1):88
- Glass TA, Goodman SN, Hernán MA, Samet JM (2013) Causal inference in public health. *Annu Rev Public Health* 34:61–75. <https://doi.org/10.1146/annurev-publhealth-031811-124606>
- Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37(3):424–438
- Greenland S (2015) Concepts and pitfalls in measuring and interpreting attributable fractions, prevented fractions, and causation probabilities. *Ann Epidemiol* 25(3):155–161. <https://doi.org/10.1016/j.anepidem.2014.11.005>
- Hart J, Garshick E, Dockery D, Smith T, Ryan L, Laden F (2011) Long-term ambient multi-pollutant exposures and mortality. *Am J Respir Crit Care Med* 183:73–78
- Hausman DM, Woodward J (2004) Modularity and the causal markov condition: a restatement. *Br J Philos Sci* 55(1):147–161. <https://doi.org/10.1093/bjps/55.1.147>
- Hausman DM, Woodward J (1999) Independence, invariance, and the Causal Markov condition. *Br J Philos Sci* 50(4):521–583. <https://doi.org/10.1093/bjps/50.4.521>
- Heinze-Deml C, Peters J, Meinshausen N (2017) Invariant causal prediction for nonlinear models. <https://arxiv.org/pdf/1706.08576.pdf>

- Hernan M, VanderWeele T (2011) On compound treatments and transportability of causal inference. *Epidemiology* 22:368
- Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58:295–300
- Hill J (2016) Atlantic causal inference conference competition: IS your SATT where it's at? <http://jenniferhill7.wixsite.com/acic-2016/competition>
- Höfler M (2005) The Bradford Hill considerations on causality: a counterfactual perspective. *Emerg Themes Epidemiol* 2:11
- Holt J, Leach AW, Johnson S, Tu DM, Nhu DT, Anh NT, Quinlan MM, Whittle PJL, Mengersen K, Mumford JD (2017) Bayesian networks to compare pest control interventions on commodities along agricultural production chains. *Risk Anal.* <https://doi.org/10.1111/risa.12852>
- Hoover KD (2012) Causal structure and hierarchies of models. *Stud History Philos Sci C* 43(4):778–786. <https://doi.org/10.1016/j.shpsc.2012.05.007>
- IARC (2006) IARC monographs on the evaluation of carcinogenic risk to humans: preamble. International Agency for Research on Cancer (IARC), Lyons, France. <http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf>
- Imai K, Keele L, Tingley D, Yamamoto T (2011) Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am Polit Sci Rev* 4:105
- Iserman R, Münchhof M (2011) Identification of dynamic systems: an introduction with applications. Springer, New York, NY
- Jonsson A, Barto B (2007) Active learning of dynamic Bayesian networks in Markov decision processes. In: SARA'07 proceedings of the 7th international conference on abstraction, reformulation, and approximation, Whistler, Canada, 18–21 July 2007. Springer, Berlin, pp 273–284
- Kahneman D (2011) Thinking fast and slow. Farrar, Straus, and Giroux, New York
- Khakzad N, Reniers G (2015) Risk-based design of process plants with regard to domino effects and land use planning. *J Hazard Mater* 299:289–297. <https://doi.org/10.1016/j.jhazmat.2015.06.020>
- Keele L, Tingley D, Yamamoto T (2015) Identifying mechanisms behind policy interventions via causal mediation analysis. *J Policy Anal Manage* 34(4):937–963
- Kenny DA (1979) Correlation and causality. Wiley, New York
- Khakzad N, Khan F, Amyotte P (2013) Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. *Process Saf Environ Prot* 91(1–2):46–53
- Kleinberg S, Hripcsak G (2011) A review of causal inference for biomedical informatics. *J Biomed Inform* 44(6):1102–1112
- Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge, MA
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28(5):1–26. <https://www.jstatsoft.org/article/view/v028i05/v28i05.pdf>
- Lähdesmäki H, Hautaniemi S, Shmulevich I, Yli-Harri O (2006) Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. *Signal Process* 86(4):814–834. <https://doi.org/10.1016/j.sigpro.2005.06.008>
- Lagani V, Triantafyllou S, Ball G, Tegnér J, Tsamardinos I. (2016) Chapter 2: probabilistic computational causal discovery for systems biology. In: Geris L, Gomez-Cabrero D (eds) Uncertainty in biology: a computational modeling approach. Springer
- Lancet (2017) www.thelancet.com/pb-assets/Lancet/stories/commissions/pollution-2017/Pollution_and_Health_Infographic.pdf
- Lee S, Honavar V (2013) m-transportability: transportability of a causal effect from multiple environments. In: Proceedings of the twenty-seventh AAAI conference on artificial intelligence. www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/viewFile/6303/7210 (Bareinboim and Pearl, 2013; Lee and Honavar, 2013)
- Leu SS, Chang CM (2013) Bayesian-network-based safety risk assessment for steel construction projects. *Accid Anal Prev* 54:122–133. <https://doi.org/10.1016/j.aap.2013.02.019>

- Lepeule J, Laden F, Dockery D, Schwartz J (2012) Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009. *Environ Health Perspect* 120:965–970
- Li J, Ma S, Le T, Liu L, Liu J (2017) Causal decision trees. *IEEE Trans Knowl Data Eng* 29(2):257–271
- Lo WC, Shie RH, Chan CC, Lin HH (2016) Burden of disease attributable to ambient fine particulate matter exposure in Taiwan. *J Formos Med Assoc* 116(1):32–40
- Lok JJ (2017) Mimicking counterfactual outcomes to estimate causal effects. *Ann Stat* 45(2):461–499. <https://doi.org/10.1214/15-AOS1433>
- Machado D, Costa RS, Rocha M, Ferreira EC, Tidor B, Rocha I (2011) Modeling formalisms in systems biology. *AMB Express* 1:45. <https://doi.org/10.1186/2191-0855-1-45>
- Maglogiannis I, Zafiroopoulos E, Platis A, Lambrinoudakis C (2006) Risk analysis of a patient monitoring system using Bayesian network modeling. *J Biomed Inform* 39(6):637–647
- Maldonado G (2013) Toward a clearer understanding of causal concepts in epidemiology. *Ann Epidemiol* 23(12):743–749
- Mauá DD (2016) Equivalences between maximum a posteriori inference in Bayesian networks and maximum expected utility computation in influence diagrams. *Int J Approx Reason* 68 (C):211–229
- McClellan RO (1999) Human health risk assessment: a historical overview and alternative paths forward. *Inhal Toxicol* 11(6–7):477–518
- Mengshoel OJ, Chavira M, Cascio K, Poll S, Darwiche A, Uckun S (2010) Probabilistic model-based diagnosis: an electrical power system case study. *IEEE Trans Syst Man Cybern Part A Syst Hum* 40(5):874–885
- Menzies P (2012) The causal structure of mechanisms. *Stud Hist Phil Biol Biomed Sci* 43(4):796–805. <https://doi.org/10.1016/j.shpsc.2012.05.000>
- Murray CJ, Lopez AD (2013) Measuring the global burden of disease. *N Engl J Med* 369(5):448–457. <https://doi.org/10.1056/NEJMra1201534>
- Nadkarni S, Shenoy PP (2004) A causal mapping approach to constructing Bayesian networks. *Decis Support Syst* 38(2):259–281. [https://doi.org/10.1016/S0167-9236\(03\)00095-2](https://doi.org/10.1016/S0167-9236(03)00095-2)
- National Research Council (2012) Deterrence and the death penalty. Washington, DC: The National Academies Press. doi: <https://doi.org/10.17226/13363>
- Ogarrio JM, Spirtes P, Ramsey J (2016) A hybrid causal search algorithm for latent variable models. *JMLR Workshop Conf Proc* 52:368–379. www.ncbi.nlm.nih.gov/pmc/articles/PMC5325717/
- Neyman J (1923) Sur les applications de la theorie des probabilités aux expériences agricoles: Essai des principes. Master's thesis (trans: Dabrowska DM, Speed TP) Excerpts reprinted in English, *Statistical Science*, vol 5, pp 463–472
- Nowzohour C, Bühlmann P (2016) Score based causal learning in additive noise models. *Statistics* 50(3):471–485
- Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, Keogh JP, Meyskens FL, Valanis B, Williams JH, Barnhart S, Hammar S (1996) Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *N Engl J Med* 334(18):1150–1155
- Pang M, Schuster T, Filion KB, Schnitzer ME, Eberg M, Platt RW (2016) Effect estimation in point-exposure studies with binary outcomes and high-dimensional covariate data—a comparison of targeted maximum likelihood estimation and inverse probability of treatment weighting. *Int J Biostat* 12(2). <https://doi.org/10.1515/ijb-2015-0034>
- Papana A, Kyrtsov C, Kugiumtzis D, Diks C (2017) Assessment of resampling methods for causality testing: a note on the US inflation behavior. *PLoS One* 12(7):e0180852. <https://doi.org/10.1371/journal.pone.0180852>
- Pearl J (1993) Comment: graphical models, causality and intervention. *Stat Sci* 8:266–269 <https://doi.org/10.1214/ss/1177010894>
- Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, Cambridge

- Pearl J (2001) Direct and indirect effects. In: Proceedings of the seventeenth conference on uncertainty in artificial intelligence, Morgan Kaufmann, San Francisco, CA, pp 411–420
- Pearl J (2009) Causal inference in statistics: an overview. *Stat Surv* 3:96–146. https://projecteuclid.org/download/pdfview_1/euclid.ssu/1255440554
- Pearl J (2010) An introduction to causal inference. *Int J Biostat* 6(2):7
- Pearl J (2014) Reply to commentary by Imai, Keele, Tingley, and Yamamoto to concerning causal mediation analysis. *Psychol Methods* 19(4):488–492
- Peters J, Bühlmann P, Meinshausen N (2016) Causal inference using invariant prediction: identification and confidence intervals. *J R Stat Soc Ser B* 78(5):947–1012
- Petersen ML, Sinisi SE, van der Laan MJ (2006) Estimation of direct causal effects. *Epidemiology* 17(3):276–284
- Petitti DB (1991) Associations are not effects. *Am J Epidemiol* 133(2):101–102. <https://academic.oup.com/aje/article-abstract/133/2/101/118425/Associations-Are-Not-Effects?redirectedFrom=PDF>
- Peyraud N, Givry S, Franc A, Robin S, Sabbadin R, Schiex T, Vignes M (2015) Exact and approximate inference in graphical models: Variable elimination and beyond. <https://arxiv.org/pdf/1506.08544.pdf>
- Poole DL, Mackworth AK (2017) Artificial intelligence: foundations of computational agents, 2nd edn. Cambridge University Press. <http://artint.info/2e/html/ArtInt2e.html>
- Prüss-Üstün A, Mathers C, Corvalán C, Woodward A (2003) Introduction and methods: Assessing the environmental burden of disease at national and local levels, Environmental burden of disease series No. 1. World Health Organization (WHO), Geneva, Switzerland. www.who.int/quantifying_ehimpacts/publications/en/9241546204chap4.pdf?ua=1
- Relton C, Torgerson D, O'Cathain A, Nicholl J (2010) Rethinking pragmatic randomised controlled trials: introducing the “cohort multiple randomised controlled trial” design. *BMJ* 340:c1066. <https://doi.org/10.1136/bmj.c1066>. <http://www.bmj.com/content/340/bmj.c1066>
- Rhomberg LR, Chandalia JK, Long CM, Goodman JE (2011) Measurement error in environmental epidemiology and the shape of exposure-response curves. *Crit Rev Toxicol* 41(8):651–671. <https://doi.org/10.3109/10408444.2011.563420>
- Richardson TS, Rotnitzky A (2014) Causal etiology of the research of James M. Robins. *Stat Sci* 29 (4):459–484. <https://doi.org/10.1214/14-STS505>
- Rigaux C, Ancelet S, Carlin F, Nguyen-thé C, Albert I (2013) Inferring an augmented Bayesian network to confront a complex quantitative microbial risk assessment model with durability studies: application to *Bacillus cereus* on a courgette purée production chain. *Risk Anal* 33 (5):877–892. <https://doi.org/10.1111/j.1539-6924.2012.01888.x>
- Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3:143–155
- Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55. <https://doi.org/10.2307/2335942>
- Rothenhausler D, Heinze C, Peters J, Meinschausen N (2015) BACKSHIFT: learning causal cyclic graphs from unknown shift interventions. arXiv pre-print <https://arxiv.org/pdf/1506.02494.pdf>. See also the BACKSHIFT R package at <https://cran.r-project.org/web/packages/backShift/backShift.pdf>
- Rubin D (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66(5):688–701
- Rubin D (1978) Bayesian inference for causal effects: the role of randomization. *Ann Stat* 6:34–58
- Rubin DB (2004) Direct and indirect causal effects via potential outcomes. *Scand J Stat* 31:161–170
- Sanchez-Graillet O, Poesio M (2004) Acquiring Bayesian networks from text. In: Proceedings of the fourth international conference on language resources and evaluation (LREC'04), Lisbon, Portugal, May 26–28. European Language Resources Association (ELRA), Paris, France. [www.lrec-conf.org/proceedings/lrec2004/](http://lrec-conf.org/proceedings/lrec2004/)
- Savageau M, Voit E (1987) Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. *Math Biosci* 87(1):83–115

- Schaffter T, Marbach D, Floreano D (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* 27(16):2263–2270
- Schreiber T (2000) Measuring information transfer. *Phys Rev Lett* 85(2):461–464. <https://doi.org/10.1103/PhysRevLett.85.461>
- Shachter RD (1986) Evaluating influence diagrams. *Oper Res* 34(6):871–882
- Shachter RD, Bhattacharjya D (2010) Solving influence diagrams: exact algorithms. In: Cochran J et al (eds) Wiley encyclopedia of operations research and management science. Wiley, New York. www.it.uu.se/edu/course/homepage/aism/st11/Shachter10.pdf
- Schwartz S, Gatto NM, Campbell UB (2011) Transportability and causal generalization. *Epidemiology* 22(5):745–746
- Schwartz J, Laden F, Zanobetti A (2002) The concentration-response relation between PM(2.5) and daily deaths. *Environ Health Perspect* 110(10):1025–1029
- Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A (2006) A linear non-Gaussian acyclic model for causal discovery. *J Mach Learn Res* 7:2003–2030
- Shpitser I, Pearl J (2008) Complete identification methods for the causal hierarchy. *J Mach Learn Res* 9(Sep):1941–1979
- Simon HA (1953) Chapter III: Causal ordering and identifiability. In: Hood WC, Koopmans TC (eds) Studies in econometric method, Cowles Commission for Research in Economics Monograph No. 14. Wiley, New York, NY, pp 49–74
- Simon HA (1954) Spurious correlation: a causal interpretation. *J Am Stat Assoc* 49(267):467–479
- Simon HA, Iwasaki Y (1988) Causal ordering, comparative statics, and near decomposability. *J Econ* 39:149–173. <http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=34081>
- Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, Shete S, Etzel CJ (2007) A risk model for prediction of lung cancer. *J Natl Cancer Inst* 99(9):715–726
- Suppes P (1970) A probabilistic theory of causality. North-Holland Publishing Company, Amsterdam, Holland
- Tashiro T, Shimizu S, Hyvärinen A, Washio T (2014) ParceLiNGAM: a causal ordering method robust against latent confounders. *Neural Comput* 26(1):57–83. https://doi.org/10.1162/NECO_a_00533
- Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT (2016) Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *Int J Epidemiol* 45(6):1887–1894
- Theocharous G, Murphy K, Kaelbling LP (2004) Representing hierarchical POMDPs as DBNs for multi-scale robot localization. In: Proceedings of the IEEE international conference on robotics and automation ICRA’04
- Triantafillou S, Tsamardinos I (2015) Constraint-based causal discovery from multiple interventions over overlapping variable sets. *J Mach Learn Res* 16:2147–2205
- Trovati M (2015) Extraction of Bayesian networks from large unstructured datasets. In: Trovati M, Hill R, Anjum A, Zhu S, Liu L (eds) Big-data analytics and cloud computing. Springer, Cham
- Tudor RS, Hovorka R, Cavan DA, Meekling D, Hejlesen OK, Andreassen S (1998) DIAS-NIDDM—a model-based decision support system for insulin dose adjustment in insulin-treated subjects with NIDDM. *Comput Methods Prog Biomed* 56(2):175–191
- VanderWeele TJ, Vansteelandt S (2009) Conceptual issues concerning mediation, interventions and composition. *Stat Its Interface* 2:457–468
- Voortman M, Dash D, Drudzsel MJ (2010) Learning causal models that make correct manipulation predictions with time series data. *Proc Mach Learn Res* 6:257–266. <http://proceedings.mlr.press/v6/voortman10a/voortman10a.pdf>
- Frignat P, Avila M, Duculty F, Kratz F (2015) Failure event prediction using Hidden Markov Model approaches. *IEEE Trans Reliab* 99:1–11
- Westreich D (2012) Berkson’s bias, selection bias, and missing data. *Epidemiology* 23(1):159–164. <https://doi.org/10.1097/EDE.0b013e31823b6296>

- Wibral M, Pampu N, Priesemann V, Siebenhuhner F, Seiwert H, Lindner M, Lizier JT, Vicente R (2013) Measuring information-transfer delays. *PLoS One* 8(2):e55809. <https://doi.org/10.1371/journal.pone.0055809>
- Wintle BC, Nicholson A (2014) Exploring risk judgments in a trade dispute using Bayesian networks. *Risk Anal* 34(6):1095–1111. <https://doi.org/10.1111/risa.12172>
- Wickham H (2014) Tidy data. *J Stat Softw* 59(10):1–23
- Wiener N (1956) The theory of prediction. In: Beckenbach EF (ed) *Modern mathematics for engineers*, vol 1. McGraw-Hill, New York
- Wright S (1921) Correlation and causation. *J Agric Res* 20:557–585. www.ssc.wisc.edu/soc/class/soc952/Wright/Wright_Correlation%20and%20Causation.pdf
- Wu AH, Yu MC, Thomas DC, Pike MC, Henderson BE (1988) Personal and family history of lung disease as risk factors for adenocarcinoma of the lung. *Cancer Res* 48(24 Pt 1):7279–7284
- Zhang J (2008) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif Intell* 172(16–17):1873–1896
- Zhang JL, Rubin DB (2003) Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *J Educ Behav Stat* 28:353–368. <https://doi.org/10.3102/10769986028004353>
- Zhang L, Wu X, Qin Y, Skibniewski MJ, Liu W (2016) Towards a fuzzy Bayesian network based approach for safety risk analysis of tunnel-induced pipeline damage. *Risk Anal* 36(2):278–301. <https://doi.org/10.1111/risa.12448>

Part II

**Descriptive Analytics in Public and
Occupational Health**

Chapter 3

Descriptive Analytics for Public Health: Socioeconomic and Air Pollution Correlates of Adult Asthma, Heart Attack, and Stroke Risks



Introduction

This is the first of four chapters emphasizing the application of *descriptive analytics* to characterize public and occupational health risks. Much of risk analysis addresses basic descriptive information: how big is a risk now, how is it changing over time or with age, how does it differ for people or situations with different characteristics, on what factors does it depend, with what other risks or characteristics does it cluster? Such questions arise not only for public and occupational health and safety risks, but also for risks of failures or degraded performance in engineering infrastructure or technological systems, financial systems, political systems, or other “systems of systems” (Guo and Haimes 2016). Simply knowing how large a risk is now and whether it is increasing, staying steady, or decreasing may be enough to decide whether a proposed costly intervention to reduce it is worth considering further. This chapter shows how to use basic tools of descriptive analytics, especially interaction plots (showing the conditional expected value of one variable at different levels of one or more other variables), together with more advanced methods from Chap. 2, such as regression trees, partial dependence plots, Bayesian networks (BNs), to describe risks and how they vary with other factors. A brief discussion and motivation of these methods is given for readers who have skipped Chap. 2. Chapter 4 introduces additional descriptive techniques, including plots that use non-parametric regression to pass smooth curves or surfaces through data clouds. It shows how they can be used, together with simple mathematical analysis, to resolve a puzzle that has occasioned some debate among toxicologists: that some studies have concluded that workers form disproportionately high levels of benzene metabolites at very low occupational exposure concentrations compared to higher concentrations, while other studies conclude that metabolism of benzene at low concentrations is approximately linear, and proportional to concentrations in inhaled air. Chapter 5 emphasizes the value of descriptive plots, upper-bounding analyses, and qualitative

assumptions, as well as more quantitative risk assessment modeling, in bounding the size of human health risks from use of antibiotics in food animals. Chapter 6 calculates plausible bounds on the sizes of the quantitative risks to human health of infection with a drug-resistant “super-bug” from swine farming operations. Together, these chapters illustrate how descriptive analytics can be used to obtain and present useful quantitative characterizations of human health risks despite realistic scientific uncertainties about the details of relevant causal processes.

Asthma in the United States is an important public health issue. Many physicians, regulators, and scientists have expressed concern that exposures to criterion air pollutants have contributed to a rising tide of asthma cases and symptoms. The following sections describe associations between self-reported asthma experiences and various socioeconomic factors in survey data, as well as pollution data from other sources. Interaction plots are used to investigate and visualize statistical associations among variables. We then apply Bayesian network learning algorithms and other non-parametric machine-learning algorithms to further describe these statistical dependencies and to clarify possible causal interpretations. Associations with self-reported heart attack and stroke experience confirm that well-established relations between smoking and heart attack or stroke risks are seen in this data set (Shah and Cole 2010; Oliveira et al. 2007).

Readers with limited interest in asthma, stroke, and heart attack risks may skim the rest of this chapter without impairing understanding of subsequent chapters. However, we recommend looking at the figures, as they illustrate the use of interaction plots and other diagrams to show how risks cluster and how they vary with other factors. A brief summary of the empirical findings is that self-reported heart attack and stroke experience are positively associated with each other and with self-reported asthma risks. Intriguingly, young divorced women with low incomes are at greatest risk of asthma, especially if they are ever-smokers. Income is an important confounder of other relations. (For example, in logistic regression modeling, PM2.5 is positively associated ($p < 0.06$) with both stroke risk and heart attack risk when these are regressed only against PM2.5, sex, age, and ever-smoking status, but not when they are regressed against these variables and income.) In this data set, PM2.5 is significantly negatively associated with asthma risk in regression models, with a $10 \mu\text{g}/\text{m}^3$ decrease in PM2.5 corresponding to about a 6% increase in the probability of asthma, possibly because of confounding by smoking, which is negatively associated with PM2.5 and positively associated with asthma risk. A variety of non-parametric methods are used to quantify these associations and to explore potential causal interpretations.

Data Sources

To investigate the association between air pollutants (O₃ and PM2.5) and self-reported adult asthma, stroke, and heart attack risks, we merged the following data sources: (a) The most recent 5 years of available survey response data from a survey

of over 228,000 individuals from 15 states, retrieved from the Center for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance (BRFSS) System (www.cdc.gov/brfss/questionnaires/state2013.htm); and (b) Environmental Protection Agency (EPA) data on O₃ and PM2.5 concentrations for the counties in which these individuals lived at the time of the survey, retrieved from the US EPA web site (www.epa.gov/airtrends/pm.html). Counties were used as the common key for merging annual average air pollution levels with individual response data. Table 3.1 summarizes the number of individual responses from each state for each of several questions. These responses are coded so that a response of “Yes” has a value of 1 and a value of “No” has a value of zero. Other responses, or non-responses, are coded as missing data. Thus, for example, 38% of the 8618 respondents from Arizona were male (giving a mean value of 0.38 to the variable “Sex = Male” (henceforth abbreviated as “Sex”) with values of 1 for men and 0 for women). As suggested by this example, the respondents in the BRFSS do not constitute a simple random sample of the population. The BRFSS survey supplies county weights for reweighting responses to better reflect the entire population. However, this chapter does not seek to extrapolate relations outside the surveyed population, but focuses on quantifying conditional relations within this sample, e.g., studying how probability of asthma varies by age and sex and other variables, without considering how to adjust for differences between the joint frequency distribution of these variables in the survey population and in the more general population.

Similarly, not every respondent answered all questions, and there is no guarantee that responses can be extrapolated from those who did to those who did not. Hence, we only consider questions that were answered by almost all of the 228,369 respondents. For the variables in Table 3.1, for example, over 95% of surveyed individual answered each question.

The BRFSS data consist primarily of either dichotomous (yes-no) variables such as those in Table 3.1, all of which are coded as binary (0-1) variables with 0 = no, 1 = yes; or categorical variables, including age (50–99 years), income, education, and marital status. To these we added the two continuous pollution variables obtained from EPA: average daily O₃ concentration in ppm and average daily PM2.5 concentration in micrograms of fine particulate matter per cubic meter of air. Table 3.2 lists the complete set of variables analyzed (other than survey year, month, and location) and their means and minimum and maximum values, as well as the number of individuals responding to each question. Table 3.3 shows the layout of the data (the first 21 of 228,369 records) for individual respondents. Ozone measurements were not available for the county (Apache County, AZ), year, and month of the survey (January, 2010) for these 21 individuals. The entire data set is available from the author upon request.

In Table 3.3, the three categorical variables *Income*, *Education*, and *Marital Status* have integer values for responses of 1–8, 1–6, and 1–6, respectively, with higher numerical values representing higher levels for *Income* and *Education*. *Smoking* is a binary variable that indicates whether a respondent reports having smoked at least 100 cigarettes (five packs) during his or her life to date. The

Table 3.1 Means and frequency counts (*N*) for individual responses to different questions on the BRFSS survey, for 2008–2012. Responses are broken down by states (rows)

State means	Sex = Male Sex N	Asthma ever means	Asthma ever N	FluShot means	FluShot N	Health plan means	Health plan N	Heart attack ever means	Heart attack ever N	Hispanic means	Hispanic N
AZ 0.38	8618 0.14	8592 0.55	8379 0.94	8600 0.09		8562 0.10		8559			
CA 0.40	25,528 0.13	25,505 0.52	23,146 0.93	25,515 0.07		25,499 0.16		25,463			
FL 0.37	9915 0.12	9895 0.52	9557 0.91	9887 0.10		9844 0.10		9814			
GA 0.34	1925 0.12	1919 0.53	1850 0.92	1922 0.07		1914 0.02		1915			
IL 0.36	4638 0.12	4631 0.50	4532 0.93	4634 0.07		4619 0.05		4620			
MA 0.37	49,621 0.14	49,451 0.57	46,565 0.97	49,461 0.08		49,329 0.06		49,319			
MI 0.35	10,334 0.13	10,310 0.50	10,098 0.94	10,312 0.09		10,260 0.02		10,276			
NJ 0.38	27,550 0.12	27,466 0.51	26,113 0.93	27,478 0.08		27,420 0.08		27,423			
NY 0.37	6939 0.12	6912 0.58	6706 0.94	6912 0.07		6888 0.07		6866			
NC 0.37	8935 0.12	8916 0.59	8745 0.93	8922 0.08		8894 0.02		8911			
OH 0.36	17,820 0.12	17,761 0.54	17,336 0.93	17,781 0.09		17,690 0.01		17,729			
PA 0.36	9770 0.12	9735 0.56	9472 0.94	9747 0.09		9705 0.02		9708			
TX 0.37	13,110 0.13	13,074 0.56	12,727 0.90	13,076 0.08		13,020 0.21		12,977			
VA 0.43	388 0.11	385 0.60	377 0.95	388 0.06		387 0.02		387			
WA 0.40	33,278 0.15	33,172 0.57	32,814 0.94	33,234 0.07		33,033 0.02		33,131			
All gps	228,369 0.13	227,724 0.55	218,407 0.94	227,869 0.08		227,064 0.07		227,098			

Table 3.2 Variables, number of records with complete data for each question, and mean, minimum, and maximum values

Variable	Valid N	Mean	Minimum	Maximum
Age	228,369	65.64	50.0	99.00
Sex	228,369	0.38	0.0	1.00
Income	193,321	5.66	1.0	8.00
Education	227,945	4.92	1.0	9.00
Marital status	228,087	2.12	1.0	9.00
Smoking	225,543	0.51	0.0	1.00
PM2.5	222,349	9.39	1.4	31.54
O3	177,148	0.04	0.0	0.08
Asthma ever	227,724	0.13	0.0	1.00
Heart attack ever	227,064	0.08	0.0	1.00
Stroke ever	227,606	0.05	0.0	1.00

dependent variables *Asthma Ever*, *Heart Attack Ever*, and *Stroke Ever* are answers to the question of whether a doctor, nurse, or other health professional had ever told the respondent that s/he had the corresponding condition, with answers are coded as 1 for yes, 0 for no, and blank (missing) for all other values.

Methods and Analytic Strategy

Since most of the variables in this data set other than age, PM2.5, and O3 are dichotomous or categorical, it is useful to examine associations and interactions among them using interaction plots that show how the mean value of one variable varies with the levels of one or more others. The following sections plot the main dependent variables of interest (prevalence of self-reported asthma, stroke, or heart attack) against explanatory variables such as age, income, sex, and average concentrations of O3 and PM2.5 in the counties where respondents lived at the time of the survey. Traditional 95% confidence intervals (mean plus or minus 1.96 sample standard deviations) are indicated visually as vertical bars around the mean values shown in the interaction plots. Such exploratory data analysis can reveal nonlinear patterns of association and does not require any parametric modeling assumptions. However, interaction plots are most useful for examining the relations among only a few explanatory variables and the dependent variables. We also used multiple logistic regression models to quantify associations between multiple explanatory variables and health effects, and used a non-parametric Bayesian network (BN) learning program (the *bnlearn* package in R) to discover and visualize statistical dependence relations (represented by arrows between variables) and conditional independence relations (represented by a lack of arrows between variables) among all variables simultaneously.

Table 3.3 Layout of the data, showing values of variables for the first 21 individual respondents

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
STATE_CODE	State	COUNTY_CODE	County_name	Year	Month	Age	Sex	Income	Education	Marital status	Smoking	PM2.5	O3	Asthma ever	Heart attack ever	Stroke ever	
4	AZ	1	Apache County	2010	1	95	0		4	3	1	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	73	0	8	6	1	1	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	66	0		5	1	0	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	75	1	3	5	2	1	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	70	0	4	4	3	1	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	68	0	5	5	3	1	1.81	1	0	0	0	
4	AZ	1	Apache County	2010	1	72	0		6	1	0	1.81	0	1	1	0	
4	AZ	1	Apache County	2010	1	66	1	7	6	1	0	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	62	0	5	3	1	1.81	0	0	0	0	0	
4	AZ	1	Apache County	2010	1	53	0	8	6	1	1	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	75	0	3	5	3	0	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	53	0	6	4	3	0	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	63	0	7	5	1	0	1.81	0	0	0	0	
4	AZ	1	Apache County	2010	1	92	0		5	2	0	1.81	0	0	0	0	

4	AZ	1	Apache County	2010	1	69	0	3	4	2	0	1.81	0	0	0
4	AZ	1	Apache County	2010	1	51	0	8	4	1	0	1.81	0	0	0
4	AZ	1	Apache County	2010	1	50	0	6	9	0	1.81	0	0	0	0
4	AZ	1	Apache County	2010	1	64	0	8	6	1	0	1.81	0	0	0
4	AZ	1	Apache County	2010	1	99	0	5	2	1	1.81	0	0	0	0
4	AZ	1	Apache County	2010	1	61	1	3	5	2	1	1.81	0	0	0
4	AZ	1	Apache County	2010	1	88	1	5	6	1	0	1.81	0	0	0

Each column represents a variable, and each row contains the data for one respondent. The variables have various scale types. Sex (0 = female, 1 = male), smoking (0 = no, 1 = yes), and the health outcome indicators for asthma, heart attack, and stroke (0 = no, 1 = yes) are binary variables. Income and Education are ordered categorical variables, Marital Status is a nominal variable (see text for details), and PM2.5 and O3 are continuous

Potential causal relations in observational data can be clarified using modern nonparametric methods. Many top-performing methods in recent competitions that evaluate the empirical performance of causal discovery and inference algorithms on suites of test problems (e.g., Hill 2016; NIPS 2013) use the following ideas, as discussed in more detail in Chap. 2:

- *Information principle*: Causes provide information about their effects that help to predict them and that cannot be obtained from other variables. This principle creates a bridge between well-developed statistical and machine learning methods for identifying informative variables that improve prediction of dependent variables (such as health effects), on the one hand, and the needs of causal inference, on the other (Frey et al. 2003; Aliferis et al. 2010). Only variables that help to predict an effect by providing information that is not redundant with that from other variables (e.g., measured confounders) are candidates to be its causes. This constraint allows techniques of predictive analytics to be applied as screens for potential causation (Pearl 2010).
- *Nonparametric analyses*. Multivariate non-parametric methods, most commonly, classification and regression trees (CART) algorithms, can be used to identify and quantify information dependencies among variables without having to make any parametric modeling assumptions (e.g., Frey et al. 2003; Halliday et al. 2016). Conversely, if no significant change occurs in the conditional empirical cumulative distribution function of a dependent variable as the value of an explanatory variable varies, for any combination of values of the remaining variables (so that that explanatory variable does not appear in CART trees for the dependent variable), then this lack of dependence does not support a conclusion that the explanatory variable is a cause of the dependent variable. The dependent variable is then said to be *conditionally independent* of the explanatory variable, given the values of other variables. Effects are not conditionally independent of their direct causes. CART trees can also be used to test for conditional independence, with the dependent variable being conditionally independent of variables not in the tree, given the variables that are in it, at least as far as the tree-growing algorithm can discover (Frey et al. 2003; Aliferis et al. 2010).
- *Model ensembles*. Rather than relying on any single statistical model, the top-performing causal analytics algorithms typically fit hundreds of nonparametric models (e.g., CART trees), called *model ensembles*, to randomly generated subsets of the data (Furqan and Siyal 2016). Averaging the resulting predictions of how the dependent variable depends on other variables over an ensemble of models usually yields better estimates with lower bias and error variance than any single predictive model.

Our analytic plan for clarifying potential causal relations is as follows:

1. *Identify statistical dependencies* and conditional independence relations among variables in Table 3.1 via nonparametric methods (described below). This step screens for possible causal relations using the information principle that variable X is a potential cause of variable Y only if X provides information that helps to

predict Y and that cannot be obtained from other sources (Pearl 2010). An arrow between two variables in a DAG model shows that one is informative about the other (see Fig. 3.10). To facilitate simple interpretations, we also provide partial correlation coefficients and their significance levels for every pair of variables.

2. *Quantify the association* between PM2.5 and adverse health outcomes using the Random Forest machine learning algorithm (i.e., an ensemble of nonparametric regression trees fit to different random subsets of the data) to correct for the observed values of all other variables, and compare the results to those from parametric regression modeling. This step simply quantifies the dependence (if any) between two variables, without assessing whether it is causal, taking into account model uncertainty by refusing to commit to any single model or parametric class of models. We carry it out using a partial dependence plot generated by the *randomForest* package in R, which averages the results of hundreds of regression trees fit to random subsets of the data to obtain a non-parametric estimate of how asthma varies as PM2.5 is swept over its full range of values. (The *randomForest* package documentation contains details.)

To facilitate easy replication and interpretation by other investigators without requiring skill in R, we accessed these packages and displayed the results using the Causal Analysis Toolkit (CAT) introduced in Chap. 2. The methods are described in more detail in the online documentation for the corresponding R packages. Figures 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9 were generated by the *Statistica* commercial software packages, and all other analyses and figures were generated using R via the CAT software.

A conspicuous challenge for this data set is that respondents answered questions about whether they had *ever* been told that they had asthma, heart attack, or stroke, but exposure concentrations are recorded only for the specific counties that they lived in at the time of the survey and only for the years 2008–2012. This raises the possibility that people who moved residences or who were diagnosed with these health conditions when pollution conditions were quite different from those in 2008–2012 might contribute irrelevant responses that dilute any observable relation between current pollutant concentrations and health effects. We meet this challenge by showing that the relative ranking of counties by pollution levels is fairly stable over the 5 years of the study and is significantly associated with health effects; thus, although such dilution of associations probably occurs, there appears to be sufficient signal in this large data set to overcome the noise.

Results

Dependence of Health Effects on Age and Sex

Figure 3.1 presents three interaction plots describing how risk of ever having had a heart attack (upper left), stroke (lower left), and asthma (upper right) vary with age

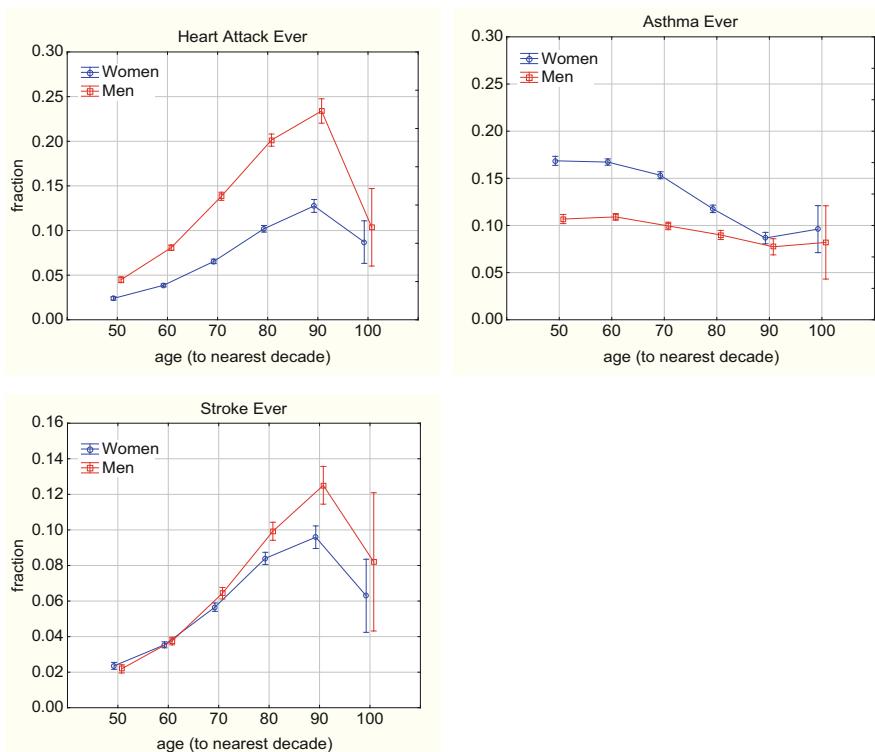


Fig. 3.1 Fraction of respondents reporting ever having had a heart attack (upper left panel), asthma (right panel) or stroke (lower panel) vs. age (horizontal axes, age in years rounded to the nearest decade). Vertical bars are 95% confidence intervals. Men and women aged 65–95 are more likely to have had a heart attack or stroke and less likely to have had asthma than people aged 50–65. Men (blue circles) are more likely than women (red squares) to have had heart attacks and strokes, but women have higher asthma risks than men

and sex. The horizontal axis shows age categories; these represent ages rounded to the nearest decade (e.g., ages 65–75 are rounded to 70; ages 75–85 are rounded to 80, and so forth). The fractions of men (red curves with square data points) and women (blue curves with round data points) that report ever being told by a medical professional that they have each health condition are plotted on the vertical axes. Separate curves are shown for men and women. A 95% confidence interval (vertical bar) is shown around each data point.

With the exception of the oldest age group (those over 95, rounded to 100, who may be exceptionally healthy), risks of ever having had a heart attack or a stroke increase dramatically with age, especially for men. By contrast, risk of ever having been diagnosed with asthma *decreases* with age and is greater for women than for men. Since the risks shown in the upper right panel of Fig. 3.1 are cumulative, i.e., they represent ever having been diagnosed with asthma, they can only decrease with age if asthma risks have been increasing over time, so that younger people are more

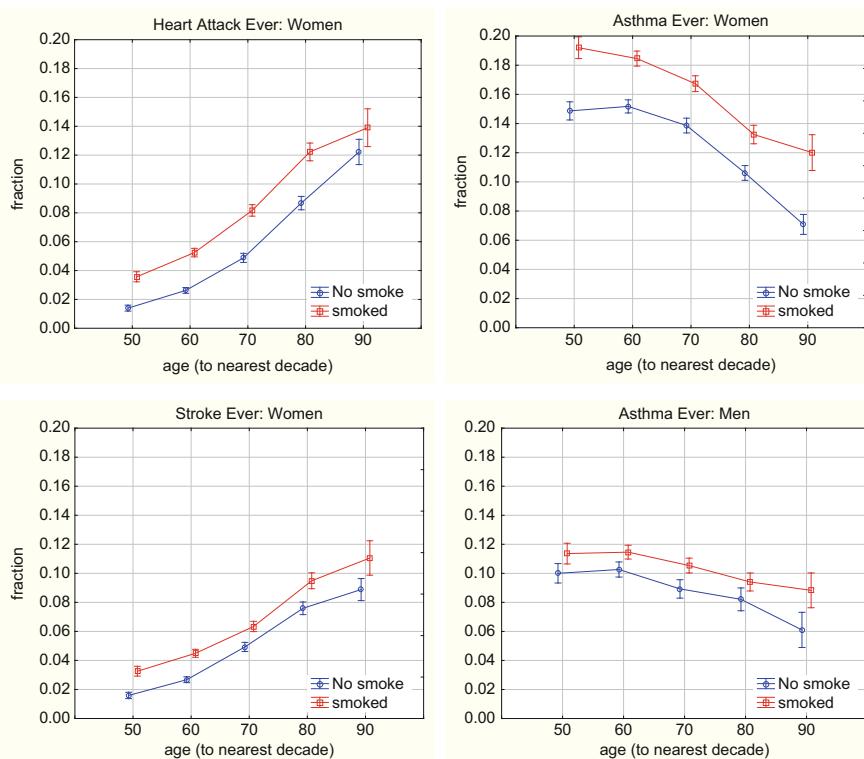


Fig. 3.2 Fraction of respondents reporting ever having had a heart attack (upper left panel), asthma (right panels) or stroke (lower left panel) vs. age (horizontal axes, age in years rounded to the nearest decade) for non-smokers (blue circles) and smokers (red squares). Vertical bars are 95% confidence intervals. Adverse health effects are greater among smokers than among non-smokers, for all age groups. These figures are for women except for the lower right panel, which shows asthma for male smokers and non-smokers

likely to have received an asthma diagnosis than older people. Given the marked effects of age and sex on disease rates, it is important to adjust for them in studying the effects of other variables on health outcomes.

Smoking Effects

Figure 3.2 shows how the fraction of female respondents reporting different health conditions varies with age and smoking status. As expected, smoking (here defined so that “No smoke” indicates fewer than 100 cigarettes (five packs) in a lifetime to date and “Smoked” indicates 100 cigarettes or more in a lifetime to date) is associated with increased risks of stroke, heart attack, and asthma at every age

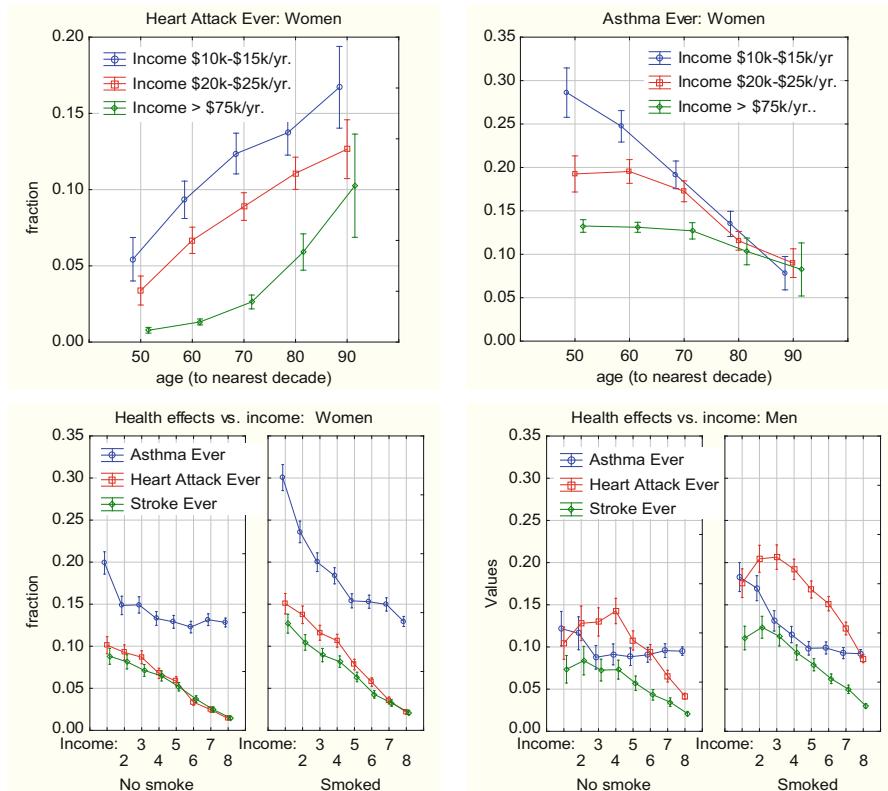


Fig. 3.3 UPPER PANELS: Fractions of women reporting ever having had a heart attack (upper left panel) or asthma (upper right panel) vs. age (horizontal axes, age in years rounded to the nearest decade) for low income (blue circles) medium income (red squares) and high income (small green circles). Vertical bars are 95% confidence intervals. LOWER PANELS: Asthma, heart attack, and stroke risks for men (lower right) and women (lower left) by income category (horizontal axes, details in text, 1 = lowest income, 8 = highest income) and smoking status (left half of panel for non smokers, right for smokers)

level. Similar effects hold for men, but heart attack risks are larger (and effects of smoking greater) and asthma risks are smaller (and effects of smoking smaller). The interaction plot for effects of smoking on age-specific asthma fraction in men is shown in the bottom right panel of Fig. 3.2, beneath the corresponding diagram for women in the upper right panel. In these and subsequent analyses, people over 95 are excluded, as they are relatively few (hence have wide confidence bands) and may have exceptionally low risks of heart attack and stroke (Fig. 3.1).

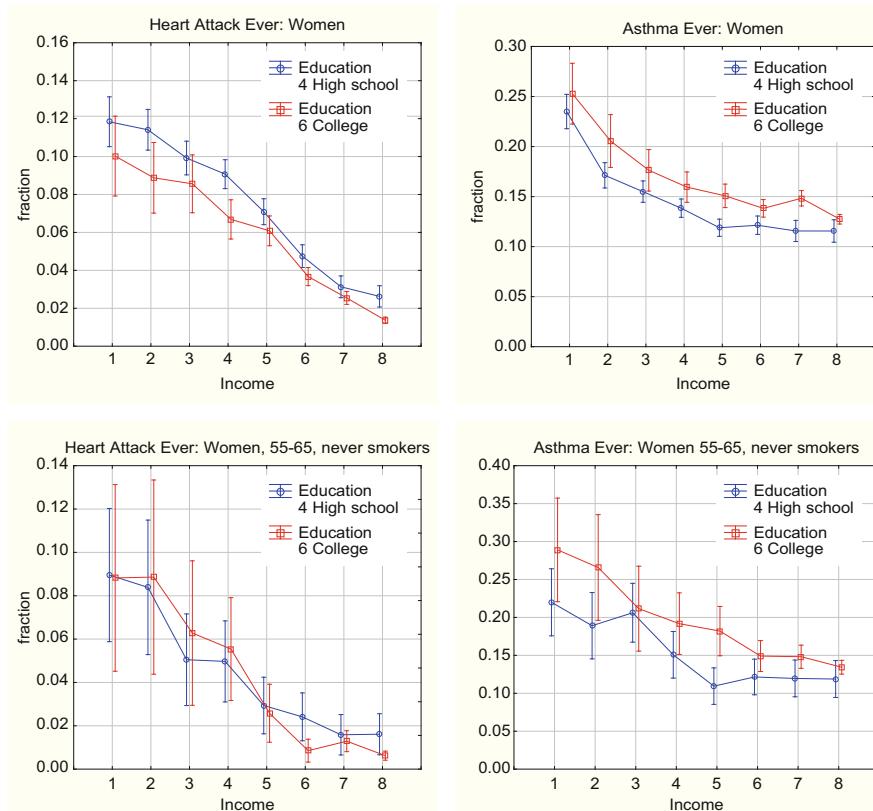


Fig. 3.4 Fractions of women reporting ever having had a heart attack (left) or asthma (right) vs. income category (horizontal axes, details in text, 1 = lowest income, 8 = highest income) and education level (blue circles for High school graduates, red squares for College graduates). The bottom panels are for younger (55–65 years old) non-smoking women. Vertical bars are 95% confidence intervals. College education is associated with lower heart attack risk but higher asthma risk for women than high school education at every level of annual income

Income Effects

Figure 3.3 shows striking effects of income on risks of adverse health effects. The top two panels show how age-specific fractions of women reporting different health effects vary with three different annual income levels: \$10k–\$15k (income code 2 on the survey); \$20k–\$25k (income code 4); and greater than \$75k (income code 8). For both heart attacks (upper left panel) and asthma (upper right panel), risks are much smaller for respondents with high incomes than for respondents with low incomes for age groups 50–80. This income effect is attenuated at older ages for asthma, although not for heart attack (or for stroke, not shown). The bottom two panels of Fig. 3.3 show all income levels across the horizontal axes and the fraction of

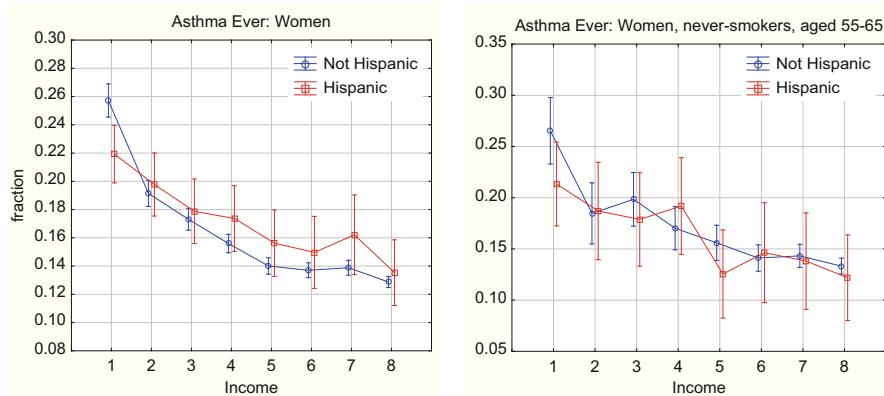


Fig. 3.5 Fractions of women reporting ever having had asthma vs. income category (horizontal axes, details in text, 1 = lowest income, 8 = highest income) and ethnicity (blue circles for not Hispanic, red squares for Hispanic). The left panel is for all women; the right panel is for younger (55–65 years old) non-smoking women. Vertical bars are 95% confidence intervals. The association between Hispanic ethnicity and asthma for all but the lowest income level (left panel) disappears after conditioning on age and smoking status (right panel)

respondents reporting all three health endpoints (heart attack, stroke, asthma) on the vertical axis, for smoking and non-smoking subpopulations (left and right sub-panels) for women and men (left and right lower panels, respectively). It is clear that the effects of higher income on dramatically reducing risks of all three health effects are greatest for smokers. The reduction in asthma risk associated with never smoking compared to ever smoking (i.e., no more than 100 cigarettes in a lifetime to date vs. more than 100 cigarettes to date) is large among respondents in the lower income categories, especially for women, but is much smaller or even non-existent at the highest income levels. This is in contrast to heart attack risk, which shows elevated risk among ever-smokers compared to never-smokers at all income levels. Thus, interactions among income, smoking, sex, and age are important for asthma risk. Table 3.4 shows health risks for different health effects risks for various combinations of these four factors, each at only two levels. The effects of smoking on asthma (0 = never, 1 = ever) are clearest at low incomes (0.19 vs. 0.29 for never- vs. ever-smoking women aged 55–65, i.e., in age decade 60, and in income code 2, \$10k–\$15k per year) but are negligible in income code 8 (>\$75k/year).

Effects of Education and Ethnicity

Education has a negative association with heart attack risk but a positive association with asthma risk for women (and men) at every income level as shown in the top row of Fig. 3.4. However, education is also associated with age (younger people are more

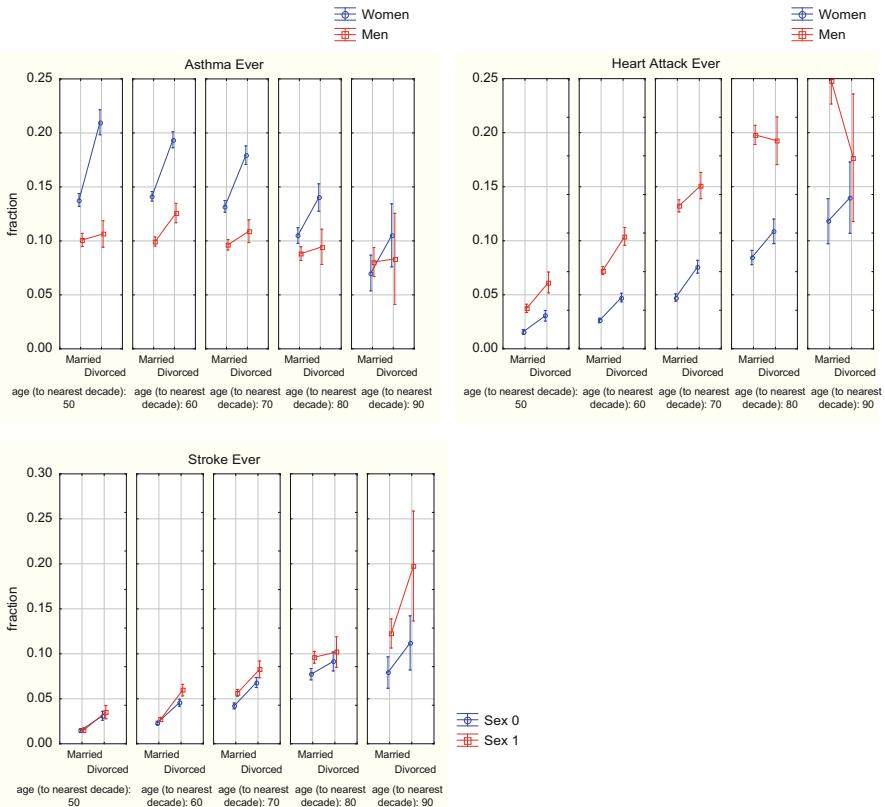


Fig. 3.6 Fractions of men (red squares) and women (blue circles) reporting ever having had asthma (upper left), heart attack (upper right) or stroke (lower left) vs. age (main horizontal axis) and married/divorced status (left and right sides). Vertical bars are 95% confidence intervals. Divorced status is associated with higher health risks than married status for both sexes and most ages

likely to have graduated from college), income (college education is associated with a nearly \$20k higher annual income than high school education at all ages), and smoking (women over 75 who are college graduates are more likely to be ever-smokers than those who are high school graduates; but women college graduates under 75 are less likely to be ever-smokers than those who are high school graduates, suggesting that being a college graduate used to be associated with smoking but has more recently become associated with not smoking). Thus, the effects of education on health risks are complicated by confounding due to age, smoking and income. The bottom row of Fig. 3.4 shows heart attack risks (left panel) and asthma risks (right panel) for different income levels and for high school and college graduates specifically for women never-smokers aged 55–65, thus controlling for possible confounding effects of age, smoking, and income by conditioning on specific levels for each. In the lower right diagram, college graduates are still more likely to report having been diagnosed with asthma than high school graduates at each income level,

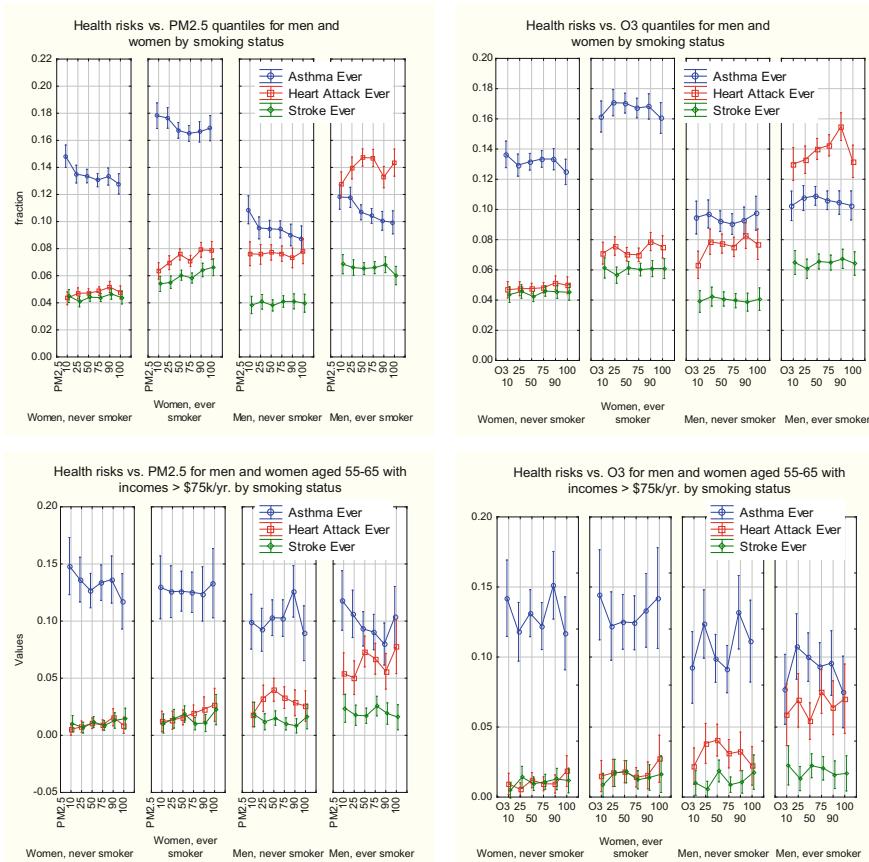


Fig. 3.7 Fractions of respondents reporting risks asthma (blue circles), heart attack (red squares) and strokes (small green circles) vs. PM_{2.5} percentile (left side) and O₃ percentile (right side) for male and female ever-smokers and never-smokers, with and without conditioning on age and income (top and bottom rows, respectively; bottom row is for people aged 55–65 in top income category). Vertical bars represent 95% confidence intervals

even after conditioning on age and never-smoked status. In the lower left diagram, however, there is no longer a clear effect of education on heart attack risk at all income levels. Similarly, although Hispanic ethnicity is associated with increased risk of asthma for women in different income groups, the association disappears after conditioning on age and smoking status (Hispanic women in this date set are more likely to be younger, and hence to have higher risk of asthma, than other women). Figure 3.5 illustrates this pattern. In summary, it appears that college education is associated with increased risk of reporting having been diagnosed with asthma, but Hispanic ethnicity *per se* is not, after conditioning on age and smoking.

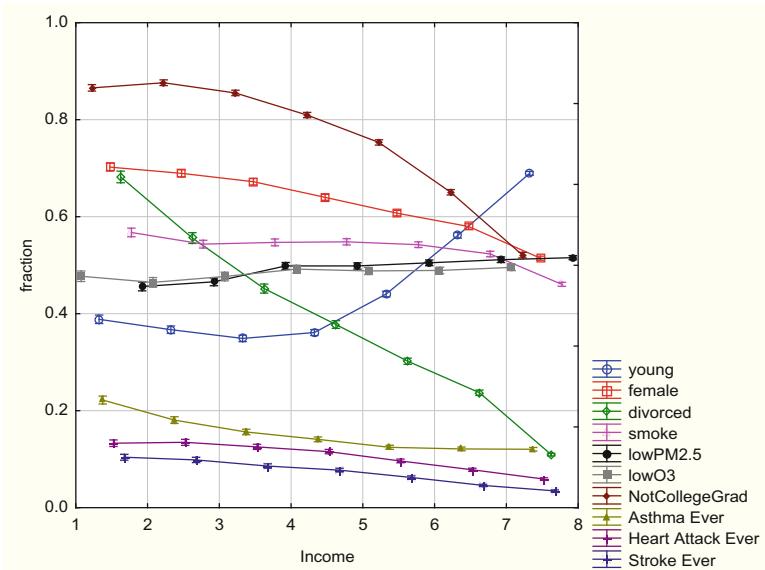


Fig. 3.8 Fractions of respondents reporting various attributes vs. income group (1 = lowest income, 8 = highest income, details in text). All factors vary with income level. Vertical bars indicating 95% confidence intervals are very narrow, reflecting large sample sizes

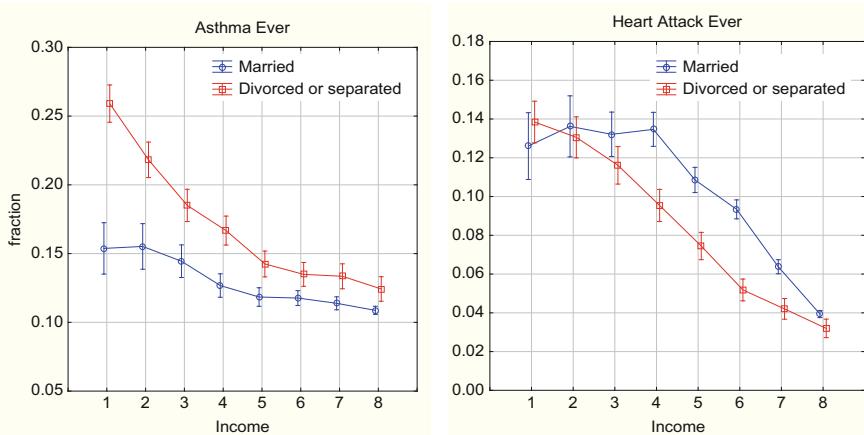


Fig. 3.9 Fractions of married (blue circles) or divorced/separated (red squares) respondents reporting asthma (left panel) or heart attack (right panel) vs. income group (1 = lowest income, 8 = highest income, details in text). Divorce or separation is positively associated with asthma risk at each income, but is negatively associated with heart attack risk at the upper income levels

Table 3.4 Health risks for different combinations of sex, age (rounded to nearest decade), income (2 = lower level, 8 = highest level), and smoking (0 = no, 1 = yes); the values of these variables are in the four left-most (shaded) columns

Sex	Age (to nearest decade)	Income	Smoking	Asthma ever means	Heart attack ever means	Stroke ever means	N
Women	60	2	0	0.19	0.07	0.07	923
Women	60	2	1	0.29	0.11	0.10	1238
Women	60	8	0	0.13	0.01	0.01	7684
Women	60	8	1	0.13	0.02	0.01	5557
Women	80	2	0	0.12	0.11	0.10	1168
Women	80	2	1	0.15	0.17	0.13	930
Women	80	8	0	0.09	0.06	0.06	768
Women	80	8	1	0.12	0.06	0.08	722
Men	60	2	0	0.13	0.11	0.07	363
Men	60	2	1	0.19	0.17	0.13	866
Men	60	8	0	0.10	0.03	0.01	5907
Men	60	8	1	0.10	0.07	0.02	5727
Men	80	2	0	0.08	0.18	0.13	169
Men	80	2	1	0.10	0.26	0.12	419
Men	80	8	0	0.08	0.12	0.07	878
Men	80	8	1	0.09	0.18	0.07	1304
All groups				0.12	0.06	0.04	34,623

Each row represents one such combination. The fractions of respondents in each row reporting ever being diagnosed with asthma, heart attack, or stroke, and the total number of respondents in each row, are shown in the four right-most columns, respectively

Effects of Marital Status

Figure 3.6 shows that divorced status is associated with higher health risks of asthma, heart attack, and stroke than married status. Divorce is most strongly associated with asthma for women; the effect is much less for men (upper left panel of Fig. 3.6). The effect of divorce on heart attack risks is greater than the effect on asthma for men 50–75 years old, but the effects are larger for asthma than for heart attacks (or stroke) for women.

Effects of Fine Particulate Matter (PM2.5) and Ozone (O3) Air Pollution

Table 3.5 shows the lower 10th percentile, lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile) and 90th percentile of the

Table 3.5 Percentiles of the frequency distributions for PM2.5 and O3 for all respondents

Variable	10th percentile	Lower quartile	Median	Upper quartile	90th percentile
PM2.5	5.6	7.0	8.9	11.4	13.95
O3	0.025	0.03	0.04	0.047	0.05

frequency distributions of PM2.5 and O3 for the BRFSS respondents for all 140 counties and 60 months covered in the data set, where each respondent is assigned the average daily PM2.5 and O3 values recorded by the EPA for the county of residence during the year and month of the interview. Monthly averages of daily PM2.5 concentrations ranged mainly between 5 and 14 $\mu\text{g}/\text{m}^3$, and O3 concentrations ranged mainly between 0.025 and 0.05 ppm. The following figures use the codes 10 (meaning values less than or equal to the 10th percentile), 25 (between the 10th and 25th percentile), 50, 75, 90, and 100 (meaning between the 90th and 100th percentile).

If air pollution levels have enough random variation over time so that pollution levels at the time of the interview contain no useful information about past pollution levels relevant for calculating associations with health effects, then associations between contemporaneous pollution levels and self-reported health conditions would not be meaningful. However, calculating the Spearman's rank correlation coefficients between the rankings of all 140 counties by average PM2.5 concentrations for all 5 years 2008–2012 reveals that all ten of the Spearman's rank correlations are positive (each of 5 years paired with each of 4 other years yields $20/2 = 10$ associations), meaning that counties that rank relatively high in PM2.5 1 year are likely to do so in other years; thus, the pollution levels at the time of the interview have the potential to be informative about levels over multi-year time spans.

Figure 3.7 plots adverse health effect prevalence against the quantiles of the PM2.5 (left side) and O3 (right side) frequency distributions for men and women ever-smokers and never-smokers. The upper row shows the fractions of individual respondents in each of these groups reporting each health condition (heart attack, stroke, asthma) at each exposure concentration quantile (corresponding concentrations are shown in Table 3.5). The lower row repeats this exercise specifically for the relatively large subpopulation of respondents aged 55–65 (age group 60, to the nearest decade) with incomes over \$75k/year. Conditioning on age and income helps to control for any potential confounding effects of these variables. For PM2.5, it appears in the diagram in the upper left quadrant of Fig. 3.6 that individual respondents from counties with relatively high ambient concentrations of PM2.5 (right ends of the asthma-vs.-PM2.5 curves) have significantly *lower* risks of asthma than individuals from counties with relatively low ambient concentrations (left ends of these curves), but that smokers (both men and women) have higher risks of heart attack when ambient PM2.5 is high than when it is low. Significant associations between ozone and asthma are not apparent. The interaction plots in the lower left quadrant of Fig. 3.7 suggest that the positive association between PM2.5 and heart attack risk and the negative association between PM2.5 and asthma risk are not explained away by confounding by age or income.

Table 3.6 Comparison of health effects prevalence in respondents aged 55–65 from counties with less than $5.9 \mu\text{g}/\text{m}^3$ (PM2.5 quantile = 10) or more than $13.95 \mu\text{g}/\text{m}^3$ (PM2.5 quantile = 100) PM2.5, matched for sex, income, and smoking

Sex	Income	Smoking	PM 2.5 quantiles	Asthma ever means	Heart attack ever means	Stroke ever means	N
F	2	0	10	0.25	0.01	0.07	72
F	2	0	100	0.18	0.05	0.05	104
F	2	1	10	0.35	0.06	0.07	103
F	2	1	100	0.28	0.10	0.07	116
F	8	0	10	0.15	0.01	0.01	780
F	8	0	100	0.12	0.01	0.01	682
F	8	1	10	0.13	0.01	0.01	579
F	8	1	100	0.13	0.03	0.02	488
M	2	0	10	0.16	0.12	0.04	26
M	2	0	100	0.24	0.10	0.05	21
M	2	1	10	0.23	0.21	0.15	73
M	2	1	100	0.15	0.18	0.16	99
M	8	0	10	0.10	0.02	0.02	604
M	8	0	100	0.09	0.03	0.02	548
M	8	1	10	0.12	0.05	0.02	592
M	8	1	100	0.10	0.08	0.02	503
All groups				0.13	0.03	0.02	5390

The four left-most (shaded) columns show different combinations of sex, income (2 = lower level, 8 = highest level), smoking (0 = no, 1 = yes) and PM2.5 (10 = bottom 10th percentile, 100 = top 10th percentile). Each row represents one such combination. The fractions of respondents in each row reporting ever being diagnosed with asthma, heart attack, or stroke, and the total number of respondents in each row, are shown in the four right-most columns, respectively

To obtain a more quantitative description of how health effects differ by PM2.5 concentration after controlling for age, sex, income, and smoking history, Table 3.6 tabulates the mean prevalence of each health condition for all combinations of each of two levels of sex, income, smoking, and PM2.5, for respondents aged 55–65 (the largest age group). While even this large data set has limited power to resolve differences when so finely partitioned, a standard binomial test rejects the null hypothesis that the proportions of respondents reporting adverse health conditions are independent of PM2.5 level in favor of the alternative hypotheses that heart attack rates are elevated at the high compared to the low level of PM2.5 ($p = 0.009$) for women ever-smokers with high incomes; and asthma rates are reduced at the high compared to the low level of PM2.5 for women never-smokers with high incomes ($p = 0.0475$). While heart attack rates are also elevated and asthma rates reduced in other groups at high compared to low PM2.5 levels—most notably, for women with low income, both ever-smokers and never-smokers—the sample sizes (N in the right-most column of Table 3.6) are too small to make these individual differences statistically significant at the conventional 0.05 level (e.g., $p = 0.13$ for the difference in asthma prevalence, 0.25 vs. 0.18, and $p = 0.07$ for the difference in heart

attack prevalence, 0.05 vs. 0.01, among female never-smokers with low income). Nonetheless, the pooled data for all of the larger groups supports the conclusion that the higher PM2.5 levels are associated with increased risk of heart attack and decreased risk of asthma.

Quantitatively, the largest reductions in asthma risk (from 0.25 to 0.18 for never-smokers and from 0.35 to 0.28 for ever-smokers) occur among women with low incomes. By contrast, at high income levels, the corresponding changes are small to negligible (from 0.15 to 0.12 for never-smokers and from 0.13 to 0.13 for ever-smokers, respectively). Thus, high income appears to greatly reduce or eliminate the association between PM2.5 and reduced asthma risk, as well as reducing the absolute risk of asthma and the positive association of smoking with asthma. From this perspective, income appears to be of central importance not only for asthma risk, as shown in Fig. 3.4, but also for modifying (specifically, reducing) the associations between other factors (smoking and PM2.5) and asthma risks.

Additional Interaction Analyses

To help visualize interactions among risk factors, we created the following dichotomous risk factors based on the preceding analyses: *young* = 1 for age less than 64 years (the median age), else 0 for respondents 64 or older; *female* = 0 for men, 1 for women; *lowIncome* = 1 for respondents with incomes below median (6 on the 1–8 scale), else 0 for respondents at or above the median income; *NotCollegeGrad* = 1 for *Education* less than 6 (graduated from college), else 0; *divorced* = 0 for married, 1 for separated or divorced; *smoke* = 1 for ever-smokers, 0 for never-smokers, *lowPM2.5* = 1 for $\text{PM2.5} < 8.91 \mu\text{g}/\text{m}^3$ (the median level of PM2.5), else 0 for greater values of PM2.5; and *lowO3* = 1 for $\text{O}_3 < 0.04 \text{ ppm}$ (the median level of O₃), else 0 for greater levels of O₃. Figure 3.8 shows how these dichotomous risk factors (other than income) and the three health condition indicators (bottom three curves, for asthma, heart attack, and stroke) vary with income levels 1–8. Other risk factors are strongly (although not always linearly) associated with income and with each other; most extremely, probability of being divorced or separated is about seven times greater for respondents in the lowest income levels compared to respondents in the top income group (level8, $>\$75k$ per year, constituting the top quartile of the income distribution), and is even higher for ever-smokers than for never-smokers in every income category (not shown). However, despite this very strong association between low income and high prevalence of divorce (or separation), the association between divorce and asthma is not explained away by confounding by low income. Asthma is significantly more prevalent among divorced or separated respondents than among married ones at every income level, as shown on the left side of Fig. 3.9, with the effect being greatest at lower income levels. For heart attacks, however, risk is lower among separated/divorced respondents at levels of 4 or more (and this is true for each sex separately, although that is not shown in Fig. 3.9).

Results of Logistic Regression Analysis

The results presented so far have not fit specific parametric models to the data. Table 3.7 shows the odds ratios for different risk factors in logistic regression models for *Asthma* and *Heart Attack* risks (developed using the R script `LR <= glm (Asthma ~ young + female + lowIncome + divorced + NotCollegeGrad + Smoking + PM2.5 + O3, family = binomial(link = 'logit')); exp(cbind(OR = coef(LR), confint(LR)))`; `summary(LR)`). By default, the lowest level of each categorical variable is used as the reference level.) These logistic regression models reinforce many of the findings from the previous interaction plots, but focus on main effects (one coefficient for each risk factor), and hence are less informative about nonlinearities and interactions among predictors.

Being divorced or separated is confirmed as a highly significant predictor of asthma risk, but not of heart attack risk. Not being a college graduate is a highly significant predictor of heart attack risk, but not of asthma risk. Smoking is a risk factor for both; ozone has a borderline significant positive association with heart attack risk (as suggested by the lower right panel of Fig. 3.7) but not with asthma risk; and PM2.5 has a significant negative association with asthma risk but no significant association (after conditioning on the other risk factors) with heart attack risk. (PM2.5 and O3 are continuous variables, unlike the other risk factors; hence the odds ratio of 0.99 for PM2.5 and asthma simply indicates that an increase of 1 $\mu\text{g}/\text{m}^3$ in PM2.5 only slightly reduces risk of asthma. For a 10 $\mu\text{g}/\text{m}^3$ increase in PM2.5, the odds ratio would be 0.94, i.e., about a 6% reduction in the probability of asthma.) The heart attack odds ratio of 11.3 for ozone is startlingly high. However, when continuous risk factors such as age and income are no longer dichotomized, but all values of all predictors are used, the odds ratio for O3 as a predictor of heart attack risk falls to 0.28 (95% CI 0.016–4.9, p value = 0.38), suggesting that the high odds ratio with dichotomized variables may be due to residual confounding or issues caused by dichotomizing continuous predictors. Income, age, sex, education, and smoking all remain highly significant predictors. The odds ratio for PM2.5 is 0.99 (95% CI 0.98–1.00, p value 0.07).

Table 3.7 Logistic regression odds ratios, 95% confidence intervals [in brackets], and *p*-values for various risk factors (left column) as predictors of Asthma (middle column) and Heart Attack (right column) risks

Risk factor	Asthma odds ratio (OR) [95% confidence limits], p value	Heart attack OR [95% confidence limits], p value
Young	1.25 [1.20, 1.3], $p < 2E - 16$	0.44 [0.41, 0.46], $p < 2E - 16$
Female	1.54 [1.48, 1.60], $p < 2E - 16$	0.40 [0.37, 0.42], $p < 2E - 16$
lowIncome	1.36 [1.30, 1.42], $p < 2E - 16$	2.0 [1.89, 2.12], $p < 2E - 16$
Divorced	1.25 [1.20, 1.31], $p < 2E - 16$	0.99 [0.93, 1.05], $p = 0.77$
NotCollegeGrad	1.01 [0.97, 1.05], $p = 0.70$	1.41 [1.33, 1.49], $p < 2E - 16$
Smoking	1.19 [1.14, 1.24], $p < 2E - 16$	1.74 [1.65, 1.84], $p < 2E - 16$
PM2.5	0.99 [0.98, 0.997], $p = 0.003$	1.00 [0.99, 1.01], $p = 0.58$
O3	1.66 [0.26, 10.6], $p = 0.59$	11.3 [0.97, 131.8], $p = 0.053$

The results for *Stroke* (not shown) are easily summarized: being young or female are associated with significantly reduced risks of stroke (ORs of 0.47 and 0.79, respectively); being low income, divorced, an ever-smoker, or not a college graduate are all associated with significantly increased risk of stroke (ORs of 2.15, 1.18, 1.36, and 1.35, respectively); and neither PM2.5 nor O3 is significantly associated with risk of stroke (OR of 1.00 for PM2.5 with 95% CI from 0.99 to 1.01; and OR of 2.0 for O3 with 95% CI from 0.10 to 42) after controlling for the other variables in Table 3.7.

Dichotomizing continuous or ordered categorical variables risks distorting regression results. For this data set, however, the key findings just summarized are robust to alternative model specifications. For example, treating *Age* as a continuous variable (or using ordered categories for age to the nearest decade), treating *Income* on a scale from 1 to 8 as either continuous or categorical, and treating *Marital Status* and *Education* as categorical variables, [e.g., using the R model specification $\text{LR} \leq \text{glm}(\text{Asthma} \sim \text{Age} + \text{female} + \text{as.factor(}\text{Income}\text{)} + \text{as.factor(}\text{Marital}\text{)} + \text{as.factor(}\text{Education}\text{)} + \text{Smoking} + \text{PM2.5} + \text{O3}, \text{family} = \text{binomial(link} = \text{'logit')})]$, does not change the conclusions that asthma risk decreases significantly with age, income, and PM2.5 concentration and increases significantly with smoking and being female. Ozone has no significant association with asthma risk. If education is entered as a continuous variable (on a scale from 1–6), then there is a significant positive association (logistic regression coefficient) between education and asthma risk (OR = 1.03, 95% CI = [1.01, 1.05], $p = 0.002$), consistent with the discussion of Fig. 3.4. Various interaction terms (not shown in the main-effects model in Table 3.7) are also significant (e.g., the female: smoking interaction for asthma has an odds ratio of 1.09, $p = 0.04$), consistent with Fig. 3.7 and previous figures.

Results of Bayesian Network and Partial Correlation Analysis

Regression models treat the different variables in a data set asymmetrically by distinguishing a single dependent variable which is to be explained or predicted from the values of one or more independent variables. By contrast, Bayesian Network (BN) models show how different variables can be used to help explain each others' values, treating each variable as both a potential explanatory variable and a potential dependent variable in relation to the rest. (See Chap. 2 for a fuller discussion.) Figure 3.10 shows the graph structures of a BN learned from the data using the default settings of the *bnlearn* package in R. The arrows in Fig. 3.10 indicate statistical dependency relations, and should not necessarily be interpreted causally. An arrow between two variables can be read as “is not independent of” (or, more briefly, as “depends on,” if it is understood that statistical dependence can go in either direction). Thus, arrows only show that the values of some variables depend on (i.e., are not statistically independent of) the values of other—a necessary but not sufficient condition for causality. In many cases, as in the relation between age and sex, or either of these and smoking, the arrows can be directed either way (via Bayes’

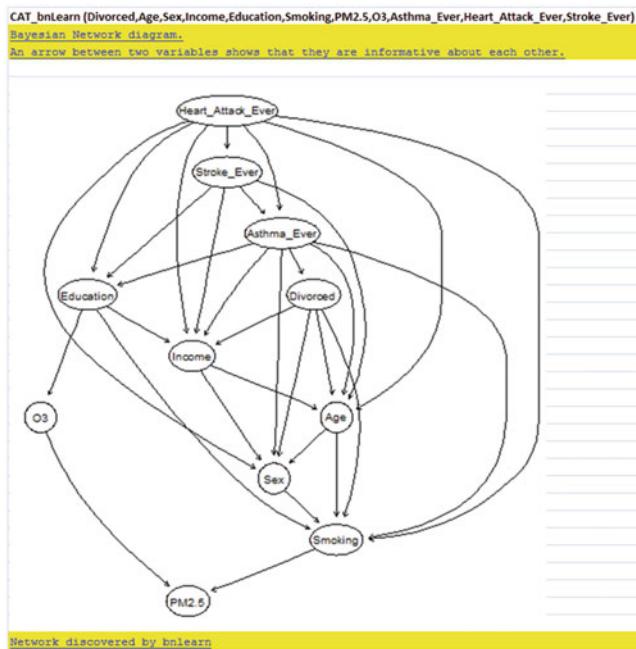


Fig. 3.10 Bayesian Network(BN) showing dependence relations among variables, generated using the *bnlearn* package in R. An arrow between two variables indicates that they are informative about each other: the frequency distribution of each variable varies significantly with the values of the variables to which it is connected, but, given those, it is conditionally independent of the values of variables to which it is not connected

rule) to correctly show that these variables have statistically dependent values, but this does not imply any clear causal interpretation. (Technically, the directions of arrows in a BN indicate one way to decompose the joint distribution of all variables into a directed acyclic graph (DAG) model with marginal distributions at input nodes (those with only outward-pointing arrows) and conditional probability distributions at other nodes.) The only implication for causality is that direct causes of a variable must be neighbors of it such a DAG, i.e., linked to it by an arrow (in either direction), in order to satisfy the information condition constraint that effects depend on—and hence are not conditionally independent of—their causes.

The structure of the BN in Fig. 3.10 shows that asthma and heart attack risks are directly linked to each other (having any of *Asthma_Ever*, *Heart_Attack_Ever*, or *Stroke_Ever* makes the others more likely) and also to smoking, sex, age, education, and income, but not to PM2.5 or O3. Insofar as health outcomes are conditionally independent of exposure variables PM2.5 and O3, given the values of other variables such as smoking and education (which also depend on each other), this data set does not support a causal interpretation for exposure-response associations, but rather explains them in terms of mutual dependence of exposure and response variables on other (confounding) variables. For example, higher education levels (college and

above) are associated with lower ozone and PM2.5 exposure concentrations, and also with higher incomes, and lower heart attack risks than lower levels of education, but this is not because lower ozone and PM2.5 directly cause these beneficial outcomes, as they are not adjacent to them in the DAG model.

Figure 3.10 in conjunction with the signs of different dependencies also suggests an explanation for the negative association between PM2.5 and asthma in regression modeling. As measured by partial correlation coefficients (i.e., correlation coefficients for each pair of variables after adjusting for others by multiple linear regression), shown in Table 3.8, PM2.5 exposure is negatively associated with smoking, and smoking is positively associated with asthma; also, PM2.5 (and O₃) are negatively associated with education, and education is positively associated with asthma. Since the only paths connecting PM2.5 and asthma in Fig. 3.10 involve smoking and education, and the associations are negative along each path, the total association between PM2.5 and asthma is negative. (The magnitudes of many of these correlation coefficients are attenuated by the use of binary variables such as *Smoking*, *Sex*, and *Asthma_Ever*, but they are still significantly different from zero.)

Results of Regression Tree and Random Forest Analyses

A sense of the magnitudes of the multivariate dependencies and interactions among variables can be gained from nonparametric regression tree analyses, the outputs of which are shown in Fig. 3.11 for the three adverse health effects, *Asthma_Ever* (upper left), *Heart_Attack_Ever* (upper right), and *Stroke_Ever* (bottom). Including all of the relevant variables (especially age, income, and education, which all have many levels) makes the trees too crowded to read, but these partial trees suffice to illustrate that heart attack risk and stroke risk are conditionally independent of PM2.5 given the values of other variables, while asthma has a negative association with PM2.5 for men (the *Sex > 0* branches of the trees). Each regression tree shows combinations of value ranges for the explanatory variables that lead to significantly different conditional distributions for the dependent variable. The trees are read as follows. Each shaded “leaf” node at the bottom of the tree shows the conditional mean value of the dependent variable, given the ranges of values for the variables in the path leading to that leaf, e.g., asthma risk is $y = 0.129$ among the 16,983 women with *Smoking* = 0 (for nonsmokers, represented in the tree as *Smoking* ≤ 0) and *Divorced* = 0 (for non-divorced women, coded as *Divorced* ≤ 0) for the left-most node in the tree for *Asthma_Ever* (upper left). This compares to a risk of 0.20 among the 4528 women who are divorced smokers. Non-leaf nodes show the *p* values from *F* tests for rejecting the null hypothesis that the conditional distributions are not different from each other on the left and right of each split.

Growing regression trees on different random subsets of the data can produce different trees. A more robust approach averages the results of ensembles of hundreds of trees fit to random subsets of the data to predict how the dependent varies as a single independent variable is varied; this generates a partial dependence plot.

Table 3.8 Partial correlation coefficients (top) between pairs of variables (rows and columns) after adjusting for all other variables via multiple linear regression, and their p -values (bottom)

Partial correlations		Outputs are: Estimated partial correlations, p-values										
		Age	Sex	Income	Education	Smoking	PM2.5	O3	Asthma_Ever	Heart_Attack_Ever	Stroke_Ever	Divorced
Age	1.00000	-0.0252	-0.1955	0.0108	0.0182	-0.00097	0.0029	-0.0713	0.1152	0.07784	-0.13404	
Sex	-0.02524	1.0000	0.0976	0.0168	0.0162	0.00380	0.0077	-0.0695	0.1096	0.00749	-0.05086	
Income	-0.19552	0.0976	1.0000	0.4729	-0.0034	-0.00245	-0.0101	-0.0567	-0.0673	-0.06152	-0.20353	
Education	0.01081	0.0168	0.4729	1.0000	-0.0492	-0.00979	-0.0174	0.0094	-0.0198	-0.00264	0.09719	
Smoking	0.01818	0.1062	-0.0034	-0.0492	1.0000	-0.03484	-0.0032	0.0329	0.0578	0.01876	0.07458	
PM2.5	-0.00097	0.0035	-0.0024	-0.0098	-0.0348	1.00000	0.0252	-0.0078	-0.0073	0.00463	0.00737	
O3	0.00286	0.0077	-0.0101	-0.0174	-0.0032	0.02525	1.0000	-0.0017	-0.0035	-0.00227	0.00707	
Asthma_Ever	-0.07132	-0.0695	-0.0567	0.0094	0.0329	-0.00783	-0.0017	1.0000	0.0399	0.03553	0.01527	
Heart_Attack_Ever	0.11523	0.1096	-0.0673	-0.0198	0.0578	-0.00730	-0.0035	0.0399	1.0000	3.16353	-0.00570	
Stroke_Ever	0.07784	0.0075	-0.0615	-0.0026	0.0188	0.00463	-0.0023	0.0335	0.1688	1.00000	0.00032	
Divorced	-0.13404	-0.0509	-0.2035	0.0972	0.0746	0.00737	0.0071	0.0153	-0.0057	0.00032	1.00000	
p.value												
Age	0.0e + 00	1.0e - 10	0.0e + 00	5.7e - 03	3.3e - 06	8.0e - 01	4.6e - 01	1.2e - 74	1.7e - 192	1.3e - 88	2.6e - 26	
Sex	1.0e - 10	0.0e + 00	1.9e - 138	1.8e - 05	1.5e - 163	3.3e - 01	4.9e - 02	6.2e - 71	3.9e - 174	5.5e - 02	8.7e - 3'	
Income	0.0e + 00	1.9e - 138	0.0e + 00	0.0e + 00	3.8e - 01	5.3e - 01	9.4e - 03	9.2e - 48	1.2e - 66	5.7e - 56	0.0e + 0	
Education	5.7e - 03	1.8e - 05	0.0e + 00	0.0e + 00	2.3e - 36	1.2e - 02	8.2e - 06	1.6e - 02	4.3e - 07	5.0e - 01	2.9e - 13	
Smoking	3.3e - 06	1.5e - 163	3.8e - 01	2.3e - 36	0.0e + 00	4.6e - 19	4.2e - 01	4.0e - 17	1.2e - 49	1.6e - 06	1.8e - 8:	
PM2.5	8.0e - 01	5.3e - 01	1.2e - 02	4.0e - 19	0.0e + 00	1.0e - 10	4.5e - 02	6.2e - 02	2.4e - 01	5.9e - 0:		
O3	4.6e - 01	4.9e - 02	8.2e - 06	4.2e - 01	1.0e - 10	0.0e + 00	6.6e - 01	3.8e - 01	5.6e - 01	7.0e - 0:		
Asthma_Ever	4.2e - 74	6.2e - 71	9.2e - 48	1.6e - 02	4.0e - 17	4.5e - 02	6.6e - 01	0.0e + 00	1.5e - 24	9.1e - 18	9.2e - 0	
Heart_Attack_Ever	1.7e - 192	3.9e - 174	1.2e - 66	4.3e - 07	1.2e - 49	6.2e - 02	3.8e - 01	1.5e - 24	0.0e + 00	1.4e - 0:		
Stroke_Ever	1.3e - 83	5.5e - 02	5.7e - 56	5.0e - 01	1.6e - 06	2.4e - 01	5.6e - 011	9.1e - 18	0.0e + 00	9.3e - 0:		
Divorced	2.6e - 260	8.7e - 39	0.0e + 00	2.9e - 137	1.8e - 81	5.9e - 02	7.0e - 02	9.2e - 05	1.4e - 05	9.3e - 01	0.0e + 0	

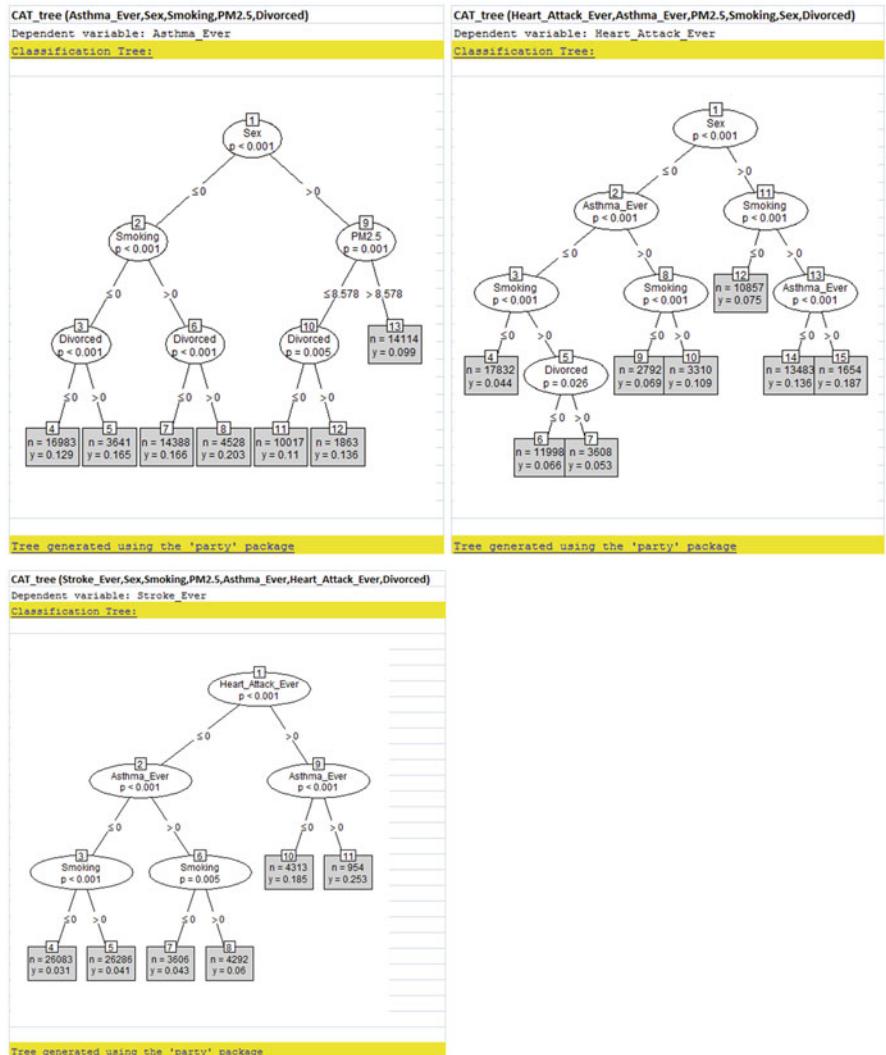


Fig. 3.11 Regression trees for adverse health effects (asthma in the upper left tree, heart attack in upper right, stroke at bottom. For binary variables, $>0 = \text{yes}$, $\leq 0 = \text{no}$.) Each tree shows how the fraction of respondents reporting each condition (the y values in the shaded leaf nodes) depend on the conditions specified by the paths leading to the leaf nodes. The number of respondents (the numbers n in the shaded leaf nodes), p values for tests of significantly different y values at intermediate nodes, and conditions defining each branch, are also shown. See text for further explanation and interpretation

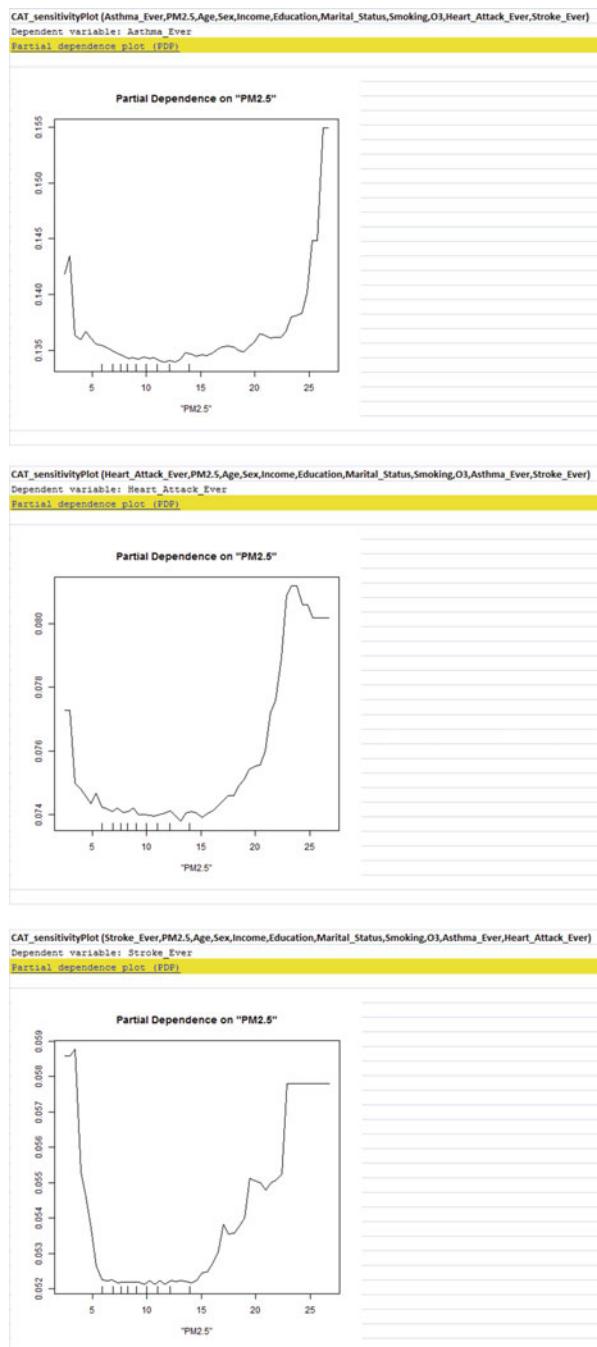
Figures 3.10 and 3.11 and Tables 3.7 and 3.8 show that logistic regression, Bayesian Networks, regression trees, and partial correlation analyses all found no significant positive association between PM2.5 (or O₃) and any of the three adverse health effects after conditioning on other variables, despite significant unconditional correlations created by confounders such as income. However, one can still generate partial dependence plots to estimate any residual associations. Because of interactions and residual confounding (e.g., due to differences in incomes or education or smoking within a single discrete coding level of these categorical variables), such residual patterns are expected. Quantifying their sizes puts a bound on the plausible magnitude of any undetected causal relations that might contribute to them. Accordingly, Fig. 3.12 shows partial dependence plots for each health effect vs. PM2.5 (generated by the *randomForest* package with a default ensemble size of 500 trees. A 10% simple random sample of all 142,081 records with complete data for these variables was used, to satisfy the memory requirements for the package.) After conditioning on levels of other variables (via trees), disease probabilities vary only slightly as PM2.5 changes. *Asthma_Ever* has values in the range 0.145 ± 0.01 , *Heart_Attack_Ever* has values in the range 0.077 ± 0.003 , and *Stroke_Ever* has values in the range 0.0055 ± 0.0004 as PM2.5 concentrations range from less than 5 $\mu\text{g}/\text{m}^3$ to over 60 $\mu\text{g}/\text{m}^3$. The plots suggest U-shaped residual associations between PM2.5 and disease risks, but whether these slight variations reflect causal impacts, residual confounding, selection biases, or random noise is unclear. In any event, an increase in PM2.5 concentration across its range by more than 50 $\mu\text{g}/\text{m}^3$ changes corresponding health risks relatively little, if at all.

Discussion

Many of the findings from the data set examined here are unsurprising in light of recent literature, but there are some exceptions. Low socioeconomic status has previously been recognized as a risk factor for both heart disease (Carlsson et al. 2016) and asthma attributed to air pollution, although, in the latter case, the need for a larger data set to examine interactions, as in Figs. 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9, has also been recognized (Burte et al. 2016). Divorce has been identified as a risk factor for heart attacks (Dupre et al. 2015), but the link between divorce (or separation) and asthma risk (Fig. 3.9) has been less well recognized. The very strong association between higher income and reduced asthma risk in the BRFSS data set and the link between younger age and higher asthma risk in this data set have been commented on previously (Zahran and Bailey 2013).

Others of our findings are more surprising. Coogan et al. (2016) report a negative relation between education and adult-onset asthma in African American women, and this contrasts with our finding of a positive association between education and asthma risk in the general BRFSS population (Fig. 3.4). PM2.5 is usually significantly positively associated with heart attack risk, as might be expected based on the left side of Fig. 3.7 (e.g., To et al. 2015), and is sometimes associated with asthma

Fig. 3.12 Partial dependence plots (concentration-response curves) for the fractions of respondents reporting asthma (top panel), heart attack risk (middle panel), and stroke (bottom panel) for different levels of PM2.5, after conditioning on observed levels of other variables



risk (www.epa.gov/region4/sesd/pm25/p2.html). Indeed, a logistic regression model for heart attack risk as the dependent variable with only sex, age, smoking status, and PM2.5 included as independent variables does show a significant positive association between heart attack risk and PM2.5 (odds ratio of 1.09 per 10 $\mu\text{g}/\text{m}^3$ of PM2.5, $p = 0.026$). However, adding *lowIncome* as an independent variable removes this significant association between PM2.5 and heart attack risk disappears (OR = 1.00, [0.996, 1.01], $p = 0.28$). (Similarly, PM2.5 is positively associated with stroke risk ($p = 0.059$) when only age, sex, and smoking are included as covariates, but not when income is also included ($p = 0.56$).) That PM2.5 is significantly positively associated with heart attack risk when income is not included as a predictor, but not when it is included, suggests the importance of controlling for income, as well as other potential socioeconomic confounders, in assessing and interpreting exposure-response relations. As shown in Fig. 3.8, both lower PM2.5 and lower prevalence of adverse health effects occur at higher income levels; thus, the associations between them must control for income as a confounder in assessing exposure-health effect relations. This has not been done in some previous studies that reported significant positive associations between PM2.5 and health risks such as heart attack or stroke hazard rates (e.g., To et al. 2015).

The finding of a significant negative association between PM2.5 and asthma risk also appears to be mostly new (but see Krstić 2011) and, *a priori*, not very plausible. In a recent literature review, Heinzerling et al. (2016) reported positive associations between ultrafine particulate matter and asthma in children, but no significant effects in multivariate models that account for potential confounding by co-pollutants. Strickland et al. (2016) found a positive association between daily county-level pediatric emergency department visits for asthma or wheeze and same-day PM2.5 concentrations in Georgia, as estimated from satellite data, but without adjustment for temperature or other possible confounders. Chen et al. (2016) identify PM2.5 as one of several risk factors for both childhood and adult asthma in China. They note that China, with relatively high levels of particulate air pollution, has had lower asthma prevalence than in developed countries, but it has been increasing in recent decades. To our knowledge, a significant negative association between PM2.5 and asthma risk in regression models (and in the partial correlations in Table 3.8) after adjusting for income and other covariates is a new and unexpected finding that deserves further scrutiny in this and other large data sets.

Figure 3.10 shows that there are multiple paths of statistical dependencies leading from most other variables to asthma risk (at the bottom of the upper left panel of Fig. 3.10). For example, being female is a risk factor for asthma not only via a direct dependency link between them, but also via paths involving age, education, and income, e.g., because being female is associated with lower income, especially for divorcees, and lower income, in turn is a risk factor for asthma. These multiple paths serve as a reminder that any single regression coefficient or odds ratio linking a predictor to a health risk variable is likely to conflate direct and indirect effects. Such coefficients do not necessarily show how changing one variable would change another that depends on it. However, asthma risk is not conditionally independent of any of the variables that points into it in Fig. 3.10, given the values of the other

variables. This leaves open the possibility that some of these arrows could represent causal relations, such that changing the values of one or more of the variables pointing into *Asthma* would cause a change in asthma prevalence. To investigate this possibility further, it would be useful to examine how asthma prevalence changes when one or more of the variables pointing into it in Fig. 3.10 is changed, perhaps using longitudinal data from natural experiments; this is beyond the intended scope of the present investigation, however.

Study Limitations and Uncertainties

A limitation of this study is that it considers exposures and responses over a 5-year window, 2008–2012, which is too brief an interval to permit study of the temporal relationship between changes in pollutant exposure levels and changes in health responses that might take place on a time scale of decades. A panel data design in which exposures and health histories for the same individual are assessed repeatedly over time, e.g., annually, for decades could add important additional information to clarify the causal interpretations (or lack of causality) for associations identified in the foregoing analyses. Bayesian Networks only show which variable are informative about each other, i.e., statistical dependence and conditional independence relationships among variables, but not the direction or magnitude of information flows between causes and effects. Inferring these from longitudinal data requires relatively lengthy time series and additional methods, such as Granger causality, transfer entropy and directed information flow (Janzing et al. 2013). When only relatively short (e.g., 5-year) time series data are available, as in this study, comparative evaluations suggest that Bayesian Network methods out-perform longitudinal methods (Zou and Feng 2009), but it would be desirable to apply both to longer time series in future work.

A second limitation is that the variables for which data are available may not include all causally relevant factors. There may be important confounders and modifying factors that are not captured in the data set analyzed here. For example, Krstić (2011) identifies apparent temperature (a measure of air temperature, relative humidity and wind speed), geographical latitude of residence and vitamin D status (reflecting exposure to sunlight) as additional relevant variables for understanding associations between asthma prevalence and PM_{2.5} air pollution (PM_{2.5}) in 97 major metropolitan/micropolitan areas of the continental U.S. None of these variables has been included in the current study, suggesting that any causal interpretation of the associations reported here could be erroneous because the effects of other potentially important causes have been omitted.

More generally, the analytic methods used in this study to clarify potential causal interpretations of observed associations complement the basic idea of potential outcomes or counterfactual causal modeling models, that differences in causes make their effects differ from what they otherwise would have been, with the related idea that differences in causes help to predict differences in effects. Although this has

the advantage that it can be tested using observational data without the need to make hypothetical modeling assumptions about what would have been if exposures or other conditions had been different from those observed, it does not protect against the possibility that one variable might be informative about another due to the presence of unmeasured variables that affect or are affected by both. Nonparametric methods for estimating probabilistic dependency dependencies—specifically, algorithms for learning regression trees and Bayesian Networks (BNs) from data, which are now mature and widely available—also help to avoid potential biases due to model specification errors or incorrect modeling assumptions.

Conclusions

The picture of asthma risk that emerges from the preceding data analyses is largely a socioeconomic one. Young divorced women with low incomes are at greatest risk of asthma, especially if they are ever-smokers or have a history of heart attacks or strokes. Income is an important confounder of other relations. For example, in logistic regression modeling, PM2.5 is positively associated with both stroke risk and heart attack risk when these are regressed only against PM2.5, sex, age, and ever-smoking status, but not when they are regressed against these variables and income. Unexpectedly, PM2.5 is significantly negatively associated with asthma risk in multiple linear regression, logistic regression, and regression tree models, with a $10 \mu\text{g}/\text{m}^3$ decrease in PM2.5 corresponding to about a 6% increase in the probability of asthma in a logistic regression model. Whether this negative association is explained by confounders and residual confounding, as Fig. 3.10 suggests, or whether it has other explanations is a worthwhile topic for further investigation. Meanwhile, the data and analyses presented here suggest that substantially reducing the burden of adult asthma may require addressing the causal web of socioeconomic conditions leading to low incomes, smoking, and divorce, especially among women.

References

- Aliferis CE, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XS (2010) Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *J Mach Learn Res* 11:171–234
- Burte E, Nadif R, Jacquemin B (2016) Susceptibility factors relevant for the association between long-term air pollution exposure and incident asthma. *Curr Environ Health Rep* 3(1):23–39
- Carlsson AC, Li X, Holzmann MJ, Wändell P, Gasevic D, Sundquist J, Sundquist K (2016) Neighbourhood socioeconomic status and coronary heart disease in individuals between 40 and 50 years. *Heart* 102(10):775–782
- Chen Y, Wong GW, Li J (2016) Environmental exposure and genetic predisposition as risk factors for asthma in China. *Allergy Asthma Immunol Res* 8(2):92–100. <https://doi.org/10.4168/aair.2016.8.2.92>

- Coogan PF, Castro-Webb N, Yu J, O'Connor GT, Palmer JR, Rosenberg L (2016) Neighborhood and individual socioeconomic status and asthma incidence in African American women. *Ethn Dis* 26(1):113–122
- Dupre ME, George LK, Liu G, Peterson ED (2015) Association between divorce and risks for acute myocardial infarction. *Circ Cardiovasc Qual Outcomes* 8(3):244–251
- Frey L, Fisher D, Tsamardinos I, Aliferis CF, Statnikov A (2003) Identifying Markov blankets with decision tree induction. In: Proceedings of the third IEEE international conference on data mining, Melbourne, 19–22 Nov 2003, pp 59–66
- Furqan MS, Siyal MY (2016) Random forest Granger causality for detection of effective brain connectivity using high-dimensional data. *J Integr Neurosci* 15(1):55–66
- Guo Z, Haimes YY (2016) Risk assessment of infrastructure system of systems with precursor analysis. *Risk Anal* 36(8):1630–1643. <https://doi.org/10.1111/risa.12559>
- Halliday DM, Senik MH, Stevenson CW, Mason R (2016) Non-parametric directionality analysis—extension for removal of a single common predictor and application to time series. *J Neurosci Methods* 268:87–97
- Heinzerling A, Hsu J, Yip F (2016) Respiratory health effects of ultrafine particles in children: a literature review. *Water Air Soil Pollut* 227:32
- Hill J (2016) Atlantic causal inference conference competition: is your SATT where it's at? <http://jenniferhill7.wixsite.com/acic-2016/competition>
- Janzing D, Balduzzi D, Grosse-Wentrup M, Scholkopf B (2013) Quantifying causal influences. *Ann Stat* 41(5):2324–2358
- Krstić G (2011) Asthma prevalence associated with geographical latitude and regional insolation in the United States of America and Australia. *PLoS One* 6(4):e18492. <https://doi.org/10.1371/journal.pone.0018492>
- NIPS (Neural Information Processing Society) 2013 workshop on causality. <http://clopinet.com/isabelle/Projects/NIPS2013/>
- Oliveira A, Barros H, Maciel MJ, Lopes C (2007) Tobacco smoking and acute myocardial infarction in young adults: a population-based case-control study. *Prev Med* 44(4):311–316. Epub 2007 Jan 17
- Pearl J (2010) An introduction to causal inference. *Int J Biostat* 6(2):Article 7
- Shah RS, Cole JW (2010) Smoking and stroke: the more you smoke the more you stroke. *Expert Rev Cardiovasc Ther* 8(7):917–932
- Strickland MJ, Hao H, Hu X, Chang HH, Darroow LA, Liu Y (2016) Pediatric emergency visits and short-term changes in PM_{2.5} concentrations in the U.S. State of Georgia. *Environ Health Perspect* 124(5):690–696
- To T, Zhu J, Villeneuve PJ, Simatovic J, Feldman L, Gao C, Williams D, Chen H, Weichenthal S, Wall C, Miller AB (2015) Chronic disease prevalence in women and air pollution—a 30-year longitudinal cohort study. *Environ Int* 80:26–32
- Zahran HS, Bailey C (2013) Factors associated with asthma prevalence among racial and ethnic groups—United States, 2009–2010 behavioral risk factor surveillance system. *J Asthma* 50(6):583–589
- Zou C, Feng J (2009) Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics* 10:122. <https://doi.org/10.1186/1471-2105-10-122>

Chapter 4

Descriptive Analytics for Occupational Health: Is Benzene Metabolism in Exposed Workers More Efficient at Very Low Concentrations?



Introduction

The occupational risks to workers from noxious substances inhaled in air depend on the concentrations inhaled and on what happens to the inhaled substances—for example, whether they are swiftly detoxified and eliminated from the body without doing harm, or whether they are metabolized to form toxic concentrations of metabolites in target tissues. Descriptive analytics applied to data on inhaled concentrations and metabolites formed can be used to clarify how efficiently the body produces toxic metabolites at low exposure concentrations. This chapter applies descriptive analytics methods introduced in Chaps. 1–3, including interaction plots, nonparametric regression, CART trees, and Bayesian networks, to data on benzene metabolites in Chinese factory workers in an effort to resolve a recent puzzle in the literature on low dose benzene toxicology. For readers who do not care to pursue this topic further, we recommend quickly examining the figures to see how plots and visualizations of patterns in the data can be displayed and used to gain insight into the dependencies among variables.

Two apparently contradictory findings in the literature on low-dose human metabolism of benzene are that (1) Metabolism is approximately linear at low concentrations, e.g., below 10 ppm, consistent with decades of quantitative modeling of benzene pharmacokinetics and dose-dependent metabolism; and (2) Measured benzene exposure and metabolite concentrations for occupationally exposed benzene workers in Tianjin, China in a set of recent studies show that dose-specific metabolism (DSM) ratios of metabolite concentrations per ppm of benzene in air decrease steadily with benzene concentration, with the steepest decreases below 3 ppm. This has been interpreted as indicating that metabolism at low concentrations of benzene is highly nonlinear. This chapter reexamines the data using non-parametric methods of descriptive analytics and concludes that both findings are correct. They are not contradictory. Low-concentration metabolism can be linear,

with metabolite concentrations proportional to benzene concentrations in air, and yet DSM ratios can still decrease with benzene concentrations. The algebra of random variables allows a ratio to be negatively correlated with its denominator even when the mean of the numerator is proportional to the denominator. Toxicological interpretations of declining DSM ratios as evidence of nonlinear metabolism are unwarranted when direct plots of metabolite concentrations against benzene ppm in air show approximately straight-line relationships between them, as in the Tianjin data. Thus, relatively straightforward descriptive analytics can help to resolve what at first appears to be contradiction that has fueled heated discussions in the recent literature. Descriptive plots and analysis reveal that highly nonlinear, decreasing DSM ratios are consistent with linear metabolism.

Since the 1970s, it has been recognized that occupational inhalation exposures to hundreds of ppm of benzene for decades increase risks of acute myeloid leukemia (AML). For example, among Turkish shoe workers exposed prior to 1970, the increase in AML risk was estimated as roughly two- to fourfold, or about 1 excess case per 10,000 to 100,000 worker-lifetimes (Aksoy et al. 1974). More recently, Chinese workers occupationally exposed to benzene have been reported to have levels of benzene metabolites including phenol (PH), hydroquinone (HQ), catechol (CA), E,E-muconic acid (MA), and S-phenylmercapturic acid (SPMA), as well as levels of unmetabolized benzene excreted in urine, that are disproportionately elevated at very low estimated exposure concentrations such as 1 part per million (ppm) or less (Rappaport et al. 2010). Such findings naturally prompt curiosity and concern about exactly how great the increases in benzene metabolites are at such occupational exposure levels and whether they might be sufficient to create significantly increased health risks. The main purpose of this chapter is to apply non-parametric statistical methods to Chinese worker data from three factories in Tianjin to ascertain the relationship between relatively low levels of benzene exposure (<10 ppm) and resulting increases in benzene metabolites in workers exposed to these concentrations.

The literature on low-dose benzene metabolism and health effects has been marked by conspicuously opposing views, sometimes vehemently expressed (Rappaport et al. 2013; Price et al. 2013). Both regulation and litigation have drawn on and helped to fund scientific investigations of low-exposure benzene dose-response relationships in recent decades (Schirrmeyer and Flora 2008). Investigators funded by regulators such as the U.S. Environmental Protection Agency (EPA) and the Occupational Safety and Health Administration (OSHA), or by government institutes such as the National Institute for Environmental Health Sciences (NIEHS), the National Cancer Institute (NCI), and the National Institute for Occupational Safety and Health (NIOSH), and testifying on behalf of plaintiffs, often decry evidence of low-dose hazards where investigators funded by industry groups such as the American Petroleum Institute in the U.S. or CONCAWE in Europe, including the current author, do not. Published results are sensitive to modeling assumptions and to data interpretations of unknown validity, as discussed later. As a result, non-parametric methods that make minimal or no modeling assumptions may be especially valuable for benzene to help rise above motivated

reasoning and politicized science and interpretation of data (Thomas et al. 2014). The following sections therefore apply non-parametric methods and data-visualization to benzene data.

The remainder of this chapter is organized as follows. The next section provides background on theories and interpretations of low-exposure concentration toxicological and epidemiological data that have been used to argue for or against the hypothesis that benzene concentrations (ppm) in the range of single digits or less are harmful to human health. A data set on benzene and metabolites in workers from Tianjin, China is then analyzed to reexamine the low-concentration relationship between benzene exposure concentrations and metabolite levels using non-parametric methods.

Background: Theories and Controversies in Benzene Dose-Response

Different toxicological theories have been proposed for the potential low-dose effects of benzene on leukemia risks, with some supporting and others undermining the hypothesis that excess risks should be expected at exposure concentrations in the range of a few ppm or less. For over 30 years, it has been recognized that peripheral blood white blood cells, such as lymphocytes and mononuclear cells, are much more sensitive to benzene than are red blood cells and other cells in the myeloid lineage (e.g., Kipen et al. 1989). More recently, a team of researchers at the University of California at Berkeley applied modern biological research methods to benzene-exposed workers in China, reporting altered gene expression and enzyme activity in peripheral white blood cells at benzene concentrations as low as 0.1 ppm (McHale et al. 2011; Thomas et al. 2014). Somewhat confusingly, clusters of gene expression changes in peripheral blood mononuclear cells associated with benzene exposures have been termed “the acute myeloid leukemia (AML) pathway” (*ibid*), although they involve neither leukemia nor myeloid cells nor a pathway that connects them to AML. The biological relevance of these findings to AML is unknown, as they occur in peripheral blood cells of the lymphoid lineage rather than in the multipotent stem cells and myeloid precursor cells in the bone marrow that are thought to be target cells for AML (Walter et al. 2012).

Nonetheless, the possibility that relatively low concentrations of benzene that cause altered gene expression in peripheral white blood cells might also increase risk of AML and other leukemias has remained a topic of active research for the past decade (*ibid*). The possibility of hormesis—that relatively low levels of benzene exposure might decrease risk of AML—has been less investigated, but exposures to concentrations of less than 10 ppm of benzene have been reported to significantly reduce the clonal proliferation (colony formation potential) of myeloid progenitor cells isolated from peripheral blood (Lan et al. 2004), consistent with some theoretical models of hormesis (Cox 2006, 2009).

Other investigators have emphasized the absence of detected increases in leukemia risks at relatively low levels of benzene exposure. For example, Wong (1995) reexamined a much-studied occupational cohort in the U.S. (the Pliofilm cohort) that had often been interpreted as showing no evidence of an exposure threshold for increased leukemia risk as a function of benzene exposure and concluded that, to the contrary, “No increased risk of AML was detected for cumulative exposure to benzene below 200 ppm-years (SMR 0.91). Above 200 ppm-years, risk of AML rose drastically; reaching a significant SMR of 98.37 for >400 ppm-years.” Several investigators, including the International Agency for Research on Cancer (IARC) have suggested that occupational exposures to benzene might cause increased risk of multiple myeloma, acute lymphocytic leukemia, and chronic lymphocytic leukemia, based largely on interpreting causally associations in epidemiological studies that have been described as inconclusive (Vlaanderen et al. 2012). Weed (2010) critiqued the use of meta-analysis of associations to suggest that benzene might also cause non-Hodgkin’s lymphoma (Steinmaus et al. 2008), arguing that “Causal claims... should not emerge from meta-analyses as such” and that University of California, Berkeley investigators had “performed a meta-analysis and concluded that it represented new evidence that benzene causes NHL” in spite of “a lack of consistency (i.e., significant heterogeneity), weak associations, no evidence of dose-response, no effort to provide an assessment of biological plausibility, and no new epidemiological evidence.” He used this as a case study for critical discussion of the use and misuse of meta-analysis and causal inference in occupational epidemiology. In a reply defending their interpretation of meta-analysis of correlations as evidence of causation, Steinmaus et al. (2011) respond that “We have been teaching this for many years in our course at the University of California, Berkeley, School of Public Health, titled ‘Causal inference and meta-analysis’” and indeed much of the literature postulating or asserting increased health risks at relatively low levels of benzene concentration in the past two decades has flowed from this same research group at the University of California, Berkeley, School of Public Health (Steinmaus et al. 2008, 2011; Lan et al. 2004; Rappaport et al. 2009, 2010, 2013; McHale et al. 2011; Thomas et al. 2014).

Turning from possible health effects to possible causes, Rappaport et al. (2010) postulated a model with “a hitherto unrecognized high-affinity enzyme that was responsible for an estimated 73% of total urinary metabolites [sum of phenol (PH), hydroquinone (HQ), catechol (CA), E,E-muconic acid (MA), and S-phenylmercapturic acid (SPMA)] in nonsmoking females exposed to benzene at sub-saturating (ppb) air concentrations.” If true, this hypothesis would imply that an unidentified enzyme, which the authors estimated to be responsible for more than half of all benzene metabolism at 1 ppm and close to 20% at 10 ppm, had been overlooked during the past half century of research and quantitative modeling of benzene metabolism, including during the development of validated physiologically-based pharmacokinetic (PBPK) models of benzene metabolism in humans and rodents that have successfully fit and predicted observed data without postulating any such unidentified enzyme (e.g., Knutsen et al. 2013). The evidence offered to support the hypothesis consists of fitting two parametric regression curves

to data, one to represent a one-enzyme model and the other to represent a two-enzyme model (Rappaport et al. 2010). The authors did not claim that either model had been validated by showing that it correctly described benzene metabolism in humans, nor did either model represent measurement errors in its right-hand side variables (estimated concentration of benzene in air and estimated background levels of each metabolite). Thus, although the authors interpreted the results of this curve-fitting exercise as providing “strong statistical evidence favoring two metabolizing enzymes and indicated that the higher-affinity enzyme was responsible for about 73% of all benzene metabolism at non-saturating (ppb) air concentrations,” the evidence only shows that one curve-fitting model provides a worse fit to the data than the other. This does not imply that either of them is realistic or correct. It does not justify a conclusion that the comparison provides “extremely strong evidence favoring the better model as a depiction of the true metabolism of benzene to a particular metabolite,” or that the results “provide extremely strong statistical evidence that benzene is metabolized to PH and MA via two enzymes rather than one enzyme, and that the putative high-affinity enzyme is active primarily below 1 ppm” (Rappaport et al. 2010).

Earlier work by the same group had concluded that benzene oxide-albumin adducts from Chinese workers “indicate that deviations from linear metabolism began at or below benzene exposures of 10 ppm and that pronounced saturation was apparent at 40–50 ppm” (Rappaport et al. 2002), and that “Mean trends of dose-specific levels (micromol/L/ppm benzene) of E,E-muconic acid, phenol, hydroquinone, and catechol all decreased with increasing benzene exposure, with an overall ninefold reduction of total metabolites. . .[indicating] that benzene metabolism is highly nonlinear with increasing benzene exposure above 0.03 ppm, and that current human toxicokinetic models do not accurately predict benzene metabolism below 3 ppm” (Kim et al. 2006b). Similarly, summarizing these findings, Rappaport et al. (2009) note that “Intriguingly, the exposure-specific production of major metabolites (phenol, muconic acid, hydroquinone, and catechol, in micromolar per parts per million benzene) decreased continuously with estimated exposure levels over the range of 0.03–88.9 ppm, with the most pronounced decreases occurring at benzene concentrations <1 ppm.”

These and related articles by the Berkeley group suggest that low levels of exposure to benzene are disproportionately hazardous compared to higher levels. As expressed by Rappaport et al. 2009, “Because regulatory risk assessments have assumed nonsaturating metabolism of benzene in persons exposed to air concentrations well above 10 ppm, our findings suggest that the true leukemia risks could be substantially greater than currently thought at ambient levels of exposure—about threefold higher among nonsmoking females in the general population.” This line of reasoning, which projects leukemia risks directly from modeled levels of metabolites, has proved influential with regulators: “In justifying its decision to lower the benzene content of gasoline, the U.S. EPA cited studies pointing to supralinear (greater-than-proportional) production of benzene-related protein adducts at air concentrations <1 ppm (Rappaport et al. 2002, 2005). . . Because the U.S. EPA had previously assumed that human benzene metabolism proceeded according to

nonsaturating (first-order) kinetics at exposure concentrations well above 10 ppm, saturation of metabolism below 1 ppm ‘could lead to substantial underestimation of leukemia risks’ in the general population (U.S. EPA 2007)” (*ibid*).

Against this interpretation, Price et al. (2012) reexamined the Tianjin data and the modeling by Kim et al. (2006a, b) and concluded that based on “the impacts of technical issues in the corrections for background levels of metabolites, accounting for biases in the regression modeling, and the uncertainties introduced by the use of a calibration model to estimate benzene air levels for certain workers are evaluated and . . . the Tianjin data appear to be too uncertain to support any conclusions of a change in the efficiency of benzene metabolism with variations in exposure.” In particular, Price et al. highlighted that “Defining background levels as either the levels in all workers with no occupational exposures or in workers with predicted air levels of <0.03 p.p.m. results in estimates of 2.4 fold [$<0.1\text{--}15$] and 3.3 fold [$<0.1\text{--}19$] increases, respectively.” In other words, these analyses found no significant departure from linear metabolism at low exposure concentrations, since all confidence intervals include 1.

Rappaport et al. (2013) responded vigorously. They confirmed that these alternative definitions of background levels of exposure do indeed lead to DSM ratios of metabolite concentrations in urine to benzene concentration in air that no longer decrease (and even increase) with inhaled benzene concentrations below about 0.1 ppm (*ibid*, Fig. 4.5, panels B and C), but construed this finding as unhelpful. They acknowledged that “Indeed, our analyses indicate that [these redefinitions] effectively precluded any attempt at elucidating DSM of benzene in the range of 0.03 p.p.m.” similar to the Price et al. conclusion that “The new analysis indicates that findings of increased production are. . . highly uncertain.” Thus, both sets of authors appear to agree that there was no evidence of decreasing DSM ratios at low levels of benzene exposure if all workers without occupational exposures to benzene are taken as the control group, but DSM ratios do decrease with benzene concentration if the control group is defined as the portion of the workers with no occupational exposure to benzene and with the 60 lowest levels of urinary benzene. Measured values of benzene exposures were missing for all concentrations below 0.2 ppm, the lower detection limit, making the definition of “control group” and assumptions about the levels of benzene to which they were exposed decisive for deriving or refuting the conclusion that DSM ratios decrease with benzene exposure concentrations at these low levels. A further exchange of correspondence between Price and Rappaport et al. debated the issues and the propriety of *ad hominem* comments further, but did not change the conclusions on either side (Price et al. 2013; Rappaport et al. 2013).

With this somewhat contentious background, and acknowledging funding from CONCAWE, a European organization representing petroleum companies, the following sections undertake a fresh examination of how benzene metabolites vary with benzene concentrations at low concentrations (3 ppm or less) of benzene in air using data from the Tianjin, China factory workers. In principle, this might seem a relatively simple matter to resolve, as both benzene levels below 1 ppm and corresponding metabolite concentrations have been measured in these workers, allowing direct inspection of the relationships between them. That is the approach

emphasized in the following sections. But considerable debate about benzene toxicology and interpretation of data at low doses appears to have been caused by use of alternative modeling assumptions and interpretations of data, as just described. The following sections therefore follow the constructive advice of Thomas et al. (2014), that “The use of non-parametric approaches is particularly relevant here and in epidemiological studies in general because it is impossible to know the exact functional relationships among the variables such as gene expression, dose from exposure, age, gender and smoking status of the subject, cell counts etc. Non-parametric approaches make minimal assumptions about these functional relationships and let the observed data guide the choice of the best models using rigorous statistical criteria (e.g., cross-validation). The implication of making parametric assumptions is that if these assumptions are untrue (which is almost certainly the case), the results produced can be difficult to interpret.” Accordingly, we will apply non-parametric methods to the Tianjin data, while also commenting on the extent to which the data are consistent with a linear (parametric) relationship between benzene exposure concentrations and metabolite concentrations at low doses.

Data

The data to be analyzed are from two factories in Tianjin, here called Factories 1 and 2, in which workers were occupationally exposed to benzene, and from a third factory, Factory 3, in which they were not. The data were obtained by e-mailing requests to the lead authors (Rappaport and Price) of the papers and letters just discussed; no response was received from Rappaport, but Price promptly provided a full copy of the data set, which is also available as supplemental information in Price et al. (2012). To our knowledge, the validity of these data are not disputed, and the same data have been used by both teams.

Table 4.1 shows the layout of the raw data, excluding some columns dealing with identifying individual workers and dates. Each row represents measurements for a single worker. Several workers have multiple rows, as measurements were taken for them on several different days. The variables and their units is as follows: Factory = ID of factory (1 and 2 used benzene, 3 did not); Weight = worker’s weight in kg; Height = worker’s height in cm; Gender = 0 for women, 1 for men; Subject = worker’s ID; UB = Urinary benzene (nM); UT = Urinary toluene (nM); SPMA = Urinary SPMA (μ M); PH = Urinary phenol (μ M); MA = Urinary muconic acid (μ M); CA = Urinary catechol (μ M); HQ = Urinary hydroquinone (μ M); AB = Benzene in air (ppm); AT = Toluene in air (ppm); Creat = Urinary creatinine (mM); Rep and Split indicate multiple samples on different dates from the same individual; Samdate = Date of air/urine samples; BTdate = Date of analysis of UB; PCHMdate = Date of analysis of PH, MA, CA, HQ; Sdate = Date of analysis of SPMA. The full data set can be downloaded from <http://cox-associates.com/CAT.htm>; it is data “Tianjin” in the software at that web site.

Table 4.1 Layout of the raw data for workers in three factories in Tianjin, China

Gender	Age	Smoke	Cig.	ExpCat	Factory	Weight	Height	AB	UB	MA	SPMA	PH	CA	HQ	Creat	AT	UT
1	22	1	2	1	1	61	166	1,246	382.5	16,98	0.16	106.9	17.04	23.9	1,028	3.41	68.31
1	19	0	0	1	1	52	169	1,965	446.3	18,54	0.51	199.0	21.92	24.2	1,734	10.15	107.66
1	19	0	0	1	1	52	169	6,743	1170.8	170.37	3.13	1082.9	114.06	179.1	1,734	6.77	62.36
1	20	0	15	1	1	68	170	1,570	567.6	28.83	0.38	212.2	29.00	37.1	1,851	5.73	56.41
0	23	0	0	1	1	55	158	13,821	1078.8	288.63	3.97	2016.1	192.56	223.4	1,551	6.37	59.64
1	26	1	10	1	1	59	164	14,660	642.9	61.50	0.59	338.0	49.60	53.5	1,29	8.97	48.87
1	23	1	5	1	1	73	171	2,233	784.1	17.54	0.32	208.4	35.08	26.0	0.755	12.30	130.76
1	23	1	5	1	1	73	171	2,494	383.5	8.99	0.44	102.4	23.17	9.7	0.755	10.65	71.02
1	23	1	2	1	1	62	159	2,941	127.1	68.04	0.19	550.4	36.05	82.7	1,702	8.92	40.09
1	29	0	0	1	1	50	162	2,522	101.8	32.00	0.05	241.5	24.67	28.4	1.17	10.61	36.02
1	22	1	30	1	1	82	173	1,730	79.6	7.93	0.05	66.6	32.11	16.4	0.8	9.15	47.02
0	37	0	0	1	1	63	160	1,353	919.5	16.51	0.96	178.5	28.89	22.2	1.575	45.21	175.44
0	37	0	0	1	1	63	160	1,718	318.4	23.21	1.39	219.6	31.66	24.5	1.575	23.71	150.24
1	30	1	4	1	1	55	170	1,851	96.1	87.66	0.42	492.4	63.57	56.5	1,294	24.24	30.49

Methods

Our analyses emphasize directly plotting and examining the data wherever possible using simple descriptive statistics and visualizations including histograms, interaction plots, and scatter plots. These graphs (Figs. 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.9 and 4.10), with vertical bars around data points indicating 95% confidence intervals, as well as linear correlation and multiple linear regression results and non-parametric regression curves fit to the scatter plots via locally weighted scatterplot smoothing (lowess), were all generated in *Statistica* (www.statsoft.com/Products/STATISTICA-Features), a commercial statistical software environment marketed by Quest Software Inc (formerly marketed by StatSoft). These plots suffice to establish our key findings. We also performed a more sophisticated analyses using

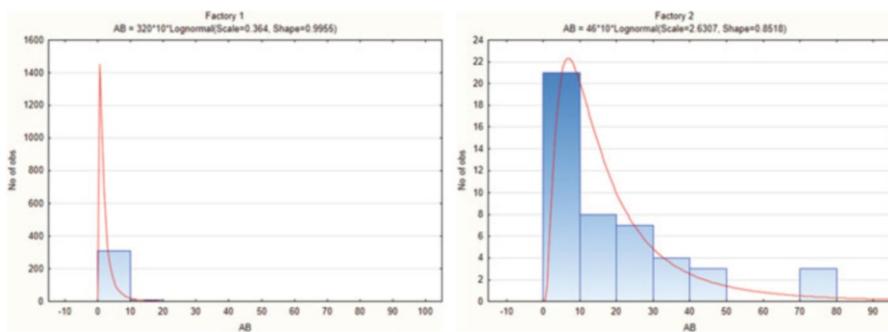


Fig. 4.1 Histograms showing the skewed distributions of air benzene (AB) in factories 1 (left) and 2 (right). Log-normal distributions (continuous curves) fit to these data appear to underestimate their skewness (heavy right tails). Note the different vertical scales. High concentrations are less frequent in Factory 1 than in Factory 2

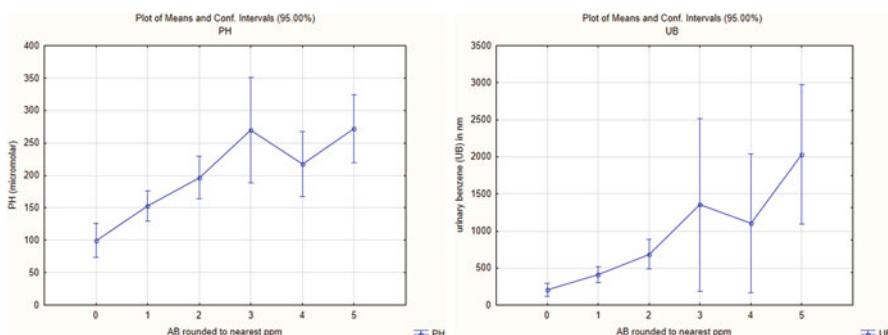


Fig. 4.2 Plot of phenol (PH) concentrations (left) and urinary benzene (right) concentrations vs. air benzene (AB) for all exposed workers in Factories 1 and 2. The plots do not clearly suggest a nonlinear or supra-linear relationship between benzene and phenol for AB between 0 and 3 ppm

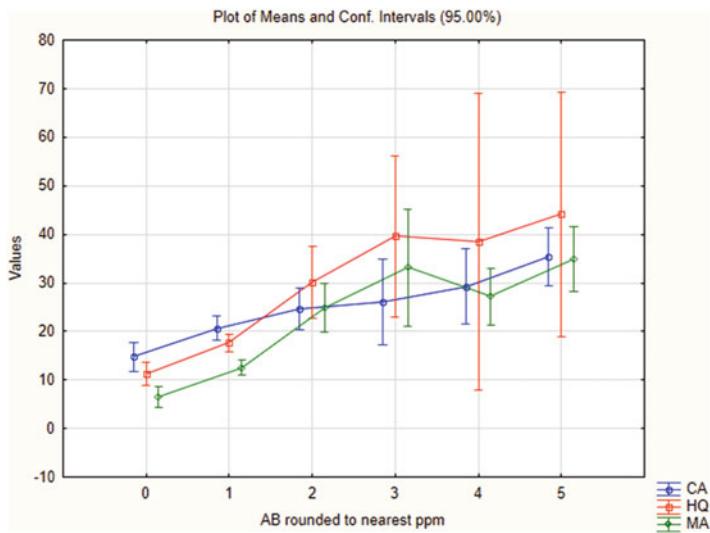


Fig. 4.3 Plot of hydroquinone (HQ), catechol (CA), and muconic acid (MA) concentrations (μM) vs. air benzene (AB) (ppm) for all exposed workers in Factories 1 and 2. The plot does not clearly suggest a nonlinear or supra-linear relationship between benzene and these metabolites for AB between 0 and 3 ppm

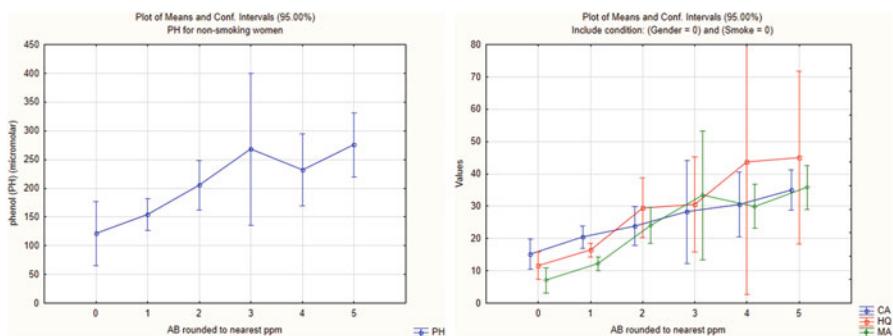


Fig. 4.4 Plots of phenol (left) and hydroquinone (HQ), catechol (CA), and muconic acid (MA) concentrations (μM) vs. air benzene (AB) (ppm) for non-smoking women in Factories 1 and 2. These plots do not show clear nonlinear or supra-linear relationship between benzene and these metabolites for AB between 0 and 3 ppm

computational statistics packages from the CRAN repository for the R project, <https://cran.r-project.org/>, as follows:

- Bayesian network learning algorithms were run using the R package *bnlearn*, (www.bnlearn.com/) with all settings at their default values. This package uses nonparametric machine learning algorithms to discover statistical dependencies and conditional independence relationships among variables, revealing which variables are informative about each other. It was used to generate Fig. 4.8.

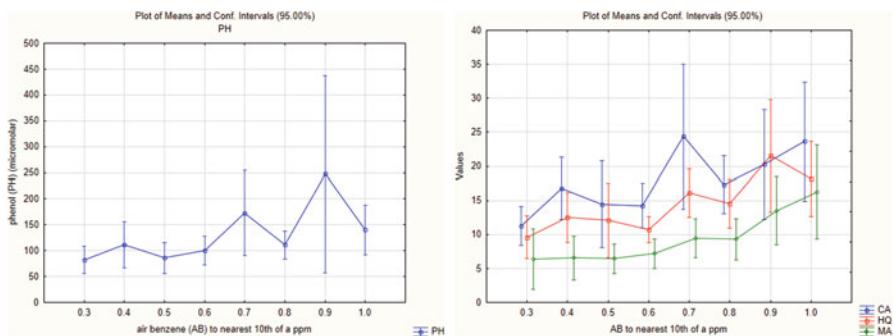


Fig. 4.5 Plots of phenol (left) and hydroquinone (HQ), catechol (CA), and muconic acid (MA) concentrations (μm) (right) vs. air benzene (AB) for workers in Factories 1 and 2 exposed to <1 ppm. These plots do not show a clear nonlinear or supra-linear relationship between benzene and these metabolites for AB between 0 and 1 ppm

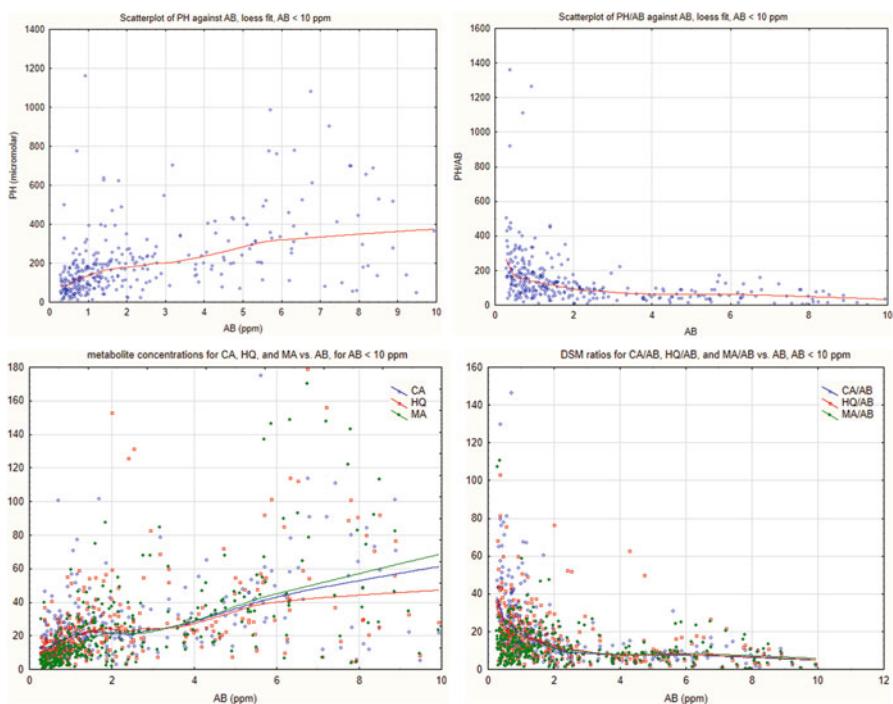


Fig. 4.6 Plots of phenol (PH) vs. air benzene (AB) (upper left) and PH/AB DSM ratio vs. AB (upper right) for workers in Factories 1 and 2 exposed to <10 ppm. The declining DSM ratios on the right are compatible with approximately linear metabolism on the left; they reflect the arithmetic fact that small denominators are associated with large ratios and large denominators are associated with small ratios

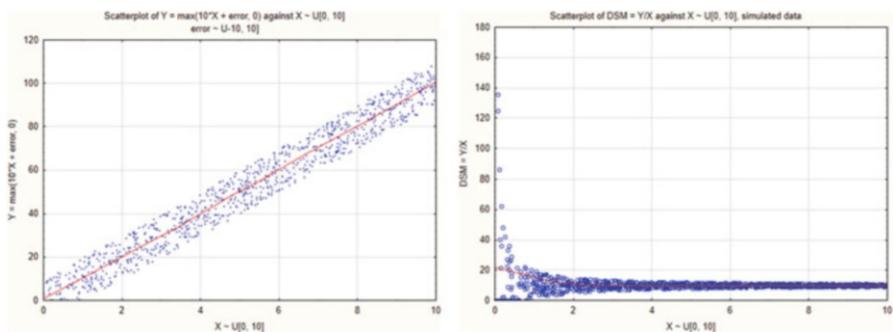


Fig. 4.7 Plots of simulated variable $Y = \max(0, 10X + \text{error})$ vs. X (left) and of the simulated DSM ratio Y/X vs. X (right), where $\text{error} \sim U[-10, 10]$. The declining simulated DSM ratios on the right are compatible with the simulated linear metabolism on the left; they reflect the arithmetic fact that small denominators are associated with large ratios and large denominators are associated with small ratios

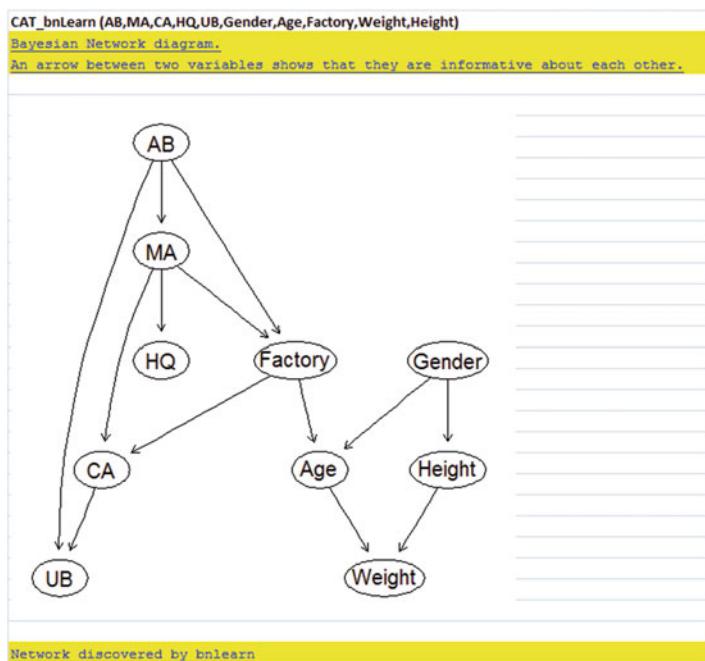


Fig. 4.8 Bayesian network (BN) learned by *bnlearn* from data for benzene-exposed workers in Factories 1 and 2. An arrow between two variables (in either direction) indicates that they are statistically dependent, i.e., informative about each other

- Random forest model ensembles were generated by the R package *randomForest*, (<https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>) to quantify multivariate statistical dependencies among variables while controlling for the levels of other variables using multiple nonparametric (classification and regression tree (CART) tree) models. Random forest provides a non-parametric alternative to parametric regression modeling that deals with model uncertainties about which variables to include as predictors and what functional forms to specify to relate predictors to a dependent variable by fitting hundreds of CART trees to random subsets of the data and averaging their predictions. It was used to generate Table 4.6 and several results mentioned in the accompanying discussion.

Bayesian networks and random forest, as well as CART trees, are also available in most current commercial statistics and machine learning software and packages, as well as in free R packages and Python sci-kit learn; the above-cited documentation provides details for the R implementations we used. We used additional R packages (*car*, *MASS*, *leaps*) for multiple linear regression analyses; the results are briefly mentioned, but not discussed in detail, as any multiple linear regression program will produce the same results. To facilitate easy replication and extensions of our analyses, we accessed all R packages and displayed the results (Fig. 4.8 and Table 4.6) using the Causal Analysis Toolkit (CAT) described in Chap. 2.

Results

Descriptive Statistics

Table 4.2 summarizes several aspects of the frequency distributions of the variables in this data set. For variables such as weight and height, the mean, median, geometric mean, and harmonic mean are all closely similar. Figure 4.1 illustrates that air benzene concentrations (AB) have a distribution with a long right tail (skewness = 5), leading to a mean value that is more than triple the median value and more than twice the geometric mean. Metabolite distributions are even more skewed (Table 4.1).

Metabolites vs. Benzene Concentrations in Air

Perhaps the most interesting research question is whether metabolite concentrations and DSM ratios vary nonlinearly with benzene concentrations between 0 and 3 ppm. To start to address this key question without introducing modeling assumptions, Figs. 4.2 and 4.3 plot mean levels of benzene metabolites against air benzene concentrations rounded to the nearest ppm. These plots pool over all individual-days of observation, so that some individuals who were measured on multiple days

Table 4.2 Descriptive statistics for all observations (pooled individual-days of data)

Variable	Valid N	Mean	Geometric mean	Harmonic mean	Median	Minimum	Maximum	Lower quartile	Upper quartile	Std. dev.	Skewness
Gender	620	0.3			0.0	0.0	1.0	0.0	1.0	0.5	0.64
Age	620	29.8	28.7	27.6	28.0	18.0	52.0	22.0	37.0	8.3	0.55
Smoke	620	0.2			0.0	0.0	1.0	0.0	0.0	0.4	1.29
Cig.	613	2.2			0.0	0.0	40.0	0.0	0.0	5.3	2.83
ExpCat	620	0.7			1.0	0.0	1.0	0.0	1.0	0.4	-1.11
Factory	620	1.6	1.4	1.3	1.0	1.0	3.0	1.0	3.0	0.9	0.88
Weight	611	60.7	59.8	59.0	60.0	39.0	114.0	53.0	66.0	10.6	0.96
Height	611	164.5	164.3	164.2	164.0	146.0	185.0	160.0	169.0	6.9	0.48
AB	366	4.9	1.9	1.1	1.5	0.3	88.9	0.7	4.7	10.4	5.00
UB	620	1219.6	67.8	0.8	114.9	0.0	54,782.2	7.8	506.2	4947.0	8.00
MA	620	27.5	6.9	2.0	6.6	0.1	651.0	1.9	23.2	66.1	5.39
SPMA	615	1.0	0.1	0.0	0.1	0.0	71.4	0.0	0.4	4.0	10.55
PH	620	248.6	122.9	75.5	116.6	7.0	6604.9	60.7	211.8	558.4	7.69
CA	620	32.1	18.7	13.1	17.6	2.1	860.9	10.3	29.6	63.9	8.17
HQ	620	31.8	14.7	8.7	12.9	0.6	742.8	7.4	26.0	67.2	5.92
Creat	620	1.3	1.2	1.0	1.3	0.1	3.3	1.0	1.7	0.6	0.31
AT	370	14.0	7.6	4.4	7.7	0.6	174.3	3.4	14.7	20.3	3.96
UT	619	71.3	19.0	3.9	30.6	0.3	1415.9	3.1	98.2	122.0	5.16

contribute multiple data points; as discussed later, individual data points are not strongly correlated between different measurement days, and including multiple days of data separately when they are available avoids the need to choose a single summary measure. Figure 4.2 shows these plot for phenol (PH) (left panel) and urinary benzene (right panel) levels averaged over all exposed workers in Factories 1 and 2 at each level of air benzene (AB) from 0 to 5 ppm, rounded to the nearest ppm. Figure 4.3 shows a similar plot for the benzene metabolites hydroquinone (HQ), catechol (CA), and muconic acid (MA), which have similar enough values to be plotted together on the same axes. These plots do not show strongly nonlinear metabolism or supra-linearity between 0 and 3 ppm of air benzene; rather, metabolism below 3 ppm appears to be approximately linear.

Table 4.3 gives the numerical values for the data points in Figs. 4.2 and 4.3, showing the mean values of each of the benzene metabolites (and also creatinine (CT), air toluene (AT), and urinary toluene (UT), and the number of person-days of data, N) for each level of air benzene concentration, 0–5, among workers in Factory 1 and Factory 2. The “PH” column in Table 4.3 corresponds to the left plot in Fig. 4.2, and the column “UB” corresponds to the right plot.

Since the modeling work of Kim et al. (2006a, b) specifically addressed non-smoking women, it is natural to wonder whether Figs. 4.2 and 4.3 may be obscuring a true nonlinearity in the metabolism of this sub-population by averaging over all exposed workers. Figure 4.4 shows analogous plots specifically for non-smoking women. In this sub-population, there is again no evidence of supra-linear metabolism at low doses.

Since almost all observations at these low concentrations come from Factory 1, focusing on the population of non-smoking women specifically in Factory 1 leaves these plots almost unchanged.

Rappaport et al. (2009) hypothesized saturation of metabolism and supra-linearity of dose-response relationships specifically below 1 ppm of benzene in air as a conjectured mechanism whereby leukemia risks in the general population might have been underestimated. Accordingly, Fig. 4.5 focuses on metabolites for air benzene exposures below 1 ppm, down to the lowest recorded levels. The horizontal axis now has increased resolution, with concentrations rounded to the nearest tenth of a ppm. Even with this sharpened focus, there is no evidence for the conjectured saturation and nonlinear (supra-linear) metabolism in this range.

The interaction plots in Figs. 4.2, 4.3, 4.4 and 4.5 do not provide reason to reject the null hypothesis of a linear relationship between benzene metabolite concentrations in urine and benzene concentrations in air of 5 ppm or less, but neither do they provide a quantitative test of the hypothesis of linearity. To help close this gap, Table 4.4 quantifies the Pearson product-moment linear correlations between air benzene and its metabolites over the whole range of observed values. (The occasional missing values were pair-wise deleted; sensitivity analyses showed that case-wise deletion or imputation made little difference.)

All correlations greater than 0.08 in Table 4.4 are statistically significantly greater than zero ($p < 0.05$). The high correlation coefficients suggest that linear relationships between air benzene and metabolite concentrations and between MA and other

Table 4.3 Mean values of each benzene metabolite and creatinine (CT), air toluene (AT), and urinary toluene (UT), and number of person-days of data, N , for each level of air benzene concentration (AB), 0–5, for workers in Factories 1 and 2

AB rounded to nearest ppm	N	UB means	MA means	SPMA means	PH means	CA means	HQ means	Creat means	AT means	UT means
0	42	205.9	6.5	0.1	99.7	14.7	11.3	1.5	18.8	78.7
1	144	409.5	12.6	0.2	152.7	20.7	17.6	1.4	14.5	108.3
2	51	687.0	24.9	0.6	196.6	24.7	30.2	1.2	12.9	84.7
3	17	1353.2	33.2	1.3	269.7	26.0	39.6	1.3	16.1	113.2
4	17	1103.3	27.2	1.0	217.3	29.2	38.5	1.3	10.9	143.4
5	18	2033.6	35.0	0.9	271.7	35.4	44.1	1.2	8.5	62.8
All grp	289	626.3	17.3	0.4	170.8	22.3	23.1	1.4	14.2	99.3

Table 4.4 Linear correlations between benzene and its biomarkers

Variable	AB	UB	MA	SPMA	PH	CA	HQ
AB	1.000	0.56	0.77	0.66	0.71	0.68	0.70
UB	0.557	1.00	0.54	0.69	0.53	0.51	0.50
MA	0.766	0.54	1.00	0.72	0.94	0.90	0.95
SPMA	0.659	0.69	0.72	1.00	0.76	0.78	0.68
PH	0.711	0.53	0.94	0.76	1.00	0.96	0.89
CA	0.684	0.51	0.90	0.78	0.96	1.00	0.86
HQ	0.700	0.50	0.95	0.68	0.89	0.86	1.00
Creat	-0.043	0.08	0.08	0.06	0.11	0.11	0.10
AT	0.049	-0.02	-0.02	0.01	0.02	0.01	-0.01
UT	-0.089	0.03	0.04	0.05	0.05	0.08	0.04

metabolites fit the data well. Spearman rank correlation coefficients between AB and PH, CA, and HQ are 0.61, 0.60, and 0.68, respectively, indicating that non-linear increasing relationships between AB and these metabolites do not provide a better description of the data by this criterion than straight-line relationships. Benzene metabolite concentrations are even more strongly correlated with each other than with air benzene (e.g., the correlation is 89% between phenol and the phenolic metabolite hydroquinone, HQ; 94% between phenol and CA; and 96% between phenol and MA), suggesting that these metabolites increase roughly linearly with phenol (PH) and with each other, as predicted by the hypothesis of low-dose linear metabolism.

Inter-individual Variability and Declining DSM Ratios

We have arrived at an apparent puzzle: although recent articles have repeatedly claimed disproportionately large dangers from benzene metabolites at low concentrations of benzene, stating that the Tianjin data “provide extremely strong statistical evidence” of a putative high-affinity enzyme active primarily below 1 ppm (Rappaport et al. 2010) and show “highly nonlinear” benzene metabolism with increasing benzene exposure above 0.03 ppm (Kim et al. 2006b), and, more specifically, disproportionately large production of benzene metabolites at these relatively low exposure concentrations (Rappaport et al. 2009), the raw data in Figs. 4.2, 4.3, 4.4 and 4.5 show no such phenomenon. We see no highly nonlinear metabolism or saturation below 3 ppm (Kim et al. 2006b) or below 1 ppm (Rappaport et al. 2009) in these data. This section attempts to resolve the paradox.

A key point is that, unlike the direct plots of metabolite concentrations vs. air benzene in Figs. 4.2, 4.3, 4.4 and 4.5, the analyses by the Berkeley team focused on the *ratios* of metabolites to air benzene at different levels of air benzene. This amounts to studying how a ratio varies as its denominator increases. Thus, for example, Rappaport et al. (2009) note that “Intriguingly, the exposure-specific

production of major metabolites (phenol, muconic acid, hydroquinone, and catechol, in micromolar per parts per million benzene) decreased continuously with estimated exposure levels over the range of 0.03–88.9 ppm, with the most pronounced decreases occurring at benzene concentrations <1 ppm” and Kim et al. (2006b) wrote that “Mean trends of dose-specific levels (micromol/L/ppm benzene) of E, E-muconic acid, phenol, hydroquinone, and catechol all decreased with increasing benzene exposure, with an overall ninefold reduction of total metabolites. Surprisingly, about 90% of the reductions in dose-specific levels occurred below about 3 ppm for each major metabolite.” Both findings are commenting on the fact that a ratio decreases as its denominator increases. But this is not surprising: it is a consequence of the algebra of random variables that, even if Y is directly proportional to X, the ratio Y/X may still be a decreasing function of X (rather than a constant), with the steepest decline occurring for the smallest values of X, if there is substantial variance in the Y values for any X value, as is the case in the Tianjin data for benzene metabolites. It is therefore a mistake to interpret a declining ratio of Y/X as X increases as a sign of an intriguing or surprising low-dose nonlinear underlying toxicological mechanism if this is just what should be expected from linear metabolism with substantial variance in measured metabolite concentrations for each X level.

Figure 4.6 shows scatter plots of the concentrations of major metabolites of benzene against air benzene concentrations of 10 ppm or less on the left side, and corresponding DSM ratios of metabolite concentrations per ppm of air benzene vs. ppm of air benzene on the right, with phenol in the top panels and CA, HQ, and MA in the bottom panels. To aid visualization, a non-parametric regression (lowess) curve is shown for each scatter plot.

In agreement with Kim et al. (2006b) and Rappaport et al. (2009, 2010), we observe that the DSM ratios on the right side of Fig. 4.6 decline as air benzene increases, with the majority of the decline taking place at air benzene concentrations below 1 ppm. However, the left side of Fig. 4.6 shows that this decline does not correspond to any strongly nonlinear metabolism at low concentrations of air benzene. Rather, the downward slopes of the right-side DSM plots are consequences of the fact that metabolite concentrations are distributed with substantial variance and skew around their air benzene exposure concentration-dependent means (or medians, geometric means, etc.), so that the ratios of these random variables are negatively correlated with their denominators. As cautioned by Liermann et al. (2004), “Ratio data, observations in which one random value is divided by another random value, present unique analytical challenges. . .[S]everal authors have pointed out that interpreting results of analyses based on ratio data can be non-intuitive, potentially leading to unintended inference and incorrect conclusions. . . They advocate reformulating the hypothesis in terms of the numerator, using the denominator as an explanatory variable.” The left-side plots in Fig. 4.6 accomplish this reformulation.

To further clarify the logic of ratios of random variables, Fig. 4.7 presents results from an artificial data set with 1000 simulated data points. Here, the true relationship between exposure concentration X and metabolite concentration Y is known exactly:

it is specified to be purely linear, $Y = 10X$, as X ranges from 0 to 10 ppm and Y ranges from 0 to 100. However, there is a relatively modest error variance in the measured values of Y , much less than in the realistic data in Fig. 4.6 but sufficient to illustrate methodological points. It is described by addition of an error term uniformly distributed between -10 and 10 , but truncated to prevent negative concentration values. Negative values of Y are assumed to be impossible, so any negative value due to random error is rounded up to 0 ; this imparts an upward bias to errors, as is probably realistic with real-world measurements that allow large positive values but no negative values. The left side of Fig. 4.7 shows this simulated data set; a nonlinear regression (lowess) curve fit to the data provides an excellent approximation to the true linear relationship, $E(Y | X) = 10X$. The right side of Fig. 4.7 shows how the ratio Y/X varies with X . Even though this simulated data set has been constructed so that metabolism is exactly linear, the DSM ratio on the right is declining, with most of the decline occurring below 1 ppm. It is not necessary to postulate a hidden high-affinity enzyme or other toxicological mechanism to explain this pattern: it is a consequence of the algebra of random variables and of the assumed error model, rather than of toxicological mechanisms. Thus, contrary to the inferences drawn by the Berkeley group, a declining DSM ratio does not necessarily provide evidence against the null hypothesis of linear metabolism at low doses; the two are entirely compatible both in principle (Fig. 4.7) and in practice (Fig. 4.6).

Tying Up Some Loose Ends: Joint Frequency Distributions of Variables and Intra-individual Variability of DSM Ratios

Price et al. (2012) and the Berkeley group (Kim et al. 2006b; Rappaport et al. 2010, 2013) have debated how to best to define the “background” levels of metabolites to subtract from exposure-dependent increases; whether it is better to use arithmetic or geometric means; and how to estimate levels of benzene exposure from urinary benzene measurements when no measurements in air are available. In light of the wide variability in measured metabolite concentrations at each air benzene level in Fig. 4.6 (with substantial variability persisting even if the data are log-transformed), the following comments consider the implications of such distributions for previously debated issues.

Table 4.5 displays selected benzene metabolite concentrations in urine from workers at Factory 3, the control factory without occupational benzene exposure for different levels of urinary benzene (UB). It is clear that there is substantial variability in metabolite concentrations for these occupationally unexposed workers; indeed, these distributions of values overlap substantially with the corresponding distributions of metabolite levels for exposed workers from Factory 1 (CA interquartile range from 10.8 to 30.2 for Factory 1 workers compared to 7.4 to 18.2 for Factory 3; HQ interquartile range from 9.7 to 27.0 for Factory 1 compared

Table 4.5 Mean values of each benzene metabolite concentration CA (μM), HQ (μM), and MA (μM) and number of person-days of data, N , for each level of urinary benzene concentration (UB in nM) for workers in the control Factory 3

UB rounded to nearest nM	CA (μM) means	HQ (μM) means	MA (μM) means	N
0	14.0	7.7	1.3	22
1	11.3	7.5	1.4	59
2	14.8	7.2	1.8	31
3	14.0	5.6	0.9	13
4	17.8	8.8	2.0	8
5	19.3	14.5	1.4	2
6	14.9	8.1	1.4	8
7	20.3	11.3	1.9	5
8	25.1	12.8	2.0	1
10	12.8	10.7	1.4	3
19	22.5	10.3	5.0	1
20	29.1	11.5	1.3	1
23	12.7	7.0	0.6	1
29	28.2	20.8	1.6	1
30	19.0	12.7	1.0	1
67	19.8	14.4	2.7	1
100	49.1	55.0	2.8	1
168	44.2	31.6	3.2	1
All grps	14.4	8.4	1.5	160

to 4.5 to 10.6 for Factory 3; MA interquartile range from 4.4 to 24.7 and MA 10th and 90th percentiles of 2.2 and 44.4 for Factory 1 compared to 0.61 to 1.8 for Factory 3. Only for MA is the interquartile range, i.e., the interval between the lower and upper quartiles of the frequency distribution of values, clearly higher for workers in Factory 1 than the values shown in Table 4.5 for workers from Factory 3.) This suggests that simply subtracting off *any* single number to represent the background level of a metabolite is an over-simplification: background levels of metabolites are better modeled as following distributions—or, more accurately, a joint frequency distribution—in the population.

More generally, the variables may be viewed as having a joint frequency distribution. Figure 4.8 shows the structure of a Bayesian network (BN) that represents this empirical joint distribution via products of marginal frequencies of values at input nodes (those with only outward-pointing arrows, such as Gender and AB) and conditional frequencies of values, given the values of their inputs, at all other nodes.

Roughly speaking, the BN structure in Fig. 4.8 identifies two variables as informative about each other, even after conditioning on other variables, if and only if they are linked by an arrow (in either direction, by Bayes' rule). Otherwise, they are conditionally independent of each other. The BN in Fig. 4.8 was discovered automatically from the Tianjin data by the *bnlearn* package in R, accessed via the CAT add-in for Excel. (Smoking was excluded because of missing data values.)

Figure 4.8 sheds light on how best to estimate air benzene levels (AB) from other measured variables. It implies that AB should be estimated from MA and UB levels, as well as from the factory at which a worker is employed. UB levels depend on CA (which depends on MA and factory) as well as AB. Attempts to infer AB from UB should therefore adjust for levels of other variables, especially MA. That the factory at which a worker is employed is identified as an independent predictor of MA and CA as well as AB suggests that different factories might have different practices (perhaps leading to more or less dermal exposure) or that other local differences might provide explanations for metabolite levels that are unrelated to the occupational inhalation exposures we have focused on.

Figure 4.9 visualizes the result that both UB and MA are informative about AB by estimating contours of equal AB concentrations for different combinations of UB and MA. The non-parametric regression method used in Fig. 4.9 to estimate contours from data on combinations of MA, UB, and AB values for exposed workers is the default spline surface-fitting algorithm in the commercial *Statistica* package, but other non-parametric regression methods (distance-weighted least squares) and polynomial regression yield similar conclusions: AB depends on MA as well as UB. A high level of MA paired with a low level of UB still predicts a high level of AB. If AB depended only on UB, then the contours in Fig. 4.8 would be vertical lines, which is not the case.

Table 4.6 gives a more quantitative assessment of the relative importance of different variables for predicting AB using a standard non-parametric machine learning algorithm (random forest, implemented via the R package *randomForest*

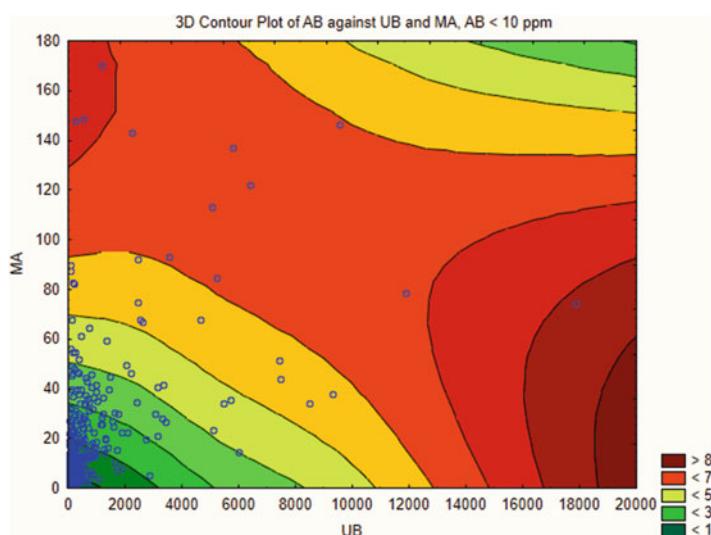


Fig. 4.9 Air benzene levels (ppm) are best predicted for workers at each factory from MA (μM) (vertical axis) as well as UB levels (nM) (horizontal axis). AB is likely to be high if MA or UB is high. Calibration models that assume that AB depends only on UB would imply vertical contours, which are not observed. Contours are estimated by regression splines in *Statistica*[®] (all settings at default values)

Table 4.6 Estimated importances of different variables for predicting air benzene (AB) using a random forest ensemble of classification and regression trees

CAT_importance (AB, MA, CA, HQ, UB, Gender, Age, Factory, Weight, Height)	
Dependent variable: AB	
<u>randomForest summary</u>	
No. of variables tried at each split: 3	
Mean of squared residuals: 39.41704	
% Var explained: 64.78	
Importance table	
From most to least important, the relative importances of these potential causes are as follows:	
Variable	Importance (%IncMSE)
MA	48.93108859
CA	41.18949572
HQ	23.62376035
UB	12.30793747
Factory	9.72260729
Gender	0.87745199
Age	0.14049447
Weight	0.07240969
Height	-1.20498238

A variable's importance is measured here as the increase in mean squared error in predicting < AB > if the variable is dropped

MA is the single most useful predictor, reducing mean squared prediction error by about 4 times as much as UB

with all options at their default values, accessed from Excel via the CAT add-in). The metric of importance in Table 4.6 is the estimated increase in mean squared prediction error caused by dropping each variable. By this criterion, MA is the single most important predictor, and calibration models in the literature that depend only on Factory and UB omit the three most important variables (MA, CA, and HQ), any or all of which would allow more accurate predictions of air benzene than UB.

A different way to quantify this is to examine how much of the total variance in air benzene is explained by each subset of predictors using the *randomForest* package via the Sensitivity Plots tab of CAT. The results are that Factory alone explains 29% of the total variance; UB and Factory together explain 40%; MA and Factory explain 56%; MA and UB explain 64%; and UB and MA and Factory explain 65%. As implied by the structure of the Bayesian network in Fig. 4.7, including the other variables as predictors, such as Gender, Age, CA and HQ, does not significantly further improve prediction of AB once UB, MA and Factory are known. Only about 65% of the variance in AB can be explained by the other variables, and including MA, UB, and Factory achieves this. For comparison, a main-effects multiple linear regression model using the same variables explains about 62% of the variance in AB ($R^2 = 63\%$, adjusted $R^2 = 62\%$), so the additional flexibility of the random forest non-parametric model ensemble to represent

nonlinearity and interactions among variables contributes relatively little improvement compared to a straightforward linear model.

Given the identified importance of MA as a predictor of AB (Table 4.6 and Fig. 4.9) and HQ and CA (Fig. 4.8), it is worth examining how AB, CA, and HQ vary with MA. Figure 4.10 uses scatter plots of these variables on the vertical axes (AB in the upper panel, HQ and CA in the lower panel) to visualize the answer. Even

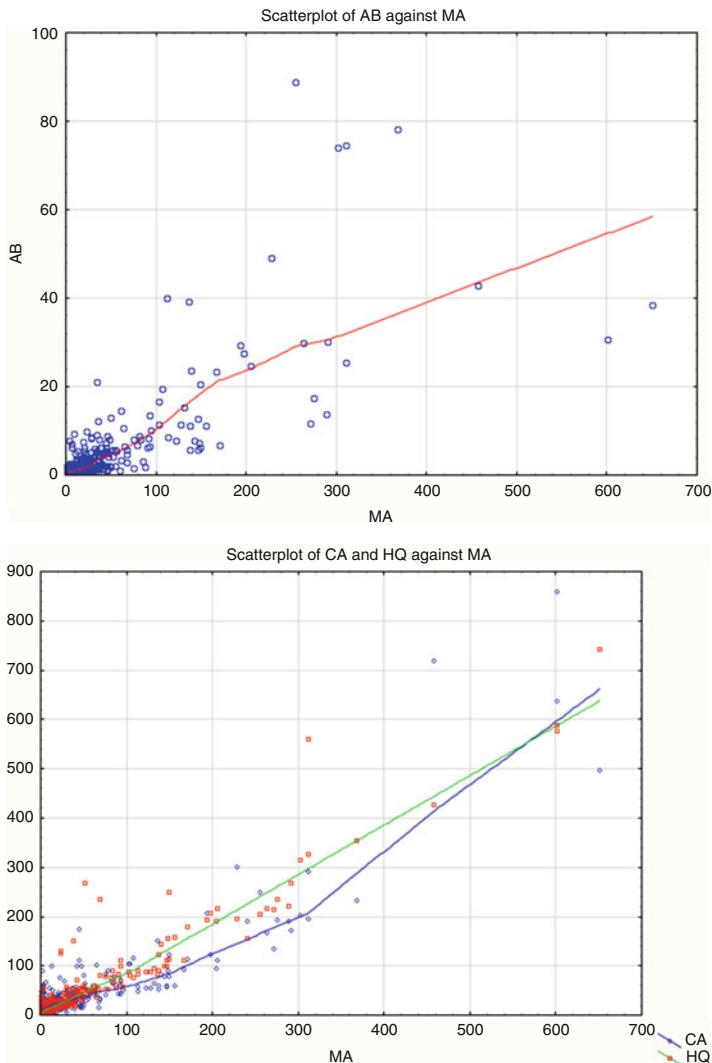


Fig. 4.10 Metabolite MA (μM) (horizontal axes) is approximately linearly related to air benzene (AB) in ppm (vertical axis, upper panel) and to metabolites HQ and CA (vertical axis, lower panel) for all AB values < 10 ppm and MA values $< 100 \mu\text{M}$. Nonparametric (lowess) regression curves are shown for each scatter plot

ignoring the less important predictors UB and Factory, it is clear that the nonparametric regression curves through these scatter plots are approximately linear for $MA < 100 \mu\text{M}$, corresponding roughly to $AB < 10 \text{ ppm}$. For HQ, this linearity extends over the entire range of observed MA values. There is no evidence of any departure from linearity for these metabolites at benzene concentrations less than 10 ppm, and no evidence of low-dose supra-linearity in production of HQ or CA.

A final question of great scientific and public healthy interest is the extent to which different individuals consistently metabolize benzene differently. If a DSM ratio such as UB/AB or a metabolite ratio such as MA/PH is exceptionally high or low for an individual on the first day it is measured, is it likely to be exceptionally high or low again the next time it is measured? Without presenting details, we can easily summarize the empirical answer: comparing the ratios on different days for individuals for whom measurements were collected more than once shows that the answer is no. The ratios on 1 day are approximately independent of ratios on successive days of measurement. The wide scatter in values of ratios appears to be due not mainly to inter-individual heterogeneity, with some individuals being consistently more or less efficient metabolizers than others, but rather to day-to-day random variation in measurement results. This emphasizes again the importance of accounting for the distributions of these random variables in understanding and quantifying the effects of benzene exposures on benzene metabolites.

Discussion and Conclusions

Figures 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10 provide a fresh look at data on low-dose metabolism of benzene (below 10 ppm of benzene in air) in workers from Tianjin China using plots and non-parametric analyses to visualize how metabolite levels vary with benzene concentrations in air. Our main conclusions, as they relate to previous discussions and debates about how to interpret the Tianjin metabolite data, are as follows.

- The data provide no reason to reject the hypothesis that benzene metabolism is linear at benzene concentrations below 10 ppm (Figs. 4.2, 4.3, 4.4, 4.5, and 4.10).
- Price et al. and the Berkeley group (e.g., Rappaport et al. 2013) appear to agree that the Berkeley group's previously published conclusions about supra-linear metabolism at low concentrations of air benzene are sensitive to assumptions about how "background" exposure levels should be defined, while disagreeing on which definition is best justified.
- The Berkeley group appears to be correct that average dose-specific metabolism (DSM), i.e., concentration of metabolite per ppm of benzene in air, is a decreasing function of its denominator, ppm of benzene in air (Fig. 4.6). The larger its denominator, the smaller the DSM ratio. Specifically, Fig. 4.6 confirms the finding of Kim et al. (2006b) that "Mean trends of dose-specific levels (micromol/L/ppm benzene) of E,E-muconic acid, phenol, hydroquinone, and

catechol all decreased with increasing benzene exposure... [and most] of the reductions in dose-specific levels occurred below about 3 ppm for each major metabolite."

- However, this arithmetic relationship between DSM ratios and their denominators does not imply "that benzene metabolism is highly nonlinear with increasing benzene exposure above 0.03 ppm, and that current human toxicokinetic models do not accurately predict benzene metabolism below 3 ppm" (Kim et al. 2006b) or "provide extremely strong statistical evidence that benzene is metabolized to PH and MA via two enzymes rather than one enzyme, and that the putative high-affinity enzyme is active primarily below 1 ppm" (Rappaport et al. 2010). Instead, this pattern is entirely consistent with linear metabolism (Figs. 4.6 and 4.7).
- To us, it seems plausible that the reason the Berkeley group's putative high-affinity enzyme (Kim et al. 2006b; Rappaport et al. 2010) still has not been found after over a decade of continued research and publications is that it does not exist: the declining DSM pattern illustrated in Fig. 4.6 is readily explained by the algebra of random variables and the presence of substantial measurement error variance (i.e., DSM ratios are negatively correlated with their denominators even if mean metabolite production is proportional to air benzene), not on any underlying toxicological phenomenon.

Our main conclusion is thus that there is actually no necessary conflict between the claim that low-dose metabolism appears to be linear (or at least that the Tianjin data provide no clear reason to think otherwise) and that DSM ratios are decreasing with air benzene: both can be true. We have also briefly examined other aspects of the Tianjin data, concluding that the distributions of levels of metabolites other than MA among workers in Factory 1 (occupationally exposed to benzene) and Factory 3 (not occupationally exposed to benzene) overlap substantially, and that unmeasured levels of benzene in air can be reconstructed much better from measured levels of MA (ideally combined with information on UB an factory) than from measured levels of UB.

The analyses we have presented emphasize the value of plotting and inspecting relationships among variables in raw data without imposing modeling assumptions and toxicological interpretations upon them. Examining the data via non-parametric methods, including interaction plots, scatter plots, non-parametric regression, Bayesian network-learning algorithms, and random forest model ensembles, has shown that metabolism of benzene at concentrations below 10 ppm is linear to an excellent approximation (e.g., Fig. 4.10 and Table 4.4). Declining DSM levels as a function of air benzene are not inconsistent with this linear metabolism (e.g., Fig. 4.6). We hope that these observations will help to resolve some of the controversies in recent literature on low-dose metabolism of benzene. We join the Berkeley group (Thomas et al. 2014) in encouraging other practitioners to apply non-parametric descriptive analytics methods, as well as data visualization, to help avoid and resolve confusions and controversies that arise from the use of models and modeling assumptions of unknown validity. The descriptive analyses in this chapter illustrate the potential value of non-parametric methods in explaining and resolving controversies in the interpretation of benzene metabolism data.

References

- Aksoy M, Erdem S, DinCol G (1974) Leukemia in shoe-workers exposed chronically to benzene. *Blood* 44(6):837–841
- Cox LA Jr (2006) Universality of J-shaped and U-shaped dose-response relations as emergent properties of stochastic transition systems. *Dose Response* 3(3):353–368. <https://doi.org/10.2203/dose-response.0003.03.006>
- Cox LA Jr (2009) Hormesis without cell killing. *Risk Anal* 29(3):393–400. <https://doi.org/10.1111/j.1539-6924.2008.01120.x>
- Kim S, Vermeulen R, Waidyanatha S, Johnson BA, Lan Q, Rothman N, Smith MT, Zhang L, Li G, Shen M, Yin S, Rappaport SM (2006a) Using urinary biomarkers to elucidate dose-related patterns of human benzene metabolism. *Carcinogenesis* 27(4):772–781
- Kim S, Vermeulen R, Waidyanatha S, Johnson BA, Lan Q, Smith MT, Zhang L, Li G, Shen M, Yin S, Rothman N, Rappaport SM (2006b) Modeling human metabolism of benzene following occupational and environmental exposures. *Cancer Epidemiol Biomark Prev* 15(11):2246–2252
- Kipen HM, Cody RP, Goldstein BD (1989) Use of longitudinal analysis of peripheral blood counts to validate historical reconstructions of benzene exposure. *Environ Health Perspect* 82:199–206
- Knutson JS, Kerger BD, Finley B, Paustenbach DJ (2013) A calibrated human PBPK model for benzene inhalation with urinary bladder and bone marrow compartments. *Risk Anal* 33(7):1237–1251. <https://doi.org/10.1111/j.1539-6924.2012.01927.x>
- Lan Q, Zhang L, Li G, Vermeulen R, Weinberg RS, Dosemeci M, Rappaport SM, Shen M, Alter BP, Wu Y, Kopp W, Waidyanatha S, Rabkin C, Guo W, Chanock S, Hayes RB, Linet M, Kim S, Yin S, Rothman N, Smith MT (2004) Hematotoxicity in workers exposed to low levels of benzene. *Science* 306(5702):1774–1776
- Liermann M, Steel A, Rosing M, Guttorp P (2004) Random denominators and the analysis of ratio data. *Environ Ecol Stat* 11(1):55–71
- McHale CM, Zhang L, Lan Q, Vermeulen R, Li G, Hubbard AE, Porter KE, Thomas R, Portier CJ, Shen M, Rappaport SM, Yin S, Smith MT, Rothman N (2011) Global gene expression profiling of a population exposed to a range of benzene levels. *Environ Health Perspect* 119(5):628–634. <https://doi.org/10.1289/ehp.1002546>.
- Price PS, Rey TD, Fontaine DD, Arnold SM (2012) A reanalysis of the evidence for increased efficiency in benzene metabolism at airborne exposure levels below 3 p.p.m. *Carcinogenesis* 33(11):2094–2099. <https://doi.org/10.1093/carcin/bgs257>
- Price PS, Rey TD, Fontaine DD, Arnold SM (2013) Letter to the editor in response to ‘Low-dose metabolism of benzene in humans: science and obfuscation’ Rappaport et al. (2013). *Carcinogenesis* 34(7):1692–1696. <https://doi.org/10.1093/carcin/bgt101>. Epub 2013 Mar 25
- Rappaport SM, Yeowell-O’Connell K, Smith MT, Dosemeci M, Hayes RB, Zhang L, Li G, Yin S, Rothman N (2002) Non-linear production of benzene oxide-albumin adducts with human exposure to benzene. *J Chromatogr B Analyt Technol Biomed Life Sci* 778(1–2):367–374
- Rappaport SM, Kim S, Lan Q, Vermeulen R, Waidyanatha S, Zhang L, Li G, Yin S, Hayes RB, Rothman N, Smith MT (2009) Evidence that humans metabolize benzene via two pathways. *Environ Health Perspect* 117(6):946–952. <https://doi.org/10.1289/ehp.0800510>
- Rappaport SM, Kim S, Lan Q, Li G, Vermeulen R, Waidyanatha S, Zhang L, Yin S, Smith MT, Rothman N (2010) Human benzene metabolism following occupational and environmental exposures. *Chem Biol Interact* 184(1–2):189–195. <https://doi.org/10.1016/j.cbi.2009.12.017>
- Rappaport SM, Kim S, Thomas R, Johnson BA, Bois FY, Kupper LL (2013) Low-dose metabolism of benzene in humans: science and obfuscation. *Carcinogenesis* 34(1):2–9. <https://doi.org/10.1093/carcin/bgs382>
- Schirrmeyer A, Flora B (2008) The coming wave of Benzene litigation. In: Presentation at national association of railroad trial counsel special litigation conference XVIII, Lake Tahoe, CA, 7–8 Feb 2008. [http://www.sdablaw.com/html/020708TheComingWaveOfBenzeneLitigation\(00051891\).pdf](http://www.sdablaw.com/html/020708TheComingWaveOfBenzeneLitigation(00051891).pdf)

- Steinmaus C, Smith AH, Jones RM, Smith MT (2008) Meta-analysis of benzene exposure and non-Hodgkin lymphoma: biases could mask an important association. *Occup Environ Med* 65 (6):371–378. <https://doi.org/10.1136/oem.2007.036913>
- Steinmaus C, Smith AH, Smith MT (2011) Regarding “meta-analysis and causal inference: a case study of benzene and non-Hodgkin lymphoma”: an incomplete analysis. *Ann Epidemiol* 21 (1):67–69
- Thomas R, Hubbard AE, McHale CM, Zhang L, Rappaport SM, Lan Q, Rothman N, Vermeulen R, Guyton KZ, Jinot J, Sonawane BR, Smith MT (2014) Characterization of changes in gene expression and biochemical pathways at low levels of benzene exposure. *PLoS One* 9(5): e91828. <https://doi.org/10.1371/journal.pone.0091828>
- Vlaanderen J, Lan Q, Kromhout H, Rothman N, Vermeulen R (2012) Occupational benzene exposure and the risk of chronic myeloid leukemia: a meta-analysis of cohort studies incorporating study quality dimensions. *Am J Ind Med* 55(9):779–785. <https://doi.org/10.1002/ajim.22087>
- Walter RB, Appelbaum FR, Estey EH, Bernstein ID (2012) Acute myeloid leukemia stem cells and CD33-targeted immunotherapy. *Blood* 119(26):6198–6208. <https://doi.org/10.1182/blood-2011-11-325050>
- Weed DL (2010) Meta-analysis and causal inference: a case study of benzene and non-Hodgkin lymphoma. *Ann Epidemiol* 20(5):347–355. <https://doi.org/10.1016/j.annepidem.2010.02.001>
- Wong O (1995) Risk of acute myeloid leukaemia and multiple myeloma in workers exposed to benzene. *Occup Environ Med* 52(6):380–384

Chapter 5

How Large Are Human Health Risks Caused by Antibiotics Used in Food Animals?



Consistent with findings in risk psychology about what triggers strong emotional responses and concern (see Chap. 12), risk of food poisoning from consumption of food contaminated with disease-causing bacteria and antibiotic-resistant “superbugs” sparks strong political passions, dramatic media headlines, and heated science-policy debates (Chang et al. 2014). A widespread concern is that use of animal antibiotics on farms creates selection pressures that favor the spread of antibiotic-resistant bacteria such as methicillin-resistant *Staphylococcus aureus* (MRSA) (discussed in Chap. 6), multi-drug resistant (MDR) *Salmonella*, or *E. coli* (CBS 2010). The most common effects of food-borne illness are diarrhea and possibly fever, vomiting and other symptoms of food poisoning. However, more serious harm, or death, may occur in vulnerable patients. This is especially likely if food-borne bacterial infections are resistant to usually recommended antibiotic therapies, as might happen if the infections are caused by bacteria from farms where antibiotics are used for purposes of growth promotion or disease prevention. Patients with immune systems compromised by chemotherapy, AIDS, organ transplants, or other sources can have risks hundreds or thousands of times greater than those of consumers with healthy immune systems. Fear that use of animal antibiotics on farms contributes to a rising tide of antibiotic-resistant bacterial infections has spurred many scientists, physicians, activists, journalists, and members of the public to call for elimination of the use of antibiotics as animal growth promoters.

These concerns and calls for action have usually left unquantified *how many* excess deaths, treatment failures, or days of illness each year are caused in the United States by antibiotic-resistant bacteria arising specifically from animal antibiotic use, as opposed to other sources. Most human cases of MRSA and other resistant infections are health care associated. They arise, for example, from inadequate hand washing and infection control in hospitals (Kallen et al. 2010). Although media reports sometimes conflate stories on food-borne resistance with statistics reflecting hospital-acquired cases (e.g., CBS 2010), these are in fact quite distinct etiologies. They can now often be discriminated by identifying specific molecular

markers for animal-associated as compared to hospital-associated strains of bacteria, allowing source-tracking based on molecular profiles of the bacteria found in infected patients. How many infections and fatalities per year arise among hospital patients, butchers, slaughterhouse workers, farmers, or the general public from livestock operations, meat handling, and consumption remains a topic of continuing interest, and such source-tracking is providing increasingly powerful molecular biological tools for obtaining answers. Responsible risk management is supported best by understanding how large the human health risks are now, and how much they would be changed by proposed interventions. The size of the risk depends on care taken to reduce microbial loads by participants throughout the food chain, including use of microbial safety controls during farming, transportation, slaughter, production and packaging, storage, retail, and food preparation and cooking.

This first part of this chapter introduces methods of quantitative risk assessment (QRA) for quantifying the number of adverse human health impacts per year caused by animal antibiotic use. Next, we summarize quantitative estimates and bounds on human health harm obtained by applying these methods to available data for several types of resistant bacteria (“drugs and bugs”) of greatest concern for public health in the United States. Finally, we discuss implications of such quantitative estimates for prudent risk management. Throughout the chapter, human health risks are expressed as expected numbers of illnesses, fatalities, illness-days, or quality-adjusted life years (QALYs) lost to illness per year (for population risks) or per capita-year (for individual risks). QRA can help to inform and improve risk management decisions and policies by predicting how changes in the food production process, such as greater or lesser use of antibiotics on the farm, will affect human health risks, including individual and population risks for such subpopulations as well as for the whole population of concern.

Methods of Quantitative Risk Assessment

This section reviews methods of quantitative microbial risk assessment (QMRA) and antimicrobial risk assessment. It expands upon and updates the brief summary in Cox (2008).

Farm-to-Fork Risk Simulation Models

When enough data and understanding are available, the effects of alternative risk management actions on risks created by bacteria in food—both antibiotic-resistant and antibiotic-susceptible strains—can be quantified by simulating microbial loads of bacteria along the chain of steps leading from production to consumption for each intervention. If the conditional frequency distribution of microbial loads leaving each step (e.g., slaughter, transportation, processing, storage, etc.) can be estimated,

given the microbial load entering that step, and given the controls applied (e.g., use of antibiotic sprays, refrigeration, etc.), then the effects on microbial loads of alternative risk management policies can be quantified and compared. Microbial loads are typically expressed in units such as colony-forming-units (CFUs) of bacteria per unit (e.g., per pound, per carcass, etc.) of food. If, in addition, dose-response relations are available to convert microbial loads in ingested foods to corresponding probabilities of illnesses, together with measures of illness severity (e.g., illness-days, QALYs lost, fatalities, etc.), then the effects of alternative risk management policies on human health can also be estimated and compared. As an example, Fig. 5.1 shows how the frequency distribution of illnesses per year caused by *Vibrio parahaemolyticus* in oysters are predicted to change if refrigeration time requirements that accomplish different degrees of reduction in microbial loads are implemented. The underlying quantitative microbial risk assessment (QMRA) model simulates the changes in microbial loads at successive stages from harvesting to consumption; its main logical structure and data inputs are shown in Fig. 5.2.

The discipline of applied microbiology supplies empirical growth curves and kill curves for log increase or log reduction, respectively, in microbial load from input to output of a step. These curves describe the output: input ratio (e.g., a most likely value and upper and lower statistical confidence limits) for the microbial load passing through a stage as a function of variables such as temperature, pH, and time.

A “farm-to-fork” simulation model can be constructed by concatenating many consecutive steps representing stages in the food production process. Each step receives a microbial load from its predecessor. It produces as output a microbial load value sampled from the conditional frequency distribution of the output

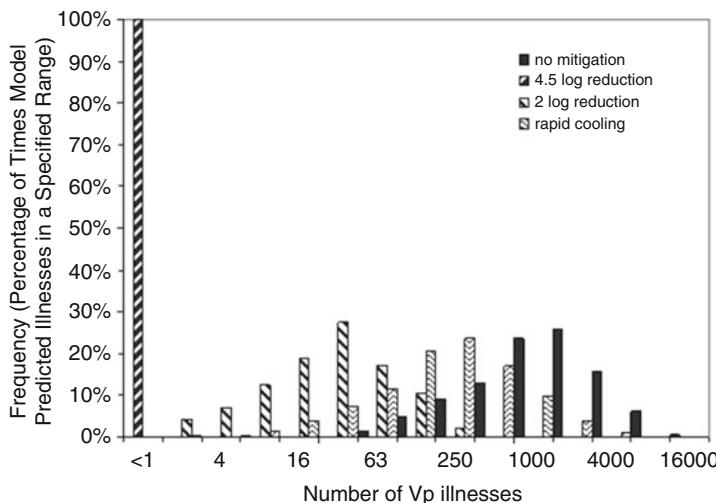


Fig. 5.1 Frequency distributions of number of *Vibrio parahaemolyticus* (Vp) illnesses per year from oysters with and without mitigation from cooling requirements. Source: www.fda.gov/Food/FoodScienceResearch/RiskSafetyAssessment/ucm184074.htm

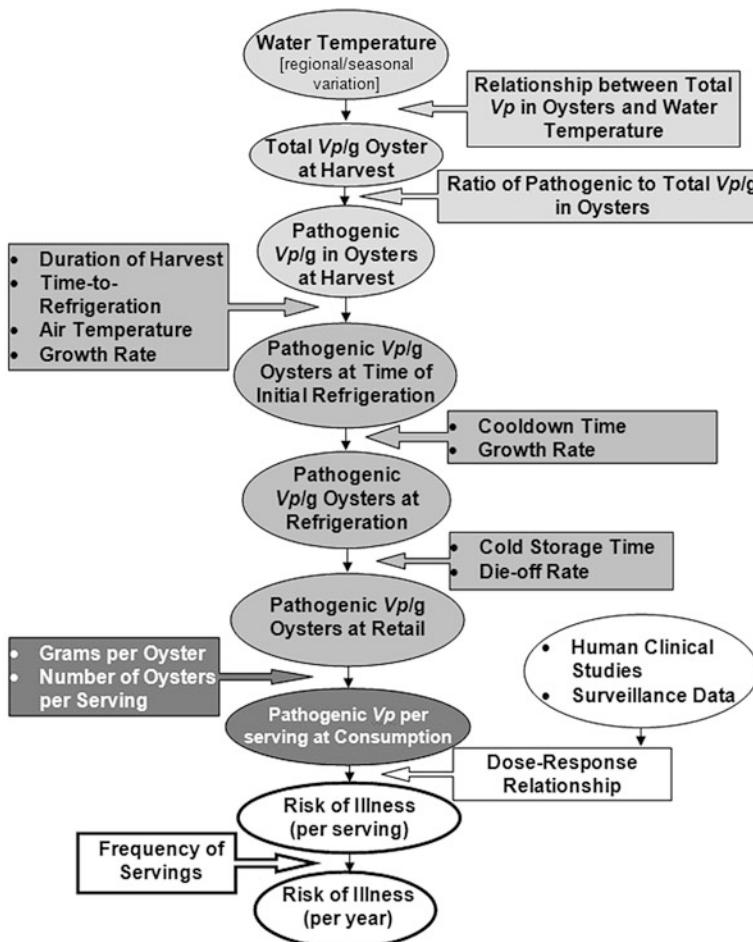


Fig. 5.2 Structure of the quantitative microbial risk assessment (QMRA) model that allows quantitative risk estimates such as those in Figure to be made. *Source:* www.fda.gov/Food/FoodScienceResearch/RiskSafetyAssessment/ucm185190.htm

microbial load, given the input microbial load, as specified by the microbial growth model describing that stage. Measured frequency distributions of microbial loads on animals (or other units of food) leaving the farm provide the initial input to the whole model. The key output from the model is a frequency distribution of the microbial load, x , of pathogenic bacteria in servings of food ingested by consumers.

Risk-reducing factors such as antimicrobial sprays and chilling during processing, freezing or refrigeration during storage, and cooking before serving are often modeled by corresponding reduction factors for microbial loads. (These may be represented as random variables, e.g., with log-normal distributions and geometric means and variances estimated from data.) The complete model is then

represented by a product of factors that increase or decrease microbial loads, applied to the empirical frequency distribution of initial microbial loads on which the factors act. Running the complete farm-to-fork model multiple times produces a final distribution of microbial loads on servings eaten by consumers. Some farm-to-fork exposure models also consider effects of cross-contamination in the kitchen, if pathogenic bacteria are expected to be spread to other foods by poor kitchen hygiene practices (e.g., failure to wash a cutting board after use.) As an example of the output from such a model, Fig. 5.3 shows an example of the distribution of microbial loads of *Salmonella* in servings of chicken well.

In summary, farm-to-fork simulation models can estimate the frequency distributions of microbial loads ingested by consumers in servings of food. As already illustrated in Fig. 5.1, QMRA can also estimate of how these frequency distributions would change if different interventions (represented by changes in one or more of the step-specific factors increasing or decreasing microbial load) were implemented. For example, enforcing a limit on the maximum time that ready-to-eat meats may be stored at delis or points of retail sale before being disposed of limits the opportunity for bacterial growth prior to consumption. Changing processing steps (such as scalding, chilling, antimicrobial sprays, etc.) can also reduce microbial loads. Such interventions shift the cumulative frequency distribution of microbial loads in food leftward, other things being held equal. If some fraction of the microbial load at each stage can be identified as resistant to antibiotics used to treat food-borne illnesses caused by consuming contaminated meat or other (possibly cross-contaminated) food, then the QMRA models can also be used to predict exposures to resistant bacteria in food.

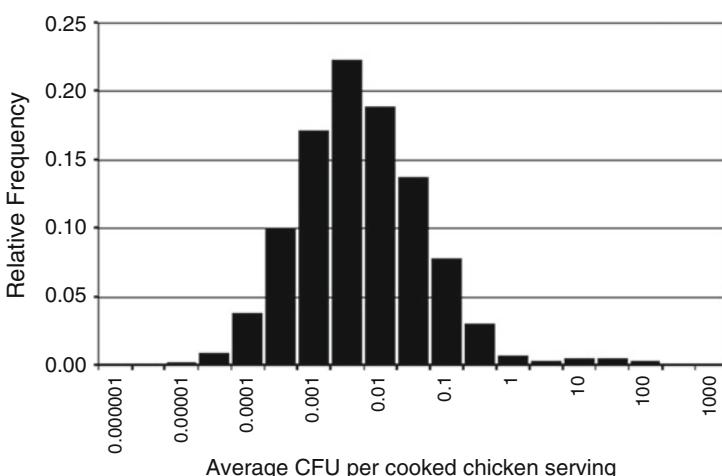


Fig. 5.3 Average dose (CFU *Salmonella*) per serving in meals prepared from contaminated broilers. Source: www.fao.org/docrep/005/y4392e/y4392e0r.htm

Dose-Response Models for Food-Borne Pathogens

Once a serving of food (e.g., chicken, oysters, hamburger, deli meats, etc.) reaches consumer, the probability that an ingested dose will cause infection and illness is described by dose-response models. Several parametric statistical models have been developed to describe the relation between quantity of bacteria ingested in food and resulting probability of illness. One of the simplest is the following exponential dose-response relation:

$$r(x) = \text{Pr}(\text{illness} \mid \text{ingest microbial load} = x \text{ CFUs}) = 1 - e^{-\lambda x}.$$

This model gives the probability that an ingested dose of x colony forming units (CFUs) of a pathogenic bacterium will cause illness. $r(x)$ denotes this probability. The function $r(x)$ is a *dose-response curve*. λ is a parameter reflecting the potency of the exposure in causing illness. Sensitive subpopulations have higher values of λ than the general population.

More complex dose-response models (especially, the widely used Beta-Poisson model) have two or more parameters, e.g., representing the population distribution of individual susceptibility parameter values and the conditional probability of illness given a susceptibility parameter. The standard statistical tasks of estimating model parameters, quantifying confidence intervals or joint confidence regions, and validating fitted models can be accomplished using standard statistical methods such as maximum likelihood estimation (MLE) and resampling methods. The excellent monograph by Haas et al. (1999) provides details and examples. It notes that “It has been possible to evaluate and compile a comprehensive database on microbial dose-response models.” Chapter 9 of this monograph provides a compendium of dose-response data and dose-response curves, along with critical evaluations and results of validation studies, for the following: *Campylobacter jejuni* (based on human feeding study data), *Cryptosporidium parvum*, pathogenic *E. coli*, *E. coli* O157:H7 (using *Shigella* species as a surrogate), *Giardia lamblia*, non-typhoid *Salmonella* (based on human feeding study data), *Salmonella typhosa*, *Shigella dysenteriae*, *S. flexneri*, *Vibrio cholerae*, Adenovirus 4, Coxsackie viruses, Echovirus 12, Hepatitis A virus, Poliovirus I (minor), and rotavirus. Thus, for many food-borne and water-borne pathogens of interest, dose-response models and assessments of fit are readily available.

Despite this base of relatively well-developed and validated dose-response models, however, two important challenges remain in developing dose-response models for specific strains of pathogenic bacteria, including antibiotic-resistant strains. Figure 5.4 illustrates the first, and Fig. 5.5 the second. In Fig. 5.4, the best-fitting model in a specific class of parametric models (the “naïve” Beta-Poisson model, which provides a widely-used approximate mathematical model for response probabilities for different doses (CFUs) ingested) provides a clearly biased description of the observed feeding trial data, under-estimating all observed response probabilities for Log Dose < 5. Figure 5.5 illustrates the problem of low-dose

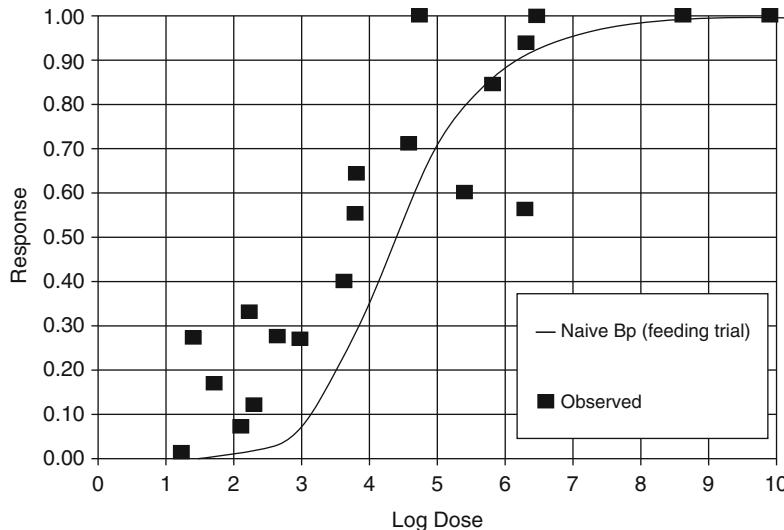


Fig. 5.4 Even the best-fitting dose-response model in a specified parametric family, such as the Beta-Poisson (BP) family, may provide a biased description of data. Here, a best-fitting dose-response model for *Salmonella* data systematically under-estimates risk at low doses. Source: www.fao.org/DOCREP/005/Y4392E/y4392e10.gif

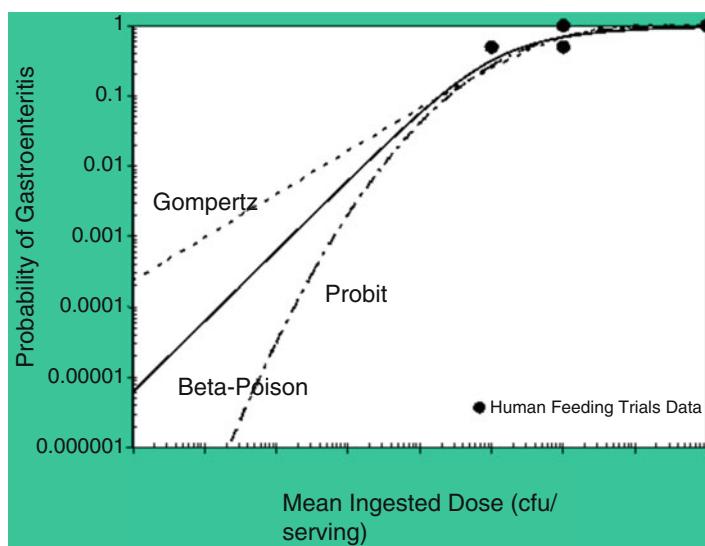


Fig. 5.5 Multiple dose-response models that fit the available experimental data equally well may make very different predictions for risks outside the range of observed data. Source: www.fda.gov/Food/FoodScienceResearch/RiskSafetyAssessment/ucm185177.htm

extrapolation, in which the dose-response relation at doses far below the range of observed data depends greatly on which specific model is assumed.

Because of these challenges, dose-response models may be highly uncertain for specific strains of pathogens, and hence risk projections based on them may also be very uncertain. Characterizing this uncertainty is a key step in QMRA that uses dose-response models.

Quantitative Risk Characterization for QMRA and Risk Management

Sampling values of exposures x from the frequency distribution predicted by a farm-to-fork model (expressed in units of bacteria-per-serving) and then applying the dose-response relation $r(x)$ to each sampled value of x produces a frequency distribution of the risk-per-serving in an exposed population. This information can be displayed in various ways to inform risk management decision-making.

For example, Fig. 5.6 shows how the (base 10 logarithm of) risk-per-serving of chicken for salmonellosis is reduced by a mitigation strategy that encourages consumers to cook chicken properly before eating it, based on the exposure sub-model in Fig. 5.3, the Beta-Poisson dose-response model in Fig. 5.5, and assumptions about how mitigation measures will affect the distribution of cooking practices.

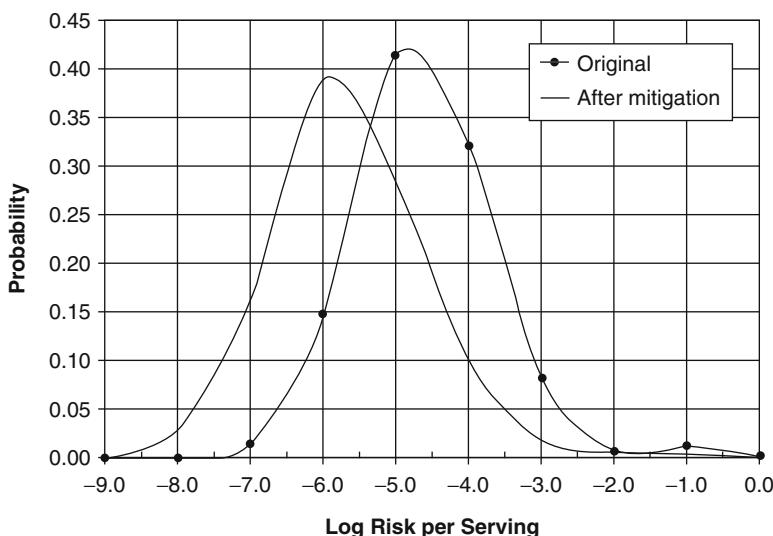


Fig. 5.6 A display showing how salmonellosis risk-per serving of chicken is reduced by better cooking practices. Source: www.fao.org/docrep/005/y4392e/y4392e0r.htm#bm27

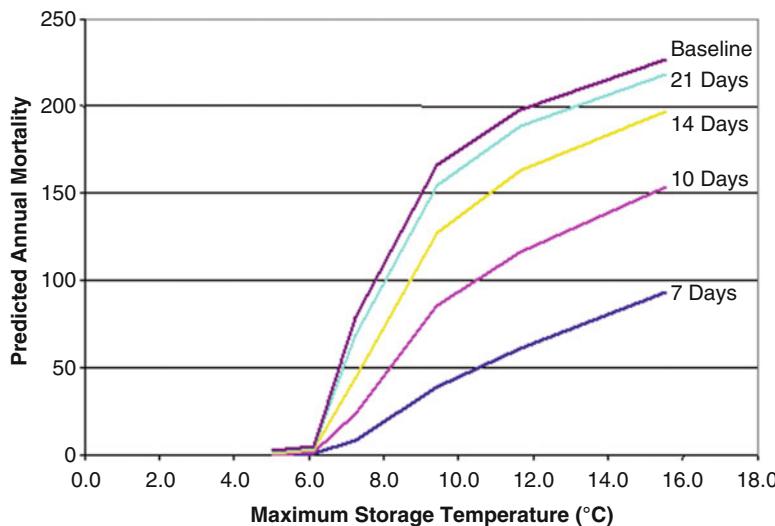


Fig. 5.7 Expected elderly mortalities per year from *Listeria monocytogenes* if different maximum storage times and temperatures are allowed. (Source: www.fda.gov/Food/FoodScienceResearch/RiskSafetyAssessment/ucm197644.htm)

Other displays showing the expected number of illnesses per year in a population, expected illnesses per capita-year in the overall population and for members of sensitive subpopulations, and frequency distributions or upper and lower confidence limits around these expected values are typical outputs of risk characterization. If particular decisions are being considered, such as a new standard for the maximum times and/or temperatures at which ready-to-eat meats can be stored before being disposed of, then plotting expected illnesses per year against the decision variables (i.e., maximum times or temperatures, in this example) provide the quantitative links between alternative decisions and their probable health consequences needed to guide effective risk management decision-making. Figure 5.7 illustrates the key concept of informing risk management decisions by showing how a measure of risk (here, expected deaths per year) varies with decisions about maximum allowed storage times and temperatures. Such displays, linking actions to their probable consequences, provide the essential information needed to inform risk management decisions.

Attribution-Based Risk Assessment and Controversies

In recent years, many efforts have been made to simplify the standard approach to quantitative microbial risk assessment just summarized, especially for application to antimicrobial-resistant bacteria. Because farm-to-fork exposure modeling and valid dose-response modeling can require data that are expensive and time consuming to

collect, or that are simply not available when risk management decisions must be made, simpler approaches with less burdensome data requirements are desirable. It is tempting to use simple multiplicative models, such as the following:

$$\text{Risk} = \text{Exposure} \times \text{Dose-Response Factor} \times \text{Consequence per case}$$

where: *Risk* = expected number of excess illness-days per year, *Exposure* is measured in potentially infectious meals ingested per year in a population, *Dose-Response Factor* = expected number of illnesses caused per potentially infectious meal ingested, and *Consequence per case* is measured in illness-days (or fatalities) caused per illness.

While such models have attractive simplicity, they embody strong assumptions that are not necessarily valid, and thus can produce highly misleading results. Specifically, the assessment of *Dose-Response Factor* requires attributing some part of the causation of illness-days to *Exposure*. Similarly, estimating the change in *Dose-Response Factor* due to an intervention that changes microbial load may require guess-work. There is often no valid, objective way to make such attributions based on available data. The risk assessment model—and, specifically, the attribution of risk to particular food sources—may then become a matter of political and legal controversy.

For example, suppose that the *Dose-Response Factor* is estimated by dividing the observed value of *Risk* in a population in one or more years by the contemporaneous values of (*Exposure* × *Consequence*). Then, this value will *always* be non-negative (since its numerator and denominator are both non-negative). The model this implies a non-negative linear relation between *Exposure* and *Risk*, even if there is no causal relation at all (or is a negative one) between them. (By analogy, one could divide the number of car accidents in Florida in a year by the number of oranges consumed in Florida that year, but the resulting “car accidents per orange consumed” ratio, although certainly positive, would not in any way imply a causal relation, or that reducing consumption of oranges would reduce car accidents per capita-year. Replacing car accidents with food-borne illnesses such as ciprofloxacin-resistant campylobacteriosis and oranges with chicken servings improves the intuitive plausibility but not the logic or credibility of such calculations.) In addition, it is a frequent observation that some level of exposure to bacteria in food protects against risk of food-borne illnesses, for example, by stimulating acquired immunity. Thus, the use of simple multiplicative model implying a necessarily non-negative linear relation between *Exposure* and *Risk* may be incorrect, producing meaningless results (or, more optimistically, extreme upper bounds on estimated risks) if the true relation is negative or non-linear.

Unfortunately, past estimates of risk of antibiotic-resistant illnesses caused by consumption of foods contaminated with resistant bacteria attributed to farm use of antibiotics have often simply assumed that some fraction of total resistant infections is caused by farm use of antibiotics (e.g., Barza and Travers 2002) or that the ratio of estimated excess resistant cases per year to servings of food per year could be interpreted causally (on a logical par with the car-accidents-per-orange-consumed

example above). For example, Chang et al. (2014) describe a case in which the United States Food and Drug Administration (FDA) used an assumption that excess cases of fluoroquinolone (FQ)-resistant campylobacteriosis were proportional to consumption of chicken exposed to FQ on the farm to estimate that between 4960 and 14,370 patients per year could have compromised treatment with ciprofloxacin (Bartholomew et al. 2003, cited in Chang et al. 2014). This model was used to support a risk management decision to withdraw fluoroquinolone use in poultry in the United States. But the subsequent decade of experience showed that the withdrawal had no detectable causal impact in reducing levels of ciprofloxacin resistance in the United States (Chang et al. 2014). As in the car crashes per orange analogy, FDA had interpreted a positive ratio as causal, and discovered only after the fact that reducing the denominator had no real-world effect on reducing the numerator.

Thus, great caution should be taken when using such simplified risk assessment models. In general, they may be useful in making rapid calculations of plausible *upper bounds* in certain situations (for example, if the true but unknown dose-response relation between exposure and risk is convex, or upward-curving), but should not be expected to produce accurate risk estimates unless they have been carefully validated (Cox 2006).

Empirical Upper-Bounding

An alternative method is available for quantifying upper bounds on the adverse human health consequences per year that could be prevented by reducing or eliminating antibiotic uses in agriculture. This method, which has advantages of simplicity, logical soundness, and reliance only on readily available data, does not attempt to simulate in detail the microbial loads traversing different pathways (e.g., food water, environment, co-selection in other bacteria, etc.) or to quantify the relevant dose-response relations. Instead, it begins with the total number of adverse events per year that might be caused by antibiotic use (e.g., total treatment failures caused by antibiotic resistance), and then uses molecular biological data to estimate upper bounds on the fraction of all such cases that might be caused by (and preventable by eliminating) animal antibiotic uses. We call this the empirical upper-bounding approach. We will illustrate it in some detail for ampicillin-resistant *E. faecium* (AREF) bacteria, and then summarize results from applying the approach to other drug-bug pairs.

Case Study: Ampicillin-Resistant *E. faecium* (AREF) Bacteria

This section, adapted from Cox et al. (2009), illustrates the empirical upper-bounding approach for potential risks to human health from use of penicillin drugs in agriculture. It illustrates how to do quantitative risk assessment when neither all

pathways from farm to consumer (or patient) nor relevant dose-response relations are known with enough confidence to permit useful simulation of microbial loads and illnesses.

Penicillin-based drugs are approved for use in food animals in the United States to treat, control, and prevent diseases and, to a lesser extent, to improve growth rates (FDA-CVM 2007; Sechen 2006; AHI 2006). Concerns that penicillin use might increase the risk of antibiotic resistance in human enterococcal infections from non-human sources, thus leading to increased morbidity and mortality (WHO 2005), have made approved feed usages of penicillins in food animals a controversial topic for several decades in the United States (IOM 1989; FDA 2000, 2003). The following sections develop a plausible upper bound on the potential for continued use of penicillin drugs in food animals to harm human health by increasing the number of antibiotic-resistant enterococcal infections in human patients. After summarizing relevant background for the hazard of greatest concern—*infection of intensive care unit (ICU) patients with ampicillin-resistant *E. faecium* (AREF) bacteria*—the following sections focus on quantifying the fraction of such resistant infections that might be prevented by discontinuing the use of penicillin drugs in food animals.

Risk to human health arises because some strains of enterococci may become opportunistic pathogens, potentially resistant to multiple drugs, that infect patients who are already seriously ill (typically in ICUs) with immune systems weakened by organ transplants, chemotherapy, AIDS, or other causes. Indeed, enterococcal infection is the second most common hospital-acquired infection in the United States (Varman et al. 2006). These infections can prolong illness and increase patient mortality. Vancomycin-resistant enterococci (VRE) are of particular concern because of their virulence and resistance to even some recently developed antibiotics. Vancomycin-resistant *E. faecium* (VREF) can cause serious and often fatal disease in vulnerable populations, such as liver transplant patients and patients with hematologic malignancies (Rice 2001).

Although many enterococcal infections, including VRE, resolve without antimicrobial treatment (Varman et al. 2006; Rice 2001), in severe cases for which antimicrobial treatment is provided, penicillin and ampicillin are often the leading choices. Most *E. faecium* infections in ICU patients in the U.S. are now resistant to vancomycin (Edmond et al. 1999; Jones et al. 2004). Patients with vancomycin resistant *E. faecium* (VREF) have worse outcomes than those with vancomycin susceptible strains—longer hospital stays and higher mortality (Webb et al. 2001). As noted by Rice (2001), *virtually all VREF are also ampicillin resistant*: “More than 95% of VRE recovered in the United States are *E. faecium*; virtually all are resistant to high levels of ampicillin.” Hence, our risk assessment treats VREF as being (at least approximately) a subset of ampicillin-resistant *E. faecium* (AREF). Since most VREF are AREF (although many AREF are not VREF), and assuming that changes in animal *penicillin* use would not significantly affect *vancomycin* resistance (consistent with historical data), we focus on human (ICU patient) infections with *vancomycin-susceptible* strains of *ampicillin-resistant E. faecium*. Presumably, this is the subpopulation that might experience decreased ampicillin

resistance if discontinuing animal penicillin drugs were to replace some AREF cases with ampicillin-susceptible cases. For patients with VREF, we assume that AREF would persist (due to the observed co-occurrence of AREF in VREF strains), so that no benefit from reduced AREF would be achieved for these patients.

Recognizing that a farm-to-fork model is not practical for AREF, due to data and knowledge gaps in release, exposure, and dose-response relations, we instead start with more readily available human data on ICU case loads and resistance rates, similar to the approach in Cox and Popken 2004. We then work backward to estimate a plausible upper bound on the annual number of human patient mortalities that might be prevented by discontinuing penicillin use in food animals.

For purposes of conservative (i.e., upper-bound) risk assessment, we define a *potentially preventable mortality* to occur whenever the following conditions hold: (1) An ICU patient dies, following (2) an *E. faecium* infection that (3) is resistant to ampicillin (AREF) (and hence might have benefited had ampicillin resistance been prevented). The infection was: (4) Vancomycin-susceptible (and hence might have also been ampicillin-susceptible, had it not been for penicillin use in food animals); (5) not known to have been contracted from the hospital environment (and hence might have been prevented by actions external to the hospital, such as elimination of AREFs from food animals); (6) could have come from food animals (i.e., has a genotype or resistance determinants of the types found in food animals). (7) The patient tolerated penicillin (i.e. was not allergic, and hence might have benefited from ampicillin, had it not been for resistance). We propose that the conjunction of these seven conditions should be interpreted as *necessary* for a mortality to have been caused (with non-negligible probability) by resistance due to use of penicillin in food animals, even though it is not *sufficient* (e.g., the infecting strain might have had some other origin than food animals, or the patient might have died anyway, even if the infection had been ampicillin-susceptible). Accordingly, the following sections estimate a plausible upper bound on annual preventable mortalities from AREF infections based on the following product of factors:

Preventable AREF mortalities per year \leq (*Total number of ICU infections per year*) \times (*fraction caused by E. faecium*) \times (*fraction of ICU E. faecium infections that are AREF and exogenous, i.e., not known to be of nosocomial origin*) \times (*fraction of these exogenous AREF cases that are vancomycin-susceptible*) \times (*fraction of vancomycin-susceptible exogenous AREF cases that might have come from food animals*) \times (*fraction of these cases that are penicillin-tolerant*) \times (*excess mortality rate for AREF cases compared to ASEF cases*).

That is, we first quantify the expected annual number of AREF cases in the U.S. that might benefit from ampicillin treatment if food animal uses of penicillin were halted (i.e., cases that are penicillin-tolerant and vancomycin-susceptible and that might have been caused by resistance determinants from food animals). Then, we multiply this number by the excess mortality rate for resistant as opposed to susceptible cases. Each of the foregoing factors can be estimated from available data, as discussed in detail in Cox et al. (2009) and summarized in Table 5.1.

Table 5.1 Summary of AREF risk calculation using empirical upper bounding

Factor	More conservative value	Less conservative value	Source
N = ICU infections/year	N = 315,000	N = 104,372.5	FDA-CVM (2004)
P _{ent} = fraction of ICU infections caused by <i>Enterococcus</i>	0.10	0.09 (Wisplinghoff et al. 2004)	FDA-CVM (2004)
P _{EF,san} = fraction of enterococcal infections caused by <i>E. faecium</i>	0.25		FDA-CVM (2004)
Fraction of enterococcal infections caused by <i>E. faecium</i> that are exogenous (non-nosocomial)	≤0.17		Cox and Popken (2004). (May be smaller now due to spread of CC-17)
Fraction of exogenous cases that are ampicillin-resistant	0.187		Willems et al. (2005)
Fraction of exogenous ampicillin-resistant cases that are vancomycin-susceptible	0.155		Jones et al. (2004)
Fraction of exogenous ampicillin-resistant vancomycin-susceptible cases possibly from food animals	0-0.069 (0.069 assumed)		Data of Leavis et al. (2006)
Fraction of exogenous ampicillin-resistant cases with penicillin-tolerant host	0.844		Lee et al. (2000)
fraction of these cases that would become ampicillin-susceptible if penicillin use in food animals were terminated	0.00-1.00 (1 is assumed)		Conservative assumption
Increase in mortality risk per case, due to ampicillin resistance	0.00-0.06 (0.06 is assumed)		Fortun et al. (2002), conservative assumption
RISK = ≤0.135 potential excess mortalities/year	315,000 * 0.10 * 0.25 * 0.17 * 0.187 * 0.14 * 0.069 * 0.844 * 0.06 = 0.135	104,372.5 * 0.09 * 0.25 *- 0.17 * 0.187 * 0.155 * 0.0- 69 * 0.844 * 0.06 ≈0.04 mortalities/year	Product of preceding factors

Source: Cox et al. 2009

Table 5.1 shows key parameter estimates, calculations, assumptions, and resulting risk estimates. When presenting point estimates, it is customary to also present interval estimates to inform decision-makers about the plausible range of estimated values. In this analysis, however, the key uncertainties have little to do with statistical sampling error, and they are not adequately characterized by confidence limits. Rather, they arise from uncertainty about the validity and conservatism of the assumptions in Table 5.1. Qualitatively, the main uncertainty is about whether a non-zero risk to human health exists from animal use of penicillin drugs. We have assumed that there is, but there is no clear empirical proof that the risk is non-zero. To bridge this knowledge gap, Table 5.1 incorporates several conservative qualitative assumptions that jointly imply that the risk is non-zero. Other quantitative parameter values presented, and their implied risk estimate of ≤ 0.135 excess mortalities/year, are intended to be realistic, data-driven values (rather than extreme upper-bounds or 95% upper confidence limits) *contingent* on these conservative qualitative assumptions. The most important conservative elements in Table 5.1 are the following qualitative assumptions:

- *Transfer of ampicillin resistance from food animal bacteria to bacteria infecting human patient occurs.* The assumption that ampicillin-resistant strains and/or determinants are transferred from strains in food animals to human ICU patients is fundamental to the assessment in Table 5.1. Such transfer has never been shown to occur, but may be possible.
- *Withdrawing animal drug use would immediately and completely prevent the problem.* Table 5.1 assumes that halting penicillin use in food animals would immediately eliminate all ampicillin resistance from the cases in Table 5.1. This is a deliberately extreme assumption. In reality, halting use might have little or no impact on the already very low levels of ampicillin resistance.
- *Resistance increases patient mortality.* The assumption that ampicillin resistance causes an increase in the mortality rates of the patients in Table 5.1 is made even though, in reality, no statistically significant difference in mortality rates has been found between resistant and non-resistant cases (Fortun et al. 2002).

With these assumptions, the calculations in Table 5.1 predict that excess mortalities per year in the entire United States population could be as high as 0.135, or about one excess mortality expected once every 7–8 years on average, if current conditions persist. This risk is concentrated among ICU patients already at high risk of such infections. With less conservative assumptions, the estimated risk falls to about 0.04 excess mortalities per year, i.e., about one excess mortality every 25 years in the United States under current conditions. The multiplicative calculation in Table 5.1 makes sensitivity analysis of these results to changes in the values of specific factors especially straightforward: the final risk estimate is directly proportional to each factor listed.

The more conservative risk estimate of 0.135 excess mortalities per year equates to an average individual risk rate in the most at-risk group (ICU patients) of approximately $0.135/315,000 = 4.3 \times 10^{-7}$ excess mortalities per ICU patient. For the United States population as a whole, this corresponds to an average

individual risk of approximately $0.135/300E6 = 4.5 \times 10^{-10}$ excess fatalities per person-year, or a lifetime risk of about $80 \times (6 \times 10^{-10}) = 3.6 \times 10^{-8}$ excess risk of mortality per lifetime (for an assumed 80-year lifetime). If the less conservative risk estimate of 0.04 excess mortalities per year is used, these individual and population risks are reduced by a factor of 0.04/0.135, or more than threefold. If one or more of the key qualitative assumptions listed above are violated, then the true risk could be as low as zero.

The main conclusion from these calculations is that not more than 0.04 excess mortalities per year (under conservative assumptions) to 0.14 excess mortalities per year (under very conservative assumptions) might be prevented in the whole U.S. population if current use of penicillin drugs in food animals were discontinued, and if this successfully reduced the prevalence of antibiotic-resistant *E. faecium* infections among intensive care unit (ICU) patients. The true risk could well be zero, if one or more of the conservative assumptions above is false.

Summary of Results from Applying Empirical Upper-Bounding Risk Assessment to Other Antibiotic-Resistant Bacteria

Antimicrobial risk analyses have now been completed for several antimicrobial-resistant bacteria of public health concern using empirical upper-bounding approaches. Among the results now available are the following.

- For streptogramins, banning virginiamycin has been estimated to prevent from 0 to less than 0.06 statistical mortalities per year in the entire United States population (Cox and Popken 2004; see also FDA-CVM 2004). More data tend to reduce such upper bounds, which in part reflect uncertainties in the data available at the time of the study.
- For macrolide-resistant campylobacter, Hurd and Malladi (2008) concluded that “the predicted risk of suboptimal human treatment of infection with *C. coli* from swine is only 1 in 82 million; with a 95% chance it could be as high as 1 in 49 million. Risks from *C. jejuni* in poultry or beef are even less.” (This analysis followed the FDA approach of interpreting simple ratios as if they applied that reducing exposures in the denominator would proportionally reduce cases in the numerator. Thus, the results may have no predictive validity if this assumption turns out to be incorrect, similar to the case of fluoroquinolones discussed by Chang et al. (2014).)
- For tetracyclines, Cox and Popken (2010) concluded that “As a case study, examining specific tetracycline uses and resistance patterns suggests that there is no significant human health hazard from continued use of tetracycline in food animals. Simple hypothetical calculations suggest an unobservably small risk (between 0 and 1.75E-11 excess lifetime risk of a tetracycline-resistant infection), based on the long history of tetracycline use in the United States without resistance-related treatment failures.”

- For MRSA, Cox and Popken (2014) “construct a conservative (plausible upper bound) probability estimate for the actual human health harm (MRSA infections and fatalities) arising from [livestock-associated] ST398-MRSA from pigs. The model provides plausible upper bounds of approximately one excess human infection per year among all U.S. pig farm workers, and one human infection per 31 years among the remaining total population of the United States. These results assume the possibility of transmission events not yet observed, so additional data collection may reduce these estimates further.”

Such quantitative risk estimates suggest that banning agricultural uses of these antibiotics might create small human health benefits (perhaps reducing compromised treatments due to resistance by a few cases per century), but are unlikely to make any measurable difference in improving public health. This finding disagrees with the passionate convictions of many experts who advocate prompt bans as urgently needed to slow the spread of resistance (Chang et al. 2014).

Since the empirical upper-bounding approach was originally developed in the early 2000s with support from the animal antibiotic industry, results such as those just cited are sometimes viewed with suspicion (*ibid.*) A virtue of quantitative risk assessment in helping to inform (and perhaps occasionally resolve) politically charged debates over what to do is that the logic, data sources, and calculations used are completely transparent and easy to summarize, as in Table 5.1, so that anyone interested can check the logic and conclusions and experiment with varying the assumptions. However, even if quantitative risk assessment proves to be too controversial to support trusted conclusions, it is often still possible to manage risks pragmatically using principles discussed next.

Managing Uncertain Food Risks via Quality Principles: HACCP

Even without QRA, it is often possible to apply process quality improvement ideas to control the microbial quality of food production processes—including both susceptible and resistant bacteria. This approach has been developed and deployed successfully (usually on a voluntary basis) using the *Hazard Analysis and Critical Control Points* (HACCP) approach summarized in Table 5.2. The main idea of HACCP is to first identify steps or stages in the food production process where bacteria can be controlled, and then to apply effective controls at those points, regardless of what the ultimate quantitative effects on human health risks may be. Reducing microbial load at points where it can be done effectively has proved very successful in reducing final microbial loads and improving food safety.

Table 5.2 Summary of seven HACCP principles

- **Analyze hazards.** Potential hazards associated with a food and measures to control those hazards are identified. The hazard could be biological, such as a microbe; chemical, such as a toxin; or physical, such as ground glass or metal fragments
- **Identify critical control points.** These are points in a food's production—from its raw state through processing and shipping to consumption by the consumer—at which the potential hazard can be controlled or eliminated. Examples are cooking, cooling, packaging, and metal detection
- **Establish preventive measures with critical limits for each control point.** For a cooked food, . . . this might include . . . minimum cooking temperature and time required to ensure the elimination of any harmful microbes
- **Establish procedures to monitor the critical control points.** Such procedures might include determining how and by whom cooking time and temperature should be monitored
- **Establish corrective actions to be taken when monitoring shows that a critical limit has not been met**—for example, reprocessing or disposing of food if the minimum cooking temperature is not met
- **Establish procedures to verify that the system is working properly**—for example, testing time-and-temperature recording devices to verify that a cooking unit is working properly
- **Establish effective recordkeeping to document the HACCP system.** This would include records of hazards and their control methods, the monitoring of safety requirements and action taken to correct potential problems. Each of these principles must be backed by sound scientific knowledge: for example, published microbiological studies on time and temperature factors for controlling food-borne pathogens

Source: USDA/FDA 1997; <http://www.cfsan.fda.gov/~lrd/bghaccp.html>

Discussion and Conclusions

This chapter has introduced and illustrated key ideas used to quantify and manage human health risks from food contaminated by bacteria, both antibiotic-resistant and antibiotic-susceptible. Somewhat similar ideas apply to other food-borne hazards, from pesticide residues to mad cow disease, i.e., risk assessment can be carried out by modeling the flow of contaminants through the food production process (together with any increases or decreases at different steps), resulting in levels of exposures in ingested foods or drinks. These are then converted to quantitative risks using dose-response functions.

The practical successes of the HACCP approach provide a valuable reminder that quantitative risk assessment (QRA) is not always a prerequisite for effective risk management. It may not be necessary to quantify a risk in order to reduce it. Reducing exposures at critical control points throughout the food production process can reduce exposure-related risk even if the size of the risk is unknown.

Where QRA can make a crucial contribution is in situations where there is doubt about whether an intervention is worthwhile. For example, QRA can reveal whether expensive risk-reducing measures are likely to produce correspondingly large health benefits. It may be a poor use of resources to implement expensive risk-reducing measures if the quantitative size of risk reduction procured thereby is very small. QRA methods such as farm-to-fork exposure modeling and dose-response modeling (Haas et al. 1999), or empirical upper-bounding approaches based on multiplicative

models (Cox 2006), can then be valuable in guiding effective risk management resource allocations by revealing the approximate sizes of the changes in human health risks caused by alternative interventions. A detailed example is given in the next chapter.

References

- Animal Health Institute (AHI) (2006) Animal health companies meet increase market need for antibiotics. News Release from the Animal Health Institute, Washington, DC. 12 Oct 2006. <http://www.ahi.org/mediacenter/documents/Antibioticuse2005REVISED.pdf>
- Bartholomew MJ, Hollinger K, Vose D (2003) Characterizing the risk of antimicrobial use in food animals: fluoroquinolone-resistant campylobacter from consumption of chicken. In: Torrence ME, Isaacson RE (eds) Microbial food safety in animal agriculture: current topics. Iowa State Press, Ames, IA, pp 293–301
- Barza M, Travers K (2002) Excess infections due to antimicrobial resistance: the “Attributable Fraction”. *Clin Infect Dis* 34(Suppl 3):S126–S130
- CBS (2010, 6/16/2010) Animal antibiotic overuse hurting humans? Katie Couric Investigates feeding healthy farm animals antibiotics. Is it creating new drug-resistant bacteria? CBS Special News Report: Katie Couric Investigates. <http://www.cbsnews.com/stories/2010/02/09/eveningnews/main6191530.shtml>
- Chang Q, Wang W, Regev-Yochay G, Lipsitch M, Hanage WP (2014) Antibiotics in agriculture and the risk to human health: how worried should we be? *Evol Appl*. <http://onlinelibrary.wiley.com/doi/10.1111/eva.12185/abstract>
- Cox LA Jr (2006) Quantitative health risk analysis methods: modeling the human health impacts of antibiotics used in food animals. Springer, New York
- Cox LA Jr (2008) Managing food-borne risks. Wiley Encyclopedia of Quantitative Risk Analysis and Assessment
- Cox LA Jr, Popken DA, Mathers J (2009) Human health risk assessment of penicillin/aminopenicillin resistance in enterococci due to penicillin use in food animals. *Risk Anal* 29 (6):796–805
- Cox LA Jr, Popken DA (2010) Assessing potential human health hazards and benefits from subtherapeutic antibiotics in the United States: tetracyclines as a case study. *Risk Anal* 30 (3):432–457
- Cox LA Jr, Popken DA (2014) Quantitative risk assessment of human MRSA risks from swine. *Risk Anal* 39(9):1639–1650
- Cox LA, Popken DA (2004) Quantifying human health risks from virginiamycin used in chickens. *Risk Anal* 24(1):271–288
- Edmond MB, Wallace SE, McClish DK, Pfaller MA, Jones RN, Wenzel RP (1999) Nosocomial bloodstream infections in United States hospitals: a three-year analysis. *Clin Infect Dis* 29 (2):239–244
- US Food and Drug Administration – Center for Veterinary Medicine (FDA-CVM) (2000) Review of agricultural antibiotics policies. http://www.fda.gov/cvm/HRESP106_157.htm#nrdc
- US Food and Drug Administration – Center for Veterinary Medicine (FDA-CVM) (2003) Guidance for industry # 152: evaluating the safety of antimicrobial new animal drugs with regard to their microbiological effects on bacteria of human health concern. US Dept of Health and Human Services, Food and Drug Administration, Center for Veterinary Medicine. <http://www.fda.gov/cvm/Guidance/fguide152.pdf>
- US Food and Drug Administration – Center for Veterinary Medicine (FDA-CVM) (2004) Risk assessment of streptogramin resistance in *Enterococcus faecium* attributable to the use of

- streptogramins in animals. Draft for Comment, 23 Nov 2004. http://www.fda.gov/cvm/Documents/SREF_RA_FinalDraft.pdf
- US Food and Drug Administration – Center for Veterinary Medicine (FDA-CVM) (2007) FDA database of approved animal drug products. FDA Center for Veterinary Medicine, VMRCVM Drug Information Lab. <http://dil.vetmed.vt.edu/>
- Fortun J, Coque TM, Martin-Davila P, Moreno L, Canton R, Loza E, Baquero F, Moreno S (2002) Risk factors associated with ampicillin resistance in patients with bacteraemia caused by *Enterococcusfaecium*. *J Antimicrob Chemother* 50(6):1003–1009
- Haas CN, Rose JB, Gerba CP (1999) Quantitative microbial risk assessment. Wiley, New York
- Hurd HS, Malladi S (2008) A stochastic assessment of the public health risks of the use of macrolide antibiotics in food animals. *Risk Anal* 28(3):695–710
- Institute of Medicine (IOM) (1989) Human health risks with the subtherapeutic use of penicillin or tetracyclines in animal feed. Report by the Committee of Human Health Risk Assessment of Using Subtherapeutic Antibiotics in Animal Feeds, Institute of Medicine, IOM-88-89. National Academy Press, Washington, DC
- Jones ME, Draghi DC, Thornsberry C, Karlowsky JA, Sahm DF, Wenzel RP (2004) Emerging resistance among bacterial pathogens in the intensive care unit—a European and North American surveillance study (2000–2002). *Ann Clin Microbiol Antimicrob* 3:14. Online journal article available at <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=509280>
- Kallen AJ, Yi M, Bulens S, Reingold A, Petit S, Gershman K, Ray SM, Harrison LH, Lynfield R, Dumyati G, Townes JM, Schaffner W, Patel PR, Fridkin SK, Active Bacterial Core surveillance (ABCs) MRSA Investigators of the Emerging Infections Program (2010) Health care-associated invasive MRSA infections, 2005–2008. *JAMA* 304(6):641–647
- Leavis HL, Bonten MJ, Willems RJ (2006) Identification of high-risk enterococcal clonal complexes; global dispersion and antibiotic resistance. *Curr Opin Microbiol* 9(5):454–460
- Lee CE, Zembower TR, Fotis MA, Postelnick MJ, Greenberger PA, Peterson LR, Noskin GA (2000) The incidence of antimicrobial allergies in hospitalized patients: implications regarding prescribing patterns and emerging bacterial resistance. *Arch Intern Med* 160(18):2819–2822
- Rice LB (2001) Emergence of vancomycin resistant enterococci. *Emerg Infect Dis* 7(2):183–187
- Sechen S (2006) The review of animal production drugs by FDA. *FDA Vet* 21(1):8–11. <http://www.fda.gov/cvm/Documents/FDAVetVolXXINo1.pdf>
- USDA/FDA (1997) HACCP principles and application guidelines. <https://www.fda.gov/Food/GuidanceRegulation/HACCP/ucm2006801.htm#princ>. Accessed 24 Apr 2018
- Varman M, Chatterjee A, Abuhamour W, Johnson WC (2006) Enterococcal infection. *Emedicine.com*. Online article available at <http://www.emedicine.com/ped/topic2703.htm> (last edited July 26, 2006)
- Webb M, Riley LW, Roberts RB (2001) Cost of hospitalization for and risk factors associated with vancomycin-resistant *Enterococcus faecium* infection and colonization. *Clin Infect Dis* 33 (4):445–452
- World Health Organization (WHO) (2005) Critically important antibacterial agents for human medicine for risk management strategies of non-human use. Report of a WHO Working Group Consultation, Canberra, Australia, 15–18 Feb 2005. http://www.who.int/entity/foodborne_disease/resistance/FBD_CanberraAntibacterial_FEB2005.pdf
- Willems RJ, Top J, van Santen M, Robinson DA, Coque TM, Baquero F, Grundmann H, Bonten MJ (2005) Global spread of vancomycin-resistant *Enterococcusfaecium* from distinct nosocomial genetic complex. *Emerg Infect Dis* 11(6):821–828
- Wisplinghoff H, Bischoff T, Tallent SM, Seifert H, Wenzel RP, Edmond MB (2004) Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clin Infect Dis* 39(3):309–317

Chapter 6

Quantitative Risk Assessment of Human Risks of Methicillin-Resistant *Staphylococcus aureus* (MRSA) from Swine Operations



Describing quantitatively *how large* a risk is provides crucial information for helping to set risk management priorities. This chapter applies descriptive analytics to assess the size of the human health risks from a particular source. It continues the theme begun in Chap. 5 of examining human health risks from antibiotic-resistant infectious bacteria, but focuses specifically on long-standing concerns expressed by the public health community, news media, and members of the general public that methicillin-resistant *Staphylococcus aureus* (MRSA) transmitted from pigs to humans may harm human health. For readers not interested in details of this application, a brief summary is that previous studies of the prevalence and dynamics of swine-associated (ST398) MRSA have sampled MRSA at discrete points in the presumed causal chain leading from swine to human patients, including sampling bacteria from live pigs, retail meats, farm workers, and hospital patients. This chapter integrates available data from several sources to construct a conservative (plausible upper-bound) probability estimate for the quantitative human health harm (MRSA infections and fatalities) arising from ST398-MRSA from pigs. The estimated plausible upper bounds are approximately one excess human infection per year among all U.S. pig farm workers, and one human infection per 31 years among the remaining total population of the U.S., assuming that bacteria transmission events not yet observed are possible. The true risks may be smaller (possibly zero for members of the general population). Putting rough numerical bounds on the size of the risk such as “less than one case per 300 million people per decade” can help to engage System 2 thinking (cognitive, slow, deliberative; see Chap. 12) about costs and benefits of risk-reducing interventions; whereas leaving the risk unquantified is more likely to engage System 1 thinking (intuitive, quick, emotional; see Chap. 12), which typically reacts very strongly and adversely to the qualitative idea of contamination of food with invisible, harmful microbes.

Introduction: How Large Is the Human Health Risk from MRSA in Swine and Pork?

Recent news stories, and many web sites and blogs, have warned that antibiotic-resistant bacteria from pig farms may threaten public and worker health. One widely discussed concern is that animal antibiotics used on farms create selection pressures that favor the spread of antibiotic-resistant “superbugs,” such as methicillin-resistant *Staphylococcus aureus* (MRSA) (CBS 2010). However, these stories have lacked a quantitative discussion of *how many* excess deaths, treatment failures, or days of illness each year are caused in the United States by MRSA arising from pig (or other livestock) production. Most human cases of MRSA arise from antibiotic use (and inadequate hand washing and infection control) in hospitals, and mainly affect hospitalized patients with severely compromised immune systems. How many cases of MRSA infections and fatalities per year arise among hospital patients, butchers, slaughterhouse workers, farmers, or the general public from livestock operations, meat handling, and consumption therefore remains to be addressed. Responsible risk management is supported best by understanding how large the human health risks are now, and how much they would be changed by proposed interventions.

Potential Human MRSA Hazards Related to Pigs and Pork

This section takes inventory of what is known about adverse human health outcomes which might be caused by exposure to MRSA from swine, and pork and identifies data on the numbers and types of adverse health consequences per year which might be caused by swine-associated (especially, sequence type 398 or ST398) MRSA. The purpose is to identify sources of risk (i.e., hazards) which pose non-negligible human health risks, so that these can be quantified further.

Consumer Exposure to MRSA via Pork Meat Poses Little Risk

Consumption of pork meat raises the logical possibility of three types of exposure and harm from MRSA: invasive infections by MRSA acquired from meat; non-infectious food poisoning by MRSA enterotoxins (as with other strains of *S. Aureus*); and colonization (with the possibility of subsequent opportunistic infection) by MRSA from pork products. Each of these is briefly considered next. We find that only the last, colonization, appears to pose a potentially significant risk in the U.S.

- *Invasive infection by MRSA consumed in pork products.* No case of a livestock-associated (ST398) MRSA *infection* transmitted via the food supply (or otherwise) has ever been found in the United States (Limbago 2010).

Worldwide, only one case of invasive infection by ST398 MRSA after ingestion of contaminated food has ever been identified, in a highly immuno-compromised individual in a Dutch hospital (Kluytmans 2010). Given the large quantities of pork consumed for decades without other cases being identified, the threat to the U.S. population for this pathway appears to be near zero.

- *Enterotoxicosis by ST398 MRSA.* Although staphylococcal enterotoxicosis is a common form of food poisoning caused by *Staphylococcus aureus* toxins in food, the pig-associated (ST398) strains of MRSA do not produce these toxins, as they lack the enterotoxin genes needed to express them (Argudin et al. 2009). No association between consumption of pig meat and increased risk of such food poisoning has been found (EFSA 2009). Thus, the threat of food poisoning by ST398 MRSA in pork products also appears to be insignificant or non-existent.
- *Colonization with ST398 MRSA encountered via the food handling process.* Although the probability of this type of colonization appears to be very low for the average consumer, as discussed next, it is somewhat higher for people in meat-handling professions such as butchers and slaughter house workers (Van Cleef et al. 2010; Gilbert et al. 2012). This assessment therefore includes quantification of the risks of colonization arising from food handling.

Direct Exposure to Pigs Can Increase Risk of Colonization with MRSA

Several studies have found elevated risk of MRSA colonization in people working closely with pigs and other livestock (Feingold et al. 2012), including farm workers (Smith et al. 2013), veterinarians (Cuny et al. 2009), and slaughterhouse workers (Van Cleef et al. 2010). However, subsequent infections have not been documented in the U.S. (Limbago 2010; IARTF 2011), although there have been some reports of minor soft tissue infections in other countries (Declercq et al. 2008; Denis et al. 2009). Because MRSA colonization has been observed among pig farm workers on MRSA positive farms, even in the U.S. (Smith et al. 2013), the possibility of subsequent infections cannot be ruled out. These risks, too, are therefore quantified in the following sections.

Hospital Outbreaks of ST398 MRSA Are Extremely Rare

The only known death associated with ST398 MRSA occurred in a medically compromised person in a Spanish hospital who sometimes worked on a pig farm (Lozano et al. 2011). There have also been limited reports of person-to-person transmission of ST398 MRSA in Dutch hospitals (van Rijen et al. 2008; Bootsma et al. 2011); however, there has been only one reported instance of an outbreak of

ST398 MRSA among patients (Wulf et al. 2007), and the study report did not clearly differentiate between colonization and infection. Because person-to-person transmission has been suggested, the possibility of future outbreaks remains, and is quantified in the following sections.

Community Outbreaks of ST398 Have Not Been Observed

No community outbreaks of ST398 MRSA have been documented, so community outbreaks are probably even less likely than hospital outbreaks. This may be due to the relatively transient nature of most ST398 colonizations. Although some authors have recently argued that living near pigs raises the risks of colonization (Feingold et al. 2012), the statistical methodology behind such ecological associations is suspect, and such reported associations are often meaningless—random spatial trends alone can create statistically significant regression coefficients; and spatial location can be a confounder for spatially distributed exposure and response variables (Cox et al. 2013). Due to its extremely low likelihood, we have not attempted to compute a risk for community outbreaks of ST398 MRSA.

ST398 MRSA Is Found in Retail Pork

Turning to food handling hazards, MRSA, including ST398, is occasionally found in retail meats including pork. Recent studies conducted in various parts of the US tested retail pork products for the presence of ST398 MRSA. Table 6.1 summarizes relevant data from these studies.

MRSA Colonization from Food Handling Is Possible

A study in the Netherlands of workers who came in frequent contact with raw meat (institutional kitchens and cold meat processing) found no MRSA colonization in 95 participants (de Jonge et al. 2010) although MRSA was present on 2 of 10 (20%)

Table 6.1 ST398 MRSA prevalence data from retail pork in the U.S.

Study	Samples	ST398 MRSA positive	Rate
Kelman et al. (2011)	300	1	0.0033
Waters et al. (2011)	26	1	0.0385
Davies (2009)	143	7	0.0490
Molla et al. (2012)	135	5	0.0370
O'Brien et al. (2012)	395	26	0.0658
Overall	999	40	0.0400

of pork samples from the same locations and 3 of 26 samples of veal and chicken. A recent study of butchers in Hong Kong did find cases of MRSA colonization (Boost et al. 2012). Seventeen of 300 butchers (5.7%) carried MRSA, and ten of these carried strains commonly found in local swine (3.3% of total). A separate study of pig carcasses at markets in Hong Kong found that 16 of 100 carcasses tested positive for swine-associated strains of MRSA (Guardabassi et al. 2009). However, MRSA carriage in the Hong Kong general population is rare (~1.4%) and is largely associated with working in health care (O'Donoghue and Boost 2004). Therefore, it is plausible that meat handling can be a risk factor for MRSA colonization in Hong Kong, and possibly elsewhere.

In summary, our hazard identification suggests that the major non-negligible risks from pig-associated MRSA arise from colonization, with potential for subsequent infection. The following sections seek to quantify these risks.

Quantifying Pig-Associated MRSA Colonization Potential

Quantitative Estimation of Colonization Potential for Professional Food Handlers in the U.S. from Pork Meat

To quantify uncertainty about the true proportion of ST398 MRSA-colonized food handlers in the U.S., we apply Bayesian conjugate prior analysis. We first assume a standard uniform prior distribution for the true but unknown proportion of colonized food handlers, and then adjust this starting assumption by conditioning on the Dutch data above (showing 0 colonizations among 95 workers). The uniform prior distribution is a deliberately conservative assumption, in that it implies that, in the absence of data, half of all food handlers are expected to be colonized, which is an order of magnitude greater than the empirically observed fractions just reviewed. Thus, we begin with a deliberately exaggerated (risk-over-estimating) assumption, and then use data to revise it downward via Bayes' Rule. As the most recent and relevant available data source, we use the Netherlands study (de Jonge et al. 2010) showing $s = 0$ positive observations among $n = 95$ workers. Bayesian conditioning of the uniform prior on these data yields a posterior distribution (a Beta($s + 1$, $n - s + 1$) = Beta(1, 96) posterior distribution) with a mean of $(s + 1)/(n + 2) = 0.0103$.

To determine a fraction attributable to pork handling alone, we note that pork comprises approximately 55% of total meats consumed in the EU (USCB 2012). MRSA prevalence on pork in the Netherlands is actually slightly *lower* than the average for all retail meats. A large-scale Dutch study found MRSA on 264 of 2217 (11.9%) total retail meat samples, 10.7% of retail pork samples, but even higher rates on several other retail meats: turkey (35.3%), chicken (16.0%), veal (15.2%) and also beef (10.6%), lamb and mutton (6.2%), fowl (3.4%) and game (2.2%) (de Boer et al. 2009). Based on these figures, it is reasonable to assume that somewhat less

than 55%, perhaps roughly 50%, of MRSA prevalence among Netherlands food handlers could be due to pork. To account for a high degree of uncertainty, we will further assume that the true fraction can be described by a uniform probability distribution with 0.50 as a mean, and [0.30, 0.70] as a plausible range. These endpoints are somewhat arbitrary, but suffice to indicate that the true fraction might differ from the point estimate of 50%.

For the U.S., Table 6.1 summarizes data from recent retail pork sampling studies. ST398 MRSA prevalence in retail pork in the U.S. is about 37.38% (0.04/0.107) of that in the Netherlands. In addition, pork as a fraction of total meat processed in the U.S. is about half (49.11%) (0.2720/0.5538) of that in the EU (USCB 2012). A base case estimate of ST398 MRSA prevalence from pork among U.S. meat handlers is thus about (0.0103 posterior mean for the fraction of food handlers colonized in Netherlands) \times (0.50 fraction from pork) \times (0.3738 ratio of prevalence in U.S. pork compared to prevalence in Netherlands pork) \times (0.4911 pork-processing/consumption ratio) \approx 0.0009.

According to the U.S. Bureau of Labor Statistics, in 2010 there were approximately 382,000 persons in the U.S. employed as “Butchers and Other Meat, Poultry, and Fish Processing Workers” (includes Slaughterers and Meat Packers) (USBLS 2012). Based on the estimates above, we would expect that about $382,000 \times 0.0009 = 357$, would be colonized with MRSA attributable to handling pork.

Quantitative Estimation of Colonization Potential from Consumer Food Handling

This section applies a series of multiplicative factors to convert the MRSA colonization risk among professional meat handlers to a risk to those consumers who also handle pork, based on relative exposure durations per year. Pork is eaten approximately 98.5 times annually per capita in the United States. Approximately 21% of that pork is from fresh cooked products, with another 31% coming from ham (NPPB 2010), implying an average of approximately 51.22 servings of fresh pork or ham per person per year. An average household size of 2.58 (per 2011 U.S. census data) gives an average of about 51.22 fresh servings per person-year/2.58 person-years per household-year \approx 19.85 raw pork preparation events per household per year. If each preparation event requires the food handler to be in contact with pork, pork juices, or related working surfaces for up to 15 min (a generous estimate), then the average annual contact time per preparer is approximately 4.96 h. Comparing this to a professional meat handler, who is likely to be in contact with meat, juices, and related working surfaces for approximately 2000 h/year, approximately 0.2720 of which is related to pork (as previously discussed), the average U.S. consumer preparer has approximately $4.96/(2000 \times 0.2720) = 0.0091$ as much exposure time to pork as a professional meat handler. Assuming that colonization events

(e.g., from a concentration of MRSA on meat) are relatively rare, and independent from exposure to exposure, a Poisson model for colonization is appropriate, and colonization risk increases approximately linearly with exposure time. The implied plausible upper bound colonization risk to the average consumer preparer is then $(0.0028 \text{ MRSA colonizations from pork per worker year for professional meat handlers}) \times 0.0091 = 2.55 \text{ colonizations per 100,000 consumer preparer-years}$ (2.55×10^{-5}). This does not attempt to correct for any differences in MRSA exposure concentrations or immunity between professional and consumer meat handlers, but assumes that either can be colonized if suitable rare conditions (e.g., high concentrations of MRSA) are encountered on the meat being handled. Applied to the approximate current number of U.S. households, approximately 315M/2.58 gives an estimate of $(315,000,000/2.58) * (2.55 \times 10^{-5}) = 3111$ additional annual colonizations (but not infections) of ST398 MRSA due to consumers handling pork. To place this in context, the best estimate of total MRSA colonizations in the U.S. is approximately two million (Graham et al. 2006).

Table 6.2 assembles the above distributions and factors into a probability model whose output is a distribution of probabilities for the average annual per person risk of being colonized with MRSA from food handling.

The corresponding model equations are:

$$\text{Food Handling Colonizations} = \begin{pmatrix} \text{Dutch food handler prevalence} \times \text{pork attributable fraction} \\ \times \text{U.S. prevalence ratio} \times \text{U.S. processing ratio} \\ \times (\text{U.S. food workers} + \text{Consumer exposure time ratio} \times \text{U.S. population}) \end{pmatrix} \quad (6.1)$$

The mean value for the equation is approximately $(9.45\text{E-}4) * 382,000 + (0.86\text{E-}6) * 122,093,000 = 361$ occupational cases of colonization + 1050 consumer cases of colonization ≈ 1411 total cases of ST398 MRSA colonization from pork meat per year in the U.S.

Table 6.2 Probability model for colonization risk from handling pork

Component	Distribution	Mean
Dutch food handler prevalence	Beta(1, 96)	0.0103
Pork attributable fraction in Netherlands	Uniform(0.30, 0.70)	0.50
U.S. to Netherlands pork prevalence ratio	Constant (ratio of survey study averages)	0.3738
U.S. to Netherlands pork processing ratio	Constant (ratio of historical processing fractions)	0.4911
Average U.S. Meat Handler Risk/year/person	Product of Above (apply to 382,000 workers)	9.45E-4
Consumer to food handler time of exposure ratio	Constant (U.S. pork preparation rates)	0.0091
Average U.S. Consumer risk/year/person	Meat Handler Risk * .0091 (apply to 315M/2.58 households)	8.6E-6

Table 6.3 ST398 MRSA colonization among pig farm workers

Location	# Positive	# Farm workers tested	Reference
Ontario, Canada	5	9	Khanna et al. (2008)
Germany	97	113	Cuny et al. (2009)
Belgium	47	94	Denis et al. (2009)
Iowa and Illinois	9	14	Smith et al. (2009)
Iowa and Illinois	27	51	Smith et al. (2013)
Total	185	281	

ST398 MRSA Colonizes Pig Farm Workers

ST398 MRSA in pigs can transiently colonize pig farm workers. Cases of ST398 MRSA transient colonization, with occasional cases of infection (the latter observed only in Europe), occur mostly in individuals having direct contact with livestock, especially pigs. Table 6.3 summarizes the available worldwide data on the proportion of colonized pig farm workers on farms where pigs have been colonized.

Pooling these samples for workers on MRSA-positive farms yields an empirical ratio of 185/281. To approximate uncertainty regarding the true proportion, we again use a Bayesian updating of a uniform prior distribution as described earlier. Accordingly, the probability that a pig farm worker in direct contact with pigs is colonized with MRSA is estimated by a $\text{Beta}(s + 1, n - s + 1) = \text{Beta}(186, 97)$ posterior distribution, with a mean of $(s + 1)/(n + 2) = 186/283 = 0.66$ and a standard deviation of 0.028. This mean and standard deviation make a normal approximation applicable, as the endpoints (0 and 1) are many standard deviations away from the mean.

Estimating the Number of U.S. Pig Farm Workers

To estimate the number of individuals in the U.S. in close contact with pigs, it is necessary to consider the number of pig farms by herd size, since the average number of workers per pig is known to decrease with herd size. Otto et al. (1998) performed an economic analysis of Iowa hog production that estimated direct workers required (for farrow-to-finish operations) as a function of herd size as follows: 150 pigs—1.4 workers, 300 pigs—3 workers, 1200 pigs—10 workers, and 3400 pigs—21 workers. We obtained the latest available values for the number of herds by size from the National Agricultural Statistical Services (USDA-NASS 2009). Using approximation and interpolation, we set the labor requirements within the USDA herd size breakouts as shown in column 2 of Table 6.4. Subsequent calculations shown in the table determine the estimate for total workers on pig farms in the U.S.

To model uncertainty, we assume that this estimate of 326,505 for the risk population may be off by as much as 20% (approx. 65,301) in either direction.

Table 6.4 Computation of total pig farm workers in the U.S.

Herd size	Workers/herd	Percent of herds	Number of herds	Total workers ^a
1–199	1.4	73.05	55,110	77,155
200–499	3	6.00	4524	13,572
500–999	6.5	4.76	3588	23,322
1000–1999	10	5.32	4013	40,130
2000–4999	21	7.10	5356	113,476
5000+	21	3.78	2850	59,850
Total		100.0	75,442	326,505

^aWorkers/Herd × Number of Herds

Table 6.5 MRSA positive farms in the U.S. and Canada

Location	# Farms MRSA positive	# Farms tested	Reference
Ontario, Canada	9	20	Khanna et al. (2008)
Canada	2	28	Weese et al. (2009)
Canada	5	46	Weese et al. (2011)
Midwestern U.S.	12	40	Frana et al. (2013)
Iowa/Illinois	4	9	Smith et al. (2013)
Minnesota	0	9	Davies (2010)
Ohio/N. Carolina	0	6	Smith et al. (2013)
Total	30 (19%)	158	

This subjective uncertainty is expressed as a uniform probability distribution ranging from 261,204 to 391,806. This range is admittedly subjective—other ranges could be considered—but it suffices to explore the sensitivity of results to uncertainty in the size of the heavily exposed worker population.

Estimating the Proportion of Farms with MRSA

No wide-scale survey of MRSA prevalence on farms has been performed in the U.S. However, Table 6.5 summarizes recent relevant data points for the U.S. and Canada. (ABF farms were not included.)

The overall proportion of farms in the U.S. and Canada positive for ST398 MRSA is $30/158 = 19\%$. To approximate uncertainty regarding the true proportion, we again use Bayesian updating of a uniform prior distribution, yielding a Beta (31, 129) distribution, with a mean of $(s + 1)/(n + 2) = 31/160 \approx 0.19$, for the distribution of the fraction of MRSA positive farms.

Table 6.6 MRSA colonization model parameters

Parameter	Distribution	Mean value
Number of U.S. pig farm workers	Uniform(261,204–391,806)	326,505
Fraction of MRSA positive farms	Beta(38, 126)	0.19
P(Colonization MRSA positive farm)	Normal(0.65, 0.028)	0.66
Estimated U.S. colonizations	Product of above	40,944

Probability Model for ST398 MRSA Colonization

Table 6.6 assembles the probability distributions derived above into a product-form probability model for the number of annual ST398 MRSA colonizations among U.S. pig farm workers.

The equation for the model is:

$$\begin{aligned} \text{Farm Worker Colonizations} = & \text{No. Pig farm workers} \\ & \times \text{Fraction MRSA positive farms} \quad (6.2) \\ & \times P(\text{Colonization}| \text{MRSA positive}) \end{aligned}$$

Estimating the Annual Probability of MRSA Infection for Those Colonized

Transient colonization of the human nasal passages (and other mammals) with *Staphylococcus aureus* is common in the U.S., and is usually harmless. Gorwitz et al. (2008) estimated a prevalence of 28.6% in the U.S. in 2004. A fraction of these are methicillin-resistant, approximately 1.5% of the U.S. population. Fortunately, only a small fraction of those colonized with MRSA at any moment develop active MRSA infections. Although the relationship between colonization and infection is not fully understood, in one instance researchers have shown that more than 80% of bloodstream infections caused by *S. aureus* in hospitalized adults were preceded by colonization of the anterior nares with the same strain (Graham et al. 2006). For initial approximation purposes, we assume that the relationship of infected to colonized can be expressed as a fraction, that is, the number of infected/year is a given fraction of those colonized.

An approximate infection rate can be estimated from a MRSA surveillance program conducted in Iowa during and following the initial detection of ST398 MRSA in swine. In a comprehensive study of 1166 MRSA isolate submissions from this program (invasive cases only) from 1999 to 2006, no ST398 strains were found (Van De Griend et al. 2009). 343 isolates were from the latest year, 2006. According to a report from the same laboratory (IARTF 2011), in the years 2007–2009, the following additional numbers of MRSA isolates were submitted to the Iowa

University Hygienic Lab by year: 445, 447, and 455. The report states: “To date, MRSA CC398 [i.e. ST398] has not been identified among the invasive MRSA isolates submitted to SHL [State Hygienic Laboratory] from Iowa residents.” (Surveillance ended 1/1/2011 and data after 2009 was not provided.) Thus, while ST398 appears likely to be not uncommon among Iowa swine, there are no reported cases of invasive or even soft-tissue infection cases of ST398 in Iowa through the time of the most recent studies. Iowa accounts for 28% of U.S. pork production (USDA 2008). This finding is replicated at the national level. The CDC has collected over 12,000 MRSA isolates over recent years that include colonization isolates, infecting isolates, and isolates from animals and from food. The vast majority were collected through the Active Bacterial Core surveillance (ABCs) system (www.cdc.gov/abcs/index.html). However, no ST398 strains have been detected (Limbago 2010).

Probability Model for Infection Given Colonization

If ST398 MRSA was widespread among Iowa hogs by 2006 (Smith et al. 2008), then for a period of 4 years, 2006–2009, and likely longer, there were no detected invasive ST398 MRSA cases in Iowa despite the simultaneous significant prevalence among Iowa swine herds and pig farm workers. To put an upper bound on the probable human health risk that is consistent with these observations, we first estimate the mean number of human colonizations of ST398 MRSA in Iowa from 2006 to 2009 as:

$$\begin{aligned} \text{Iowa Colonization Years} = & \quad 4 \text{ years} \times 0.28 \text{ of U.S. pigs in Iowa} \\ & \times \text{U.S. Farm Worker Colonizations (from Eq. 6.2)} \end{aligned} \tag{6.3}$$

To obtain a distribution of results corresponding to Eq. 6.3, we generated 100,000 random samples from the previously described distributions of uncertain inputs. The mean of the distribution was 55,114. If an estimated 55,114 ST398 MRSA colonization-years in Iowa produced 0 invasive cases, then, using a conservative Bayesian analysis similar to that described earlier, we can estimate a plausible upper bound for the annual posterior probability of infection given colonization as belonging to a beta(1; 55,115) distribution with mean 1/55,116 = 1.81E-5.

Estimating Secondary Cases

While transmission of MRSA among hospital patients is a significant health concern worldwide, the secondary case rate for ST398 MRSA appears to be much lower than that for other types of MRSA. This may be due to the relatively transient nature of ST398 colonizations (van Cleef et al. 2009; Graveland et al. 2011; Frana et al. 2013).

Hospital Cases

Wulf et al. reported an instance of an apparent outbreak of ST398 MRSA in a Dutch hospital (Wulf et al. 2007). Bootsma et al. (2011) identified cases of ST398 transmission from patient to health care workers (2 “outbreaks” involving colonization of 1 and 2 workers) in an examination of two large datasets from Dutch hospitals. This and other anecdotal instances reported in Europe indicate that secondary infections are possible.

Van Rijen et al. (2008) compared the secondary transmission rates for typable MRSA versus non-typable MRSA (presumably ST398) in a Dutch hospital. They reported that “Sixteen patients who carried typable MRSA stayed in the hospital without precautions, for a total of 138 days. Twenty-two of 2139 persons exposed to these 16 patients were shown to be colonized with the index strain. For non-typable MRSA, during 37 exposure days for 8 patients, 0 of the 408 exposed patients and health care workers were colonized. . . . Only recently, in 2007, 1 health care worker was colonized with non-typable MRSA, acquired from a patient who had not been treated in isolation.”

Hospital Estimation Model

Based on these data, a conservative upper bound on the secondary case rate, using Bayesian analysis again, would be that the transmission rate of ST398 MRSA within the hospital is approximately $(1/410)/(22/2139) = 0.238$ that of non-ST398 strains. A similar result was reported in Wassenberg et al. (2011) who computed a relative transmission risk of 0.28. A somewhat lower relative risk was obtain by (Bootsma et al. 2011), with 0.169. The Van Rijen et al. data further implies that the average length of stay for ST398 MRSA was $(37/8) = 4.625$ days versus $(138/16) = 8.625$ days for non ST398 MRSA, a reduction ratio of $4.625/8.625 = 0.536$. Wassenberg et al. obtained values of 7 days for ST398 infections and 8 days for non ST398 (.875 reduction ratio). As discussed in the previous section, the Van Rijen et al. results also imply a rate of conversion from colonization to infection in ST398 strains that is approximately $(1/4.83) = 0.207$ that of non ST398 strains. Wassenberg et al. obtained values of 13 days of exposure before transmission for ST398 cases and 3 days of exposure for non ST398 case, implying a ratio of $3/13 = 0.231$ between the two.

We used this data on hospitalization infection dynamics and the mathematical framework of Webb et al. (2009), who developed a set of differential equations describing the epidemiological dynamics of MRSA in U.S. hospitals. Their motivation was to determine the conditions under which community acquired MRSA (CA-MRSA) would displace hospital acquired MRSA (HA-MRSA) as the dominant infection. They derived a basic reproduction number, R_0 , which corresponds to the steady state number of secondary infections per initial infection. Using baseline empirical parameter values, they computed an R_0 for community associated

Table 6.7 Adjustment factors (right column) applied to a community-acquired MRSA hospital dynamics model to obtain parameters for ST398 MRSA hospital dynamics

Parameter group	Derived from Van Rijen et al.	Wassenberg et al.	Bootsma et al.	Avg.
β_{xy} —transmission rates	0.238	0.28	0.169	0.229
φ_{xy} —infection rates	0.207	0.231	NA (colonization only)	0.219
γ_{xy} —discharge rates (1/length of stay)	1/1.536	1/0.875	NA (no distinction made)	1/0.706

MRSA (R_0^C) of approximately 0.66, and 0.69 for hospital-associated MRSA R_0^H . To model the hospital dynamics of swine-associated (ST398) MRSA, we modified the Webb et al. baseline model for community-associated MRSA based on averages of the parameters discussed above, as summarized in Table 6.7. To account for uncertainty, we created a stochastic simulation version of the hospital dynamics model. This assumed that input parameters follow log-normal distributions ($\mu = \text{avg. value}$, $\sigma = \mu/2$) with the exception of the hand hygiene compliance fraction, which was distributed uniformly between 0.40 and 0.80 versus an original baseline value of 0.60.

100,000 iterations of the simulation model yielded a median reproductive rate for community-associated ST398 MRSA, R_0^C , of approximately 0.0787 (probability $>1 = .0016$), and a median R_0^H value (hospital associated) of approximately 0.0823 (probability $>1 = .0019$). These are less than 1/8 of the corresponding values for non ST398 MRSA when using the same model with the original baseline values. Our QRA model uses a midpoint value of 0.0805 to capture the secondary hospital case rate.

Quantitative Risk Analysis Model

Equations 6.1–6.3, together with a factor for secondary cases, as derived above, can be combined into a probabilistic model for total MRSA infections in the U.S. attributable to pork. The model can be expressed as

$$\begin{aligned} &\text{Expected Number of Annual Infections} \\ &= (\text{Pork Handling Colonizations} + \text{Farm Worker Colonizations}) \quad (6.4) \\ &\times P(\text{infection} \mid \text{colonization}) \times (1 + \text{Secondary Case Rate}) \end{aligned}$$

We implemented the simulation model in the R statistical programming environment (<http://www.r-project.org/>) and generated 100,000 random values for each underlying probability distribution to obtain 100,000 random values for Food Handling Colonizations (Eq. 6.1), U.S. Farm Worker Colonizations (Eq. 6.2), and Infection Rates (Eq. 6.3). The secondary case rate of 0.0805 was determined via the

separate model just discussed (c.f. Table 6.7). Equation 6.4, applied to these randomly sampled input values, yielded a distribution of U.S. annual MRSA infections attributable to pigs and pork.

Results

The mean number of meat handler colonizations in the simulation was 358.54 [95% CI—17.29, 1108.43]. Figure 6.1 shows its uncertainty distribution. The mean number of colonizations for pork consumers was 1042.36 [95% CI—50.29, 3182.38], and Fig. 6.2 shows its uncertainty distribution. The mean number of colonizations for pig farm workers was far larger than either of these, at 41,777.03 [95% CI—28,967.03, 56,831.16]. Its uncertainty distribution is plotted in Fig. 6.3. Thus, the number of colonizations of farm workers exceeded that of pork handlers and consumers combined by a factor of about 30. The incidence of colonization among pig farm workers constitutes 96.8% of the total risk pool.

The conditional infection rate for those colonized with pork attributable MRSA is modeled by a Beta(1, 55,115) distribution as described previously. It is shown in Fig. 6.4. Finally, the distribution for the total annual number of pig/pork attributable MRSA infections (Eq. 6.4) is shown in Fig. 6.5. It has a mean of 1.00 [95% CI—0.05, 3.05]. If we allocate the mean according to the proportions of the colonization risk pools, the expected total number of annual infections in U.S. pork consumers is about 0.024/year; in professional meat handlers, about 0.008/year; and in pig farm workers, about 0.968/year.

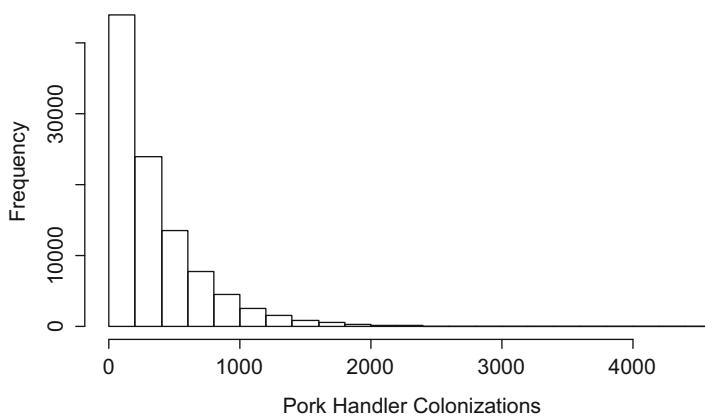


Fig. 6.1 Distribution of annual MRSA colonizations attributable to pork among professional meat handlers

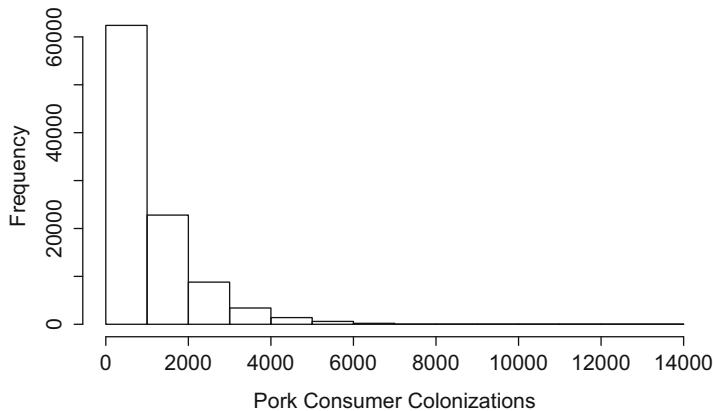


Fig. 6.2 Distribution of annual MRSA colonizations attributable to pork among pork consumers

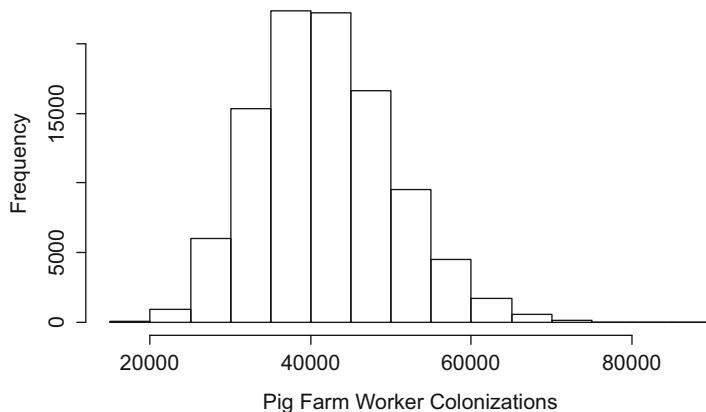


Fig. 6.3 Distribution of annual MRSA colonizations attributable to pigs among pig farm workers

Discussion and Conclusions

The size of the public health risk in the U.S. caused by MRSA from swine has not previously been quantified. Popular news stories have mentioned it in the context of 70,000 excess deaths per year from antibiotic-resistant superbugs (CBS 2010). Our conservative quantitative risk assessment indicates that MRSA from pigs and pork should be expected to cause no more than *about one infection per year in the U.S. population*, almost all among pig farm workers, under current conditions. To consumers (the general public) and professional meat handlers combined, swine- and pork-associated MRSA pose a risk of not more than *about one excess infection per 31 years* under current conditions. This corresponds to an average

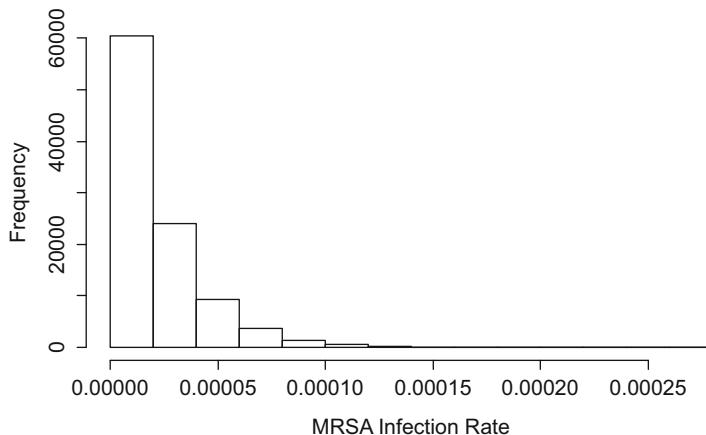


Fig. 6.4 Distribution of the MRSA infection rate for those colonized with pig/pork attributable MRSA

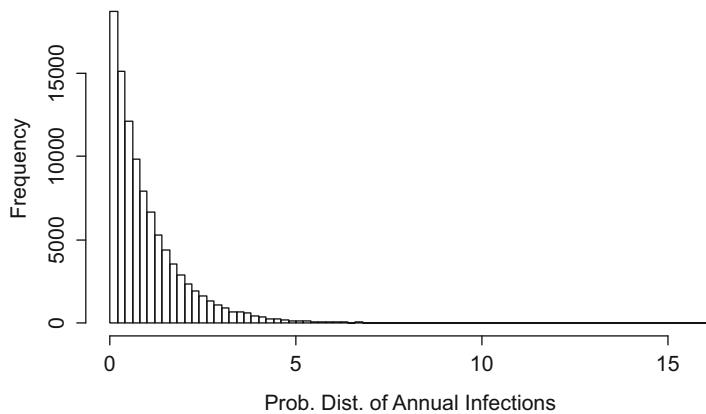


Fig. 6.5 Distribution of total annual U.S. MRSA infections attributable to pigs/pork

per-capita risk of about one case per ten billion person-years (i.e., 1 case/(315 million people \times 31 years) for the general public. Most such infections are treatable, and the excess death rate would be too small to detect. This is consistent with the historical fact that no human deaths and no serious infections have been found to have been caused among the general public or professional meat handlers by pig-associated MRSA. The fraction of all cases caused by use of antibiotics in swine is likewise undetectably small, but the finding of no statistical difference in MRSA rates between meat from hogs raised conventionally and meat from hogs raised without antibiotics (O'Brien et al. 2012) suggests that reducing antibiotic use on farms should not be expected to reduce the already small risk further.

The true risk is likely to be considerably smaller than our conservative estimates; it could be zero. We have assumed that certain events, such as invasive infection resulting from ordinary colonization (from meat handling or otherwise), can occur, even though they have not been observed. The odds of their actual occurrence have become smaller as more data has accumulated, since Bayesian conditioning on data shifts our conservative (uniform) prior distributions leftward. In addition, modeling the dynamics of MRSA in hospitals shows that ST398 MRSA risks in hospitals are unlikely to increase dramatically in future, as the ST398 type has a relatively low potential for spread (basic reproductive rate). Thus, a conservative upper-bound occupational risk of about 1 infection per pig worker per year, and a public health risk of about 1 infection per 31 years among the rest of the U.S. public, or about one excess case per 10 billion person-years, appears to be justified by current data. These estimates may continue to decline if further surveillance data accumulate showing no observed cases of ST398 infections among the general public.

The calculations in this chapter have emphasized quantitative description of the size of the human health risks from pig-derived MRSA. Similar rough upper-bound estimates of risks can also be obtained by multiplying plausible upper-bound estimates for appropriate factors in many other areas of health, safety, and environmental (HS&E) risk assessment and engineering reliability analysis. (Equation 2.11 in Chap. 2 presents a general decomposition of a probability of interest as a product of marginal and conditional probabilities in the context of Bayesian network analysis.) If the resulting upper-bound risk estimates are small enough for the risk to be safely ignored in favor of more productive risk-reducing opportunities, then such simple descriptive analyses may suffice to show that no further analyses or actions to manage these risks are currently warranted. But if the risks appear to be large enough so that it is worth considering costly interventions to manage them, then it becomes valuable to be able to predict the probable consequences of different interventions and to characterize uncertainties about them. Chapter 8 illustrates how simple calculations, together with causal attribution of observed consequences to controllable factors, can be used to accomplish such predictive modeling.

References

- Argudin MA, Fetsch A, Tenhagen BA, Kowall J, Hammerl J, Kaempe U, Hertwig S, Schroter A, Braunig J, Kasbohrer A, Appel B, Nockler K, Helmuth R, Mendoza MC, Rodicio MR, Guerra B (2009) Virulence and resistance determinants in methicillin-resistant *Staphylococcus aureus* ST398 isolates. In: 19th European congress of clinical microbiology and infectious diseases. Diseases, ESCMID Helsinki, Finland, Blackwell Publishing
- Boost M, Ho J, Guardabassi L, O'Donoghue M (2012) Colonization of Butchers with livestock-associated methicillin-resistant *Staphylococcus aureus*. Zoonoses Public Health. <https://doi.org/10.1111/zph.12034> [epub ahead of print] (Available at Last accessed on 1/28/2013)
- Bootsma MC, Wassenberg MW, Trapman P, Bonten MJ (2011) The nosocomial transmission rate of animal-associated ST398 methicillin-resistant *Staphylococcus aureus*. J R Soc Interface 8 (57):578–584

- CBS (2010, 6/16/2010) Animal antibiotic overuse hurting humans? Katie Couric Investigates feeding healthy farm animals antibiotics. Is it creating new drug-resistant bacteria? CBS Special News Report: Katie Couric Investigates. <http://www.cbsnews.com/stories/2010/02/09/eveningnews/main6191530.shtml>
- Cox LA, Popken DA, Berman DW (2013) Causal versus spurious spatial exposure-response associations in health risk analysis. *Crit Rev Toxicol* 43(S1):26–38
- Cuny C, Nathaus R, Layer F, Strommenger B, Altmann D, Witte W (2009) Nasal colonization of humans with methicillin-resistant *Staphylococcus aureus* (MRSA) CC398 with and without exposure to pigs. *PLoS One* 4(8):e6800
- Davies P (2009) Methicillin Resistant *Staphylococcus aureus* in pigs, pork products and swine veterinarians. National Pork Board - NPB Final Research Grant Report, #NPB 07-196
- Davies P (2010) Prevalence and characterization of Methicillin-resistant *Staphylococcus aureus* (MRSA) in pigs and farm workers on conventional and antibiotic-free swine farms in the USA. #08-178
- de Boer E, Zwartkruis-Nahuis JTM, Wit B, Huijsdens XW, de Neeling AJ, Bosch T, van Oosterom RAA, Vila A, Heuvelink AE (2009) Prevalence of methicillin-resistant *Staphylococcus aureus* in meat. *Int J Food Microbiol* 134(1–2):52–56
- de Jonge R, Verdier JE, Havelaar AH (2010) Prevalence of meticillin-resistant *Staphylococcus aureus* amongst professional meat handlers in the Netherlands, March-July 2008. *Euro Surveill* 15(46)
- Declercq P, Petre D, Gordts B, Voss A (2008) Complicated community-acquired soft tissue infection by MRSA from porcine origin. *Infection* 36(6):590–592
- Denis O, Suetens C, Hallin M, Catry B, Ramboer I, Dispas M, Willemans G, Gordts B, Butaye P, Struelens MJ (2009) Methicillin-resistant *Staphylococcus aureus* ST398 in swine farm personnel, Belgium. *Emerg Infect Dis* 15(7):1098–1101
- EFSA (2009) Scientific Opinion of the Panel on Biological Hazards on a request from the European Commission on Assessment of the Public Health significance of meticillin resistant *Staphylococcus aureus* (MRSA) in animals and foods. *EFSA J* 993(1):1–73
- Feingold BJ, Silbergeld EK, Curriero FC, van Cleef BA, Heck ME, Kluytmans JA (2012) Livestock density as risk factor for livestock-associated methicillin-resistant *Staphylococcus aureus*, the Netherlands. *Emerg Infect Dis* 18(11):1841–1849
- Frana TS, Beahm AR, Hanson BM, Kinyon JM, Layman LL, Karriker LA, Ramirez A, Smith TC (2013) Isolation and characterization of methicillin-resistant *Staphylococcus aureus* from pork farms and visiting veterinary students. *PLoS One* 8(1):e53738
- Gilbert MJ, Bos MEH, Duim B, Urlings BAP, Heres L, Wagenaar JA, Heederik DJJ (2012) Livestock-associated MRSA ST398 carriage in pig slaughterhouse workers related to quantitative environmental exposure. *Occup Environ Med* 69(7):472–478
- Gorwitz RJ, Kruszon-Moran D, McAllister SK, McQuillan G, McDougal LK, Fosheim GE, Jensen BJ, Killgore G, Tenover FC, Kuehnert MJ (2008) Changes in the prevalence of nasal colonization with *Staphylococcus aureus* in the United States, 2001–2004. *J Infect Dis* 197 (9):1226–1234
- Graham PL 3rd, Lin SX, Larson EL (2006) A U.S. population-based survey of *Staphylococcus aureus* colonization. *Ann Intern Med* 144(5):318–325
- Graveland H, Wagenaar JA, Bergs K, Heesterbeek H, Heederik D (2011) Persistence of livestock associated MRSA CC398 in humans is dependent on intensity of animal contact. *PLoS One* 6 (2):e16830
- Guardabassi L, O'Donoghue M, Moodley A, Ho J, Boost M (2009) Novel lineage of methicillin-resistant *Staphylococcus aureus*, Hong Kong. *Emerg Infect Dis* 15(12):1998–2000
- IARTF (2011) Report of the Iowa antibiotic resistance task force, a public health guide. <http://publications.iowa.gov/17707/1/antibioticreport.pdf>. Last accessed on 2-9-18
- Kelman A, Soong YA, Dupuy N, Shafer D, Richbourg W, Johnson K, Brown T, Kestler E, Li Y, Zheng J, McDermott P, Meng J (2011) Antimicrobial susceptibility of *Staphylococcus aureus* from retail ground meats. *J Food Prot* 74(10):1625–1629

- Khanna T, Friendship R, Dewey C, Weese JS (2008) Methicillin resistant *Staphylococcus aureus* colonization in pigs and pig farmers. *Vet Microbiol* 128(3–4):298–303
- Kluytmans JA JW (2010) Methicillin-resistant *Staphylococcus aureus* in food products: cause for concern or case for complacency? *Clin Microbiol Infect* 16(1):11–15
- Limbago B (2010) Methicillin-resistant *Staphylococcus aureus* in the United States—is there a connection between retail foods and human infection? In: 2010 scientific meeting of the national antimicrobial resistance monitoring system, US-FDA, Atlanta, GA
- Lozano C, Aspiroz C, Ezpeleta AI, Gomez-Sanz E, Zarazaga M, Torres C (2011) Empyema caused by MRSA ST398 with atypical resistance profile, Spain [letter]. *Emerg Infect Dis* 17 (1):138–140
- Molla B, Byrne M, Abley M, Mathews J, Jackson C, Fedorka-Cray PJ, Sreevatsan S, Wang P, Gebreyes W (2012) Epidemiology and genotypic characteristics of methicillin-resistant *Staphylococcus aureus* strains of porcine origin. *J Clin Microbiol* 50(11):3687–3693
- NPB (2010) Pork quick facts—the pork industry at a glance. Pork Checkoff
- O'Brien AM, Hanson BM, Farina SA, Wu JY, Simmering JE, Wardyn SE, Forshey BM, Kulick ME, Wallinga DB, Smith TC (2012) MRSA in conventional and alternative retail pork products. *PLoS One* 7(1):e30092
- O'Donoghue M, Boost M (2004) The prevalence and source of methicillin-resistant *Staphylococcus aureus* (MRSA) in the community in Hong Kong. *Epidemiol Infect* 132(6):1091–1097
- Otto D, Orazem P, Huffman W (1998) Community and economic impacts of the Iowa Hog industry. In: Miranowski J (ed) Iowa's Pork industry—dollars and scents. ISU-CAIS, Iowa City
- Smith TC, Male MJ, Harper AL, Moritz-Korolev ED, Diekema D, Herwaldt LA (2008) Isolation of methicillin-resistant *Staphylococcus aureus* (MRSA) from swine in the midwestern United States. In: International conference on emerging infectious diseases, Atlanta, GA
- Smith TC, Male MJ, Harper AL, Kroeger JS, Tinkler GP, Moritz ED, Capuano AW, Herwaldt LA, Diekema DJ (2009) Methicillin-resistant *Staphylococcus aureus* (MRSA) strain ST398 is present in midwestern U.S. swine and swine workers. *PLoS One* 4(1):e4258
- Smith TC, Gebreyes WA, Abley MJ, Harper AL, Forshey BM, Male MJ, Martin HW, Molla BZ, Sreevatsan S, Thakur S, Thiruvengadam M, Davies PR (2013) Methicillin-resistant *Staphylococcus aureus* in pigs and farm workers on conventional and antibiotic-free swine farms in the USA. *PLoS One* 8(5):e63704
- USBLS (2012) Employment by Detailed Occupation—2010 and Projected 2020. Employment by Occupation Retrieved 2/27/2013, 2013. http://www.bls.gov/emp/ep_table_102.htm
- USCB (2012) Statistical Abstract of the United States: 2012, Table 1377. Meat Consumption by Type and Country: 2009 and 2010, U.S. Census Bureau
- USDA (2008) Swine 2006 - Part IV: changes in the U.S. Pork Industry, 1990-2006. USDA - #N520.1108. http://www.aphis.usda.gov/animal_health/nahms/swine/downloads/swine2006/Swine2006_dr_PartIV.pdf. Last accessed on 12/16/2009
- USDA-NASS (2009) 2007 Census of Agriculture, USDA National Agricultural Statistical Services
- van Cleef B, Haenen A, van den Broek M, Huijsdens XW, Mulders M, Kluytmans J (2009) Acquisition and persistence of methicillin-resistant *Staphylococcus aureus* Clonal Complex 398 during occupational exposure. In: 19th European congress of clinical microbiology and infectious diseases, Helsinki, Finland
- Van Cleef BA, Broens EM, Voss A, Huijsdens XW, Zuchner L, Van Benthem BH, Kluytmans JA, Mulders MN, Van De Giessen AW (2010) High prevalence of nasal MRSA carriage in slaughterhouse workers in contact with live pigs in The Netherlands. *Epidemiol Infect* 138 (5):756–763
- Van De Griend P, Herwaldt LA, Alvis B, DeMartino M, Heilmann K, Doern G, Winokur P, Vonstein DD, Diekema D (2009) Community-associated methicillin-resistant *Staphylococcus aureus*, Iowa, USA. *Emerg Infect Dis* 15(10):1582–1589
- van Rijen MM, Van Keulen PH, Kluytmans JA (2008) Increase in a Dutch hospital of methicillin-resistant *Staphylococcus aureus* related to animal farming. *Clin Infect Dis* 46(2):261–263

- Wassenberg MW, Bootsma MC, Troelstra A, Kluytmans JA, Bonten MJ (2011) Transmissibility of livestock-associated methicillin-resistant *Staphylococcus aureus* (ST398) in Dutch hospitals. *Clin Microbiol Infect* 17(2):316–319
- Waters AE, Contente-Cuomo T, Buchhagen J, Liu CM, Watson L, Pearce K, Foster JT, Bowers J, Driebe EM, Engelthaler DM, Keim PS, Price LB (2011) Multidrug-resistant *Staphylococcus aureus* in US meat and poultry. *Clin Infect Dis* 52(10):1227–1230
- Webb GF, Horn MA, D'Agata EM, Moellering RC, Ruan S (2009) Competition of hospital-acquired and community-acquired methicillin-resistant *Staphylococcus aureus* strains in hospitals. *J Biol Dyn* 48:271–284
- Weese JS, Gow SP, Friendship R, Booker C, Reid-Smith R (2009) Methicillin-resistant *Staphylococcus aureus* (MRSA) surveillance in slaughter-age pigs and feedlot cattle. In: ASM-ESCMID conference on MRSA in animals: veterinary and public health implications, London
- Weese JS, Rousseau J, Deckert A, Gow S, Reid-Smith R (2011) Clostridium difficile and methicillin-resistant *Staphylococcus aureus* shedding by slaughter-age pigs. *BMC Vet Res* 7 (1):41
- Wulf M, Markestein A, van der Linden F, Voss A, Klaassen C, Verduin C (2007) First outbreak of methicillin-resistant *Staphylococcus aureus* ST398 in a Dutch hospital, June 2007. *Euro Surveill* 13(9):8051

Part III

Predictive and Causal Analytics

Chapter 7

Attributive Causal Modeling: Quantifying Human Health Risks Caused by Toxoplasmosis from Open System Production of Swine



Introduction

This is the first of two chapters that apply predictive analytics to two very different risk prediction problems. As in the previous two chapters, the challenge in this one is to estimate human health risks from a pathogen in swine using a combination of plausible conservative estimates of relevant risk factors and probabilistic simulation. However, our focus now shifts to predicting how risks would *change* if some fraction of swine were shifted from totally confined production systems to more humane open systems. Predicting how interventions change risk requires a causal model, as discussed in Chap. 1. As in Chaps. 5 and 6, a simple product-of-factors framework is again suitable (see Eq. 7.5). Instead of the terms describing propagation of changes along successive links in a causal chain, with the change in the quantity at each step being equal to a sensitivity or slope factor times the change in its predecessor, many of the factors in this chapter are estimated *attribution fractions*. These describe the fraction of relevant deaths or illnesses per year in the population due to (i.e., attributed to) and caused by infection with a foodborne pathogen; the fraction of them that are attributed specifically to pork consumption, and so forth. Unlike the attributable risk estimates or attributable fractions criticized in Chap. 2, which were derived purely from statistical associations, in this application the causal agent of disease, *T. Gondii*, is known and can be measured. Predictions for effects of interventions are therefore grounded in causal attribution calculations that can be compared to available data on prevalence and infectivity of the relevant causal agent. Chapter 8 will then turn to a pure prediction problem: how well the entries in one column in a table (indicating *in vivo* carcinogenicity of chemicals, or lack of it, in rodents) can be predicted from entries in other columns, representing results of relatively inexpensive high-throughput screening (HTS) assays. No causal model is required for this task: predictive analytics algorithms alone suffice.

For readers who wish to skip ahead, the main points of this chapter are as follows. Open livestock production systems, including free-range and organic livestock systems, seek to improve the welfare of animals by letting them roam in unconfined spaces. This increases their exposure to potentially harmful micro-organisms, including *T. gondii*. When transmitted through the food chain, *T. gondii* threatens human health, especially in unborn children of women infected during pregnancy, as well as the lives of patients with compromised immune systems. By contrast, conventional total confinement production systems can now keep this human health risk at or near zero. The probabilistic risk simulation model developed in the rest of this chapter quantifies the trade-off between greater use of open swine production systems and increased cases of toxoplasmosis in humans. It predicts that every 1804 pigs shifted from conventional total confinement to open production (95% confidence interval 747–9520) would cause the loss of one additional human quality-adjusted life year (QALY), and that increasing the fraction of U.S. swine raised in open/free range operations by 0.1% (approx. 65,000 pigs) would cause a loss of approximately 36 human QALYs per year, including between 1 and 2 extra adult deaths per year. Methods of causal analytics are valuable largely because they can quantify such tradeoffs and answer what-if questions about how human health risks would change for different interventions, such as if different fraction of pigs were shifted to open production systems. This tells risk managers and policy-makers what they need to know to make decisions that are well-informed about trade-offs and about the probable consequences of different choices.

Background on Toxoplasmosis

The causative agent of Toxoplasma infection, *Toxoplasma gondii*, is a common parasite in most warm-blooded animals, including people, worldwide. It is believed to be transmitted primarily via eating contaminated raw and undercooked meat (especially pork and poultry in the U.S.), as well as through contact with contaminated cat feces. More recently, drinking water and other potential environmental sources have received increased attention (Tenter et al. 2000). However, since as many as 2.8 million *Toxoplasma*-infected pigs enter the U.S. food chain each year (Hill and Dubey 2013), pork still must be considered a major risk factor. Approximately 9% of the US population (ages 12–49) are infected with the parasite, and much higher prevalence can be found in other parts of the world (Jones et al. 2007). Although risk of parasite-mediated diseases such as toxoplasmosis and trichinosis can be kept at or near zero in modern intensive swine production systems, open systems are less controllable, and increased risk of toxoplasmosis is part of the price of these alternative production systems (Davies 2011).

Most infected adult humans suffer few or no detectable ill effects from toxoplasmosis. However, infection can be deadly for immunocompromised patients and devastating for the unborn children of pregnant women. In the mid 1990s, it was estimated that about 1 in 10 AIDS patients in the US died from toxoplasmosis

(Hays 1996). For a woman infected during pregnancy, there is a 20–50% chance of the baby being born with the infection and having increased risk of blindness, mental retardation, or other medical problems. In 1994, it was estimated that about 0.8 babies in 10,000 in the US were born with the infection (Guerina et al. 1994). Of all creatures infected with *Toxoplasma gondii*, only cats shed *T. gondii* oocysts, which can withstand the external environment and spread the disease (Dubey 2013). If infected cat feces contaminate the feed of pigs, humans can subsequently become infected by eating or handling raw or undercooked pork containing the parasite.

Enhanced bio-security during pork production greatly reduces the incidence of infectious diseases. There has been a steady increase in use of bio-security measures to protect human and animal health, including disinfection, all-in/all-out livestock rotation, visitor control, rodent control, and housing control (using confined spaces that keep out all dogs, cats, birds, etc.) (USDA-APHIS 2008). Practices used in alternative pork production systems prescribe continuous outdoor access for pigs and preclude the use of confinement facilities. Although some bio-security measures are still possible, such as herd isolation, visitor control, and quick removal of diseased animals, there are also more opportunities for swine to come in contact with sources of *T. gondii* infection.

Data and Methods

To compare the *T. gondii*-related health outcomes for consumers of pork derived from pigs in total confinement systems versus pork derived from pigs in open/free range systems, we modeled the influential input parameters as probability distributions, and then linked these distributions in a probabilistic simulation model that relates health outcomes (more infections, birth defects, and deaths—expressed as Quality Adjusted Life Years) to inputs. Figure 7.1 outlines the main structure of the resulting simulation model. The fraction of pigs potentially shifted from total confinement to open/free range production systems is an exogenously specified input to the model. Varying this fraction and simulating the resulting changes allows the probable trade-off between greater animal welfare (higher fraction of pigs in open facilities) and greater human health risk to be estimated quantitatively. To account for uncertainty in the input parameters, each individual simulation run performs a random draw from all of the parameter distributions and generates the resulting linked outcomes. Performing a large number of independent simulation runs builds up distributions for the health outcomes of interest. This approach automates the process of sensitivity analysis by simultaneously combining virtually all possible combinations of inputs. The data needed to form each parameter distribution is drawn from the available scientific literature on *T. gondii* prevalence in pigs and pork, human infection rates, and human health effects of infections, as described in the sections below. These parameters are also summarized in Table 7.1.

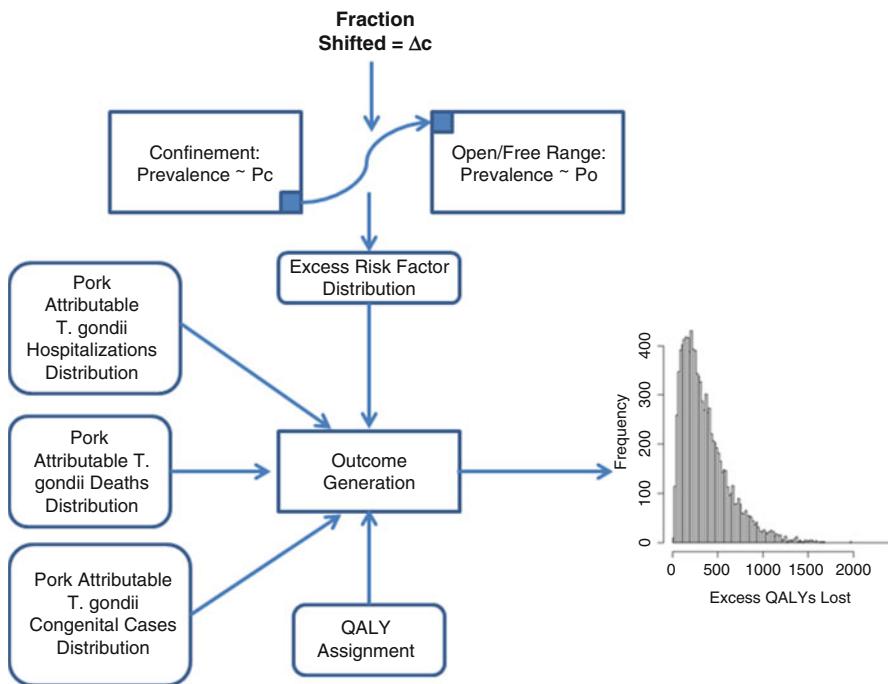


Fig. 7.1 Conceptual diagram of the Toxoplasmosis simulation model

Distributions for *T. gondii* Prevalence in Pigs

Literature and Data Review

In a nationwide survey performed in 1983–1984, 23% of all market pigs tested positive for *T. gondii* antibodies, indicating past (or present) infection (Dubey et al. 1991). A decade later, a 1995 USDA study involving pigs from 17 major pork producing states (Patton et al. 1996) found that 3.2% of market pigs tested positive, while only 0.9% tested positive in a similar 2000 study (Patton et al. 2002). The 2006 version of the USDA study (Hill et al. 2010) found a mean within-herd prevalence of *T. gondii* antibodies of 2.7%. Thus, a sharp decline in prevalence of *T. gondii* antibody prevalence has coincided with increased use of modern farming practices that emphasize biosecurity, such as confinement rearing, rodent control, hygienic feed handling, and exclusion of cats (USDA-APHIS 2008; Davies 2011).

The prevalence values above reflect nationwide averages over a variety of pork production practices and housing types. According to the USDA, in 2006, 73.6% of hog production sites in the U.S. used either total confinement (53.2%) or open buildings with no outside access (20.4%) (USDA-APHIS 2008). The rest were classified as “open building with outside access” (23.3%), lot with hut or no building (1.8%), or pasture with no hut or no building (1.3%). Not surprisingly, studies show

Table 7.1 Distribution and parameter summary

Model input	Data source(s)	Distribution	Parameters
Prevalence in total confinement swine	Published surveys (see text)	Beta	0, 0.003, 0.027 (Min, Mean, Max)
Prevalence in open/free range swine	Published surveys (see text)	Beta	0.027, 0.244, 0.901 (Min, Mean, Max)
Proportion foodborne	Covers known estimates (see text) and assuming high uncertainty	Uniform	0.30, 0.50, 0.70 (Min, Mean, Max)
Proportion from pork	Batz et al. (2012)	Normal	(0.41, 0.059) (Mean, Std. Dev)
Human baseline prevalence	Estimated prevalence for persons aged 40–49 years (Jones et al. 2007) was assumed to be the cumulative result of 45 years of constant incidence. Upper/lower limits correspond to published 95% CI	Constant	0.1137, 0.157, 0.177 (Min, Mode, Max)
Incidence	$1 - (1 - \text{Prevalence})^{(1/45)}$ applied to non-infant fraction (0.9872 per 2010 composition) of 2013 population (approx. 315 M) = 310.98 M	Degenerate	0.00268, 0.00379, 0.00432 (Min, Mode, Max)
Seroconversion rate	The symptomatic fraction was estimated at 15%. Uncertainty was based on a 50% relative increase/decrease from 0.15 on a log odds scale	PERT-Beta	0.11, 0.15, 0.21 (Min, Mode, Max)
Proportion hospitalized	2000–2006 Nationwide Inpatient Sample (NIS) using ICD—CM code 130 (Toxoplasmosis) http://www.hcup-us.ahrq.gov/nisoverview.jsp	PERT-Beta	0.017, 0.026, 0.033 (Min, Mode, Max)
Proportion who died	2000–2006 Nationwide Inpatient Sample (NIS) using ICD—CM code 130 (Toxoplasmosis)	PERT-Beta	0.0014, 0.0019, 0.0022 (Min, Mode, Max)
Under-diagnosis multiplier (deaths, hospitalizations)	Derived in (Scallan et al. 2011)	PERT-Beta	1, 2, 3 (Min, Mode, Max)
Congenital rate	Cases per 10,000 live births (3,942,000 annually)	PERT-Beta	0.2, 0.5, 0.8 (Min, Mode, Max)
QALY conversion rates	Hoffmann et al. (2012)	Factors by case type	See Table 7.3

Estimates based on Scallan et al. (2011) are shown in the center delineated block of rows. Modifications to their values are shaded grey

that the prevalence of *T. gondii* in pigs is much lower for pigs isolated from potential infection sources. In a 1995 study, samples from pigs in total confinement production systems in North Carolina were found to be positive for *T. gondii* in only 1 out of 1752 (0.06%) cases (Davies et al. 1998). A 2006 USDA swine survey found that, among pigs in total confinement facilities, 0.3% tested positive, while in other facilities (open buildings with or without outside access, hut, lot, or pasture), the rate was more than twenty-fold greater, at 6.5% (USDA-APHIS-VS-CEAH 2011). A 2008 study of over 74,000 market pigs in the United States detected *T. gondii* in 0.8% of samples overall, ranging from 0.5% in the largest systems to 2.6% in the smallest (less than 1000) (McKean et al. 2009). Larger systems are more likely to be confinement production systems. At present, it appears that pigs raised in total confinement can be *T. gondii* free, provided that there is adequate rodent control and prevention of contamination of food and water with oocysts (Dubey 2009).

Pigs that are raised using so called natural, open, organic, or free range production practices have, by definition, greater exposure to the open environment and potential sources of *T. gondii*. Gebreyes et al. (2008) compared serum samples taken from swine reared in conventional intensive indoor production systems and outdoor antimicrobial free (ABF) production systems. The study analyzed 616 samples from three different states in the US: Wisconsin, Ohio, and North Carolina. The percentage of samples testing positive for *T. gondii* was 1.1% (3/292) in swine from conventional systems versus 6.8% (22/324) in swine from ABF systems. Pigs raised on two organic farms in Michigan, (Dubey et al. 2012) were found to have antibodies to *T. gondii* in 90.1% (30/33) of cases. Dubey et al. (2002) found a *T. gondii* seroprevalence rate of 76% (19/25) in randomly selected free-range pigs on a farm in Massachusetts previously known to have *T. gondii* infections, and from 70.1% (34/48) of free range pigs on a poorly run farm in Maryland (Dubey et al. 2008). It is also worth noting that a cross-sectional study involving 3247 feral pigs in 32 states found an average *T. gondii* seroprevalence of 17.7% (Hill et al. 2014). Feral pigs could be considered the ultimate in free-range swine, however it is plausible that they would not have as much contact with *T. gondii* carriers such as rodents and cats as would open production farm pigs.

Research in the Netherlands also shows wide variation in swine seroprevalence, depending on the housing type. Kijlstra et al. (2004) compared *T. gondii* prevalence among swine from free range, organic, and intensive/conventional. The rates were 4.7% (30/635), 1.2% (8/660), and 0% (0/621) respectively. A few years later, van der Giessen et al. (2007) found rates of 5.62% (10/178) for free range, 2.74% (11/402) for organic, and 0.38% (1/265) for intensive/conventional.

Derived Distributions for Average *T. gondii* Prevalence in Pigs

Using relevant data from the studies cited above, we estimated the average *T. gondii* prevalence in two contrasting categories of pigs: (1) a randomly sampled subset of pigs raised in total confinement production systems, versus (2) a randomly sampled subset, of identical size, of pigs raised in open systems with outside access such as

Table 7.2 Beta distribution parameters for average *T. gondii* antibody prevalence in pigs

Facility type	Mean prevalence	Min (assumed)	Max (assumed)	Variance (computed)	Alpha (computed)	Beta (computed)
Total confinement	0.003 (USDA-APHIS-VS-CEAH 2011)	0	0.027 (Hill et al. 2010; see text)	2.025E-5	0.2840	2.2716
Open/free range	0.244 combined from multiple sources	0.027 (Hill et al. 2010; see text)	0.901 (Dubey et al. 2012; see text)	0.0212	1.4199	4.2991

those used in organic, natural, or free-range production. As noted above, various midway housing solutions exist, such as open buildings with no outside access, however, these are not directly relevant to our framework. To account for high degrees of estimation uncertainty, we developed beta probability distributions describing the average prevalence within each of these two groups of pigs. Beta distributions can be used to estimate continuous values that lie between 0 and 1, such as proportions and percentages. They are often used for developing distributions from limited empirical data since a minimum, maximum, and mean value are sufficient to characterize the distribution. This is especially relevant in a meta-analysis, where data and results from different studies cannot always be explicitly combined, and expert judgment may be used to inform some parameters. When appropriately parameterized, they also provide the useful property of a long “right hand tail”, which is suitable for situations where the main weight of evidence is for relatively low values, yet higher possible values must be considered, but at lower weight. The beta distribution parameters, *alpha* and *beta*, were computed from estimated values for the minimum, maximum, mean, and variance of average prevalence by using the equations described in Davis (2008). The variance was assumed to follow the standard (beta) PERT form, $\sigma^2 = \frac{(\max-\min)^2}{36}$. The minimum, maximum, and mean value parameters were estimated as described below and are summarized in Table 7.2.

Distribution for Closed/Total Confinement Systems

A plausible *upper bound* on the total confinement subset average prevalence is the value of 2.7% reported in the 2006 USDA survey as the average over *all* facility types. We reason as follows: since this average mixes in values from pigs in a variety of open facility types, where average prevalence is known to be higher, it is very likely to be an upper bound on the average for total confinement facilities alone. For a *lower bound* on the average, we use zero, reflecting the fact that zero or near zero *T. gondii* prevalence has been reported in a number of cases (Davies 2011; Kijlstra

et al. 2004; van der Giessen et al. 2007), together with the fact that elimination of this risk appears to be realistic for current total confinement systems (Davies 2011; Dubey 2009). For the *mean* of the average total confinement subset prevalence, we used the average value for total confinement facilities of 0.3% reported in the 2006 USDA survey. We did not combine this with the value of 0.06% (1/1752) reported in Davies (2011) since that study was based on swine in North Carolina alone. The McKean et al. (2009) study also cannot be used explicitly since it differentiated by herd size rather than confinement type (although the two are correlated). However, the size, geographical diversity, and relative recency of the USDA study support the validity of these estimates.

Distribution for Open/Free Range Systems

In the case of open/free range facility types, a plausible *lower bound* on the average subset prevalence is the same 2.7% average 2006 survey value used as the upper bound for total confinement facilities. Using reasoning similar to above, since the overall average is derived from a sample of which over half were pigs in total confinement facilities, the 2.7% average would likely form a lower bound on the average prevalence of pigs drawn from open/free range situations alone. To estimate the *mean* of the average for open/free range systems, we combined sample data from the available U.S. studies described above that involved outdoor ABF (Gebreyes et al. 2008), free range (Dubey et al. 2002; Dubey et al. 2008), or organic production (Dubey et al. 2012). The combined data from these studies provides an estimated mean seroprevalence rate of $(22 + 30 + 19 + 34)/(324 + 33 + 25 + 48) = 105/430 = 0.244$. This estimate seems quite plausible since it is very close to the nationwide average prevalence reported for 1983–1984 of 23% (Dubey et al. 1991), which is before the major shift to production methods that emphasize confinement and biosecurity. As an upper bound for the average subset prevalence for open/free range facilities, we used the highest observed fraction of 90.1%.

Figure 7.1 shows the beta distributions implied by the base case parameter values in Table 7.2. These represent the frequency distributions of average *T. gondii* antibody prevalence rates for subsets of pigs in each category, total confinement and open/free range, respectively.

Excess Risk Factor Distribution

Our model assumes that *T. gondii* cases in humans are proportional to the overall prevalence of *T. gondii* in pigs. In this section we derive a factor to quantify the excess risk associated with shifting a given fraction of pigs from total confinement to open/free range operations, thereby boosting the overall average prevalence level. We will later use the excess risk factor derived below as a multiplier in determining the increase in adverse health outcomes.

Let ΔC represent pigs that are shifted from total confinement operations to open/free range operations expressed as a fraction of all pigs. Then the increase in overall net prevalence of *T. gondii* in pigs, ΔP , is distributed as:

$$\Delta P \sim \Delta C^* (P_O - P_c) \quad (7.1)$$

where P_C and P_O are random variables drawn from the beta distributions for confined and open/free range operations respectively. Let:

$$P_b = \text{baseline fraction of pigs with } T. gondii \text{ antibodies} = 0.027.$$

Then the relative risk factor, RR, associated with an increase, ΔC , in the fraction of pigs not in total confinement is distributed as:

$$\text{Relative risk} = RR \sim (P_b + \Delta P) / P_b$$

And the excess (incremental) risk as a function of ΔC , denoted as $r(\Delta C)$, is $RR - 1$:

$$r(\Delta C) \sim (P_b + \Delta P) / P_b - 1 = \Delta C^* (P_O - P_c) / P_b \quad (7.2)$$

Attribution to Pork

The fraction of human toxoplasmosis cases attributed to pork consumption is the product of the fraction attributed to all food as a source, and the fraction of all food cases attributed to pork. Food sources have been estimated to cause half of all cases in the U.S. (Scallan et al. 2011). The article notes that this estimate was based both on earlier estimates (Mead et al. 1999) and on a European six-city survey of pregnant women, which used survey data and logistic regression to estimate a food-attributable fraction between 0.30 and 0.63 (Cook et al. 2000). Other mean estimates appearing in the literature include 0.32 (Cressey and Lake 2005), 0.50 (Vaillant et al. 2005), and 0.56 (Havelaar et al. 2008), based on data and perceptions in New Zealand, France, and the Netherlands respectively. An assumption that the overall food attribution fraction varies uniformly over the interval 0.30–0.70 with a mean of 0.50 seems consistent with the data above (it covers the range of previous estimates), is symmetric, and better accounts for a high degree of uncertainty than would a peaked distribution.

In the available literature, fractions of cases attributable to specific foods within the foodborne category have been based on expert judgment. Structured expert elicitation have yielded a mean attribution for pork of 0.50, with a fifth and 95th percentile of 0.21 and 0.99 respectively (Havelaar et al. 2008); and a mean attribution of 0.41 for pork in relation to other possible U.S. food sources, with a standard error of 0.059 (Batz et al. 2012). Both studies relied upon panels of scientific experts,

a panel of 16 in Havelaar et al., (11 microbiologists, 4 epidemiologists, and 1 food safety scientist) and 45 “nationally recognized food safety science experts” in Batz et al. Both the foodborne and pork-specific attribution values exhibit significant uncertainty and variability. Actual values will vary depending on food consumption patterns, climate, cultural practices, and other geographic specific dynamics. For the selection of pork within the food category we account for variability by using the mean and standard deviation parameters in the Batz et al. data to define a Normal (0.41, 0.059) probability distribution. Note that the resulting 95% confidence interval [0.29, 0.53] contains the Havelaar et al. mean estimate without introducing the statistical anomalies implied by its corresponding wide and asymmetric estimation interval. The food and pork attribution distributions are summarized in the top rows of Table 7.1.

Pork Attributable Human Case Rates: Adult Hospitalizations and Death

Estimates of human Toxoplasmosis case rates relied largely upon a single large CDC sponsored study of foodborne illness in the U.S (Scallan et al. 2011). That paper derived estimates from large national medical database systems, published research, and expert judgment. As with our study, uncertainty was captured by simulating probability distributions characterizing each component. A similar study, also using simulation modeling, was performed more recently to estimate foodborne illness rates in Canada (Thomas et al. 2013). The Canadian model utilized many estimates from Scallan et al., most notably, for toxoplasmosis, the foodborne attribution rate of 0.50. Similar to our approach, a separate model was used to estimate congenital cases. Given that the Canadian model used many of the same parameters, but was applied to a different population from that in Scallan et al., we concluded that it would have limited applicability to our situation. We used a slightly modified version of the Scallan et al. model as a submodel to ours to generate distributions of human health outcomes for a baseline prevalence level of human Toxoplasmosis. We did not do a deep re-evaluation of each individual distribution and parameter set in their model, nor did we try to combine their original parameter estimates with others. We felt that due to the extensive data, the expertise of the authors, the reasonableness of their assumptions, and the uniqueness of many of their results, their existing model would be quite suitable for the purposes of our study. An exception was the probability distribution for the proportion of cases that were foodborne, as discussed above. Also, since we are only interested in the pork attributable fraction of foodborne cases, the related distribution, also discussed above, was brought into the model. We used a slightly different form of the PERT-Beta distribution due to known mathematical shortcomings with theirs (Davis 2008). We applied the Scallan et al. study parameters only to the *non-infant* population of the U.S. in 2013 (estimated to be 310.98 M). A separate analysis,

described later in this chapter, was used to derive congenital cases. The distributions and parameters derived from their study are summarized in the center portion of Table 7.1.

Following the model of Scallan et al., with modifications as discussed above, the total pork attributable case rate distribution can be stated as the following product:

$$\begin{aligned} \text{Total Pork Attributable Cases} = & \text{ Population} \times \text{Proportion Foodborne} \\ & \times \text{Proportion from Pork} \times \text{Incidence} \\ & \times \text{Seroconversion Rate} \end{aligned} \quad (7.3)$$

Distributions for hospitalizations and deaths were then computed as:

$$\begin{aligned} \text{Hospitalizations} = & \text{ Total Pork Attributable Cases} \\ & \times \text{Proportion hospitalized} \times \text{Under-diagnosis factor} \end{aligned} \quad (7.4)$$

$$\begin{aligned} \text{Deaths} = & \text{ Total Pork Attributable Cases} \times \text{Proportion who died} \\ & \times \text{Under-diagnosis factor} \end{aligned} \quad (7.5)$$

Using the probability model described by Eqs. (7.3) through (7.5) (but without pork attribution), Scallan et al. estimated 86,686 (90% credible interval 64,861–111,912) domestically acquired *foodborne* cases of toxoplasmosis in the US, based on the 2006 population. The study also estimated 4428 hospitalizations (90% range 2634–6674) and 327 deaths (90% range 200–482) attributable to foodborne infections. Multiplying 310.98 M (estimated US non-infant population in Jan 2013) by the mean or mode of the appropriate distributions yields approximately 36,242 total cases, 1885 hospitalizations, and 138 deaths attributable to *T. gondii* in pork. These are just slightly above the mean pork attributable outcomes given in Batz et al. (2012), Table 6, who also relied upon the model of Scallan et al. More detailed estimates are presented and discussed in the [Results](#) section.

Congenital Toxoplasmosis

Congenital cases (where infection is transferred vertically from the pregnant mother to the fetus) were estimated from historical case rates and current populations. The provisional count of births in the United States for the 12-month period ending June 2012 was 3,942,000 (Hamilton and Sutton 2012). Some fraction of these infants was born with congenital toxoplasmosis. However, the only available survey data regarding the prevalence of congenital toxoplasmosis in the U.S. is largely outdated. Two prospective studies in the 1970s both reported rates of congenital toxoplasmosis of approximately 10 per 10,000 live births. More recent data regarding the rate of congenital toxoplasmosis are available from the New England Regional Newborn Screening Program (Guerina et al. 1994). All infants born in the catchment area of this program are tested for evidence of congenital toxoplasmosis; infected infants undergo clinical evaluation and treatment for 1 year. Of 635,000 infants who

underwent serologic testing in 1986–1992, 52 were infected, representing an infection rate of approximately 0.8 per 10,000 live births.

Jones et al. (2007) found that the *T. gondii* prevalence in U.S. females ages 12–49 declined from 13.4% [95% CI—11.6, 15.1] during 1988–1994 (the approximate time frame of the Guerina et al. study) to 8.2% [95% CI—6.6, 9.8] during 1999–2004, or approximately 38.8%. It has also been shown that prevalence tends to be highest in the Northeast U.S. (the area of the Guerina et al. study). For the period 1988–1994, Jones et al. (2001) estimated a prevalence of 29.2% for the Northeast, 22.8% for the South, 20.5% for the Midwest, and 17.5% for the West. Given these data, and assuming that the rate of congenital toxoplasmosis is approximately proportional to the prevalence of *T. gondii* in adult women, it seems plausible that 0.8 per 10,000 live births from the 1994 Guerina et al. study is an upper bound on the 2012 nationwide average rate. Using the observed prevalence reduction, we will assume that the mean prevalence is now $0.8 \times (1 - 0.388) \approx 0.5$ per 10,000 births. Correspondingly, the low estimate (accounting for regional differences and potential further reductions) is assumed to be 0.2 per 10,000 live births. Using a total number of births of 3,942,000 (Hamilton and Sutton 2012), the mean estimate for congenital toxoplasmosis in 2012 is then 197 cases with a range of 79–315. The mean pork-attributable count is 40.4. More precise estimates are presented and discussed in the [Results](#) section.

$$\begin{aligned} \text{Congenital Cases} = & (3,942,000 / 10,000) \times \text{Congenital Rate} \\ & \times \text{Proportion Foodborne} \times \text{Proportion from Pork} \quad (7.6) \end{aligned}$$

QALYs Lost Assignment

Data provided by Hoffmann et al. (2012) can be used to derive an average Quality Adjusted Life Year (QALY) loss per incident of illness, hospitalization, and death due to *T. gondii* for both adult and congenital cases. QALYs combine into one metric a measure of the relative impacts of health conditions on mortality, comfort, and the ability to engage in normal activities. A loss of one QALY is equivalent to the loss of one healthy human life-year, while a value of zero equates to zero loss of health. Fractional values can be used to quantify both shorter periods of time and intermediate levels of impairment. The metric has the desirable effect of placing greater weight on death or injury occurring earlier in life. QALYs were originally developed to assess alternative medical treatment options, and are increasingly used in policy and risk analysis. The averaged QALY values are shown in bold in Table 7.3. The total pork-attributable QALYs lost due to toxoplasmosis is distributed as:

$$\begin{aligned} \text{Total QALY loss} = & 1.3373e-4 \times (\text{Total Cases} - \text{Hospitalizations}) \\ & + 0.0565 \times \text{Hospitalizations} + 27.71 \times \text{Deaths} \\ & + 2.2086 \times \text{Congenital Cases} \quad (7.7) \end{aligned}$$

Analysis

This section assesses the excess risk associated with a shift in the proportion of swine from total confinement to open/free range production, with the size of the shifted fraction of swine ranging from 0 to 0.001 (0.1%). It is useful to compare the size of the shift to the current levels of pigs in the open/free range category. The number of USDA certified “organic” swine in 2006 was 7508, growing to 12,373 in 2011 (USDA-ERS 2012). The USDA organic seal verifies that producers met animal health and welfare standards, did not use antibiotics or growth hormones, used 100% organic feed, and provided animals with access to the outdoors. It does not include other labeling types such as “natural” or “free-range”, however, it does provide an idea of the magnitude versus conventionally raised swine. In contrast, for the same year, the USDA estimated the total U.S. farm hog population at 64,925,000 head (USDA-NASS 2012). Therefore, about 1 out of every 5200 pigs was raised organically in 2011, corresponding to a production shift, ΔC of approximately $1/5200 = 0.00019$. To measure the impact of a range of possible future values, we varied ΔC from 0 to 0.001 in ten increments of 0.0001 (approximately 6500 pigs each). Each successive value of ΔC was supplied as an input to a probabilistic simulation model developed in the R (version 2.15) statistical programming environment (<http://www.r-project.org/>), using the mc2d (Tools for Two-Dimensional Monte Carlo Simulations) add-in package for its four-parameter Beta distribution random number generator. We generated 10,000 random values for each underlying probability distribution to obtain 10,000 random values for each relative risk factor (Eq. 7.2) and each level of health outcome (Eqs. 7.3–7.7). Multiplying these numbers yielded the distribution of excess risk results for each outcome:

$$r_k \sim r(\Delta C) * H_k \quad (7.8)$$

Table 7.3 Average QALYs lost per case by outcome

Outcome	Base count	Total QALYs	QALYs/case
Adult illness (mild)	69,862	0	0
Adult illness (moderate)	12,396	11	8.8738e-4
Adult average illness	82,258	11	1.3373e-4
Adult hospitalization (only)	4168	64	0.0145
Adult chronic illness	260	186	0.7154
Adult average hospitalization	4428	250	0.0565
Adult death	327	9062	27.71
Congenital chronic illness	341	603	1.7683
Congenital death ^a	15	1038	69.2
Congenital case (mild)	387	0	0
Congenital average case	743	1641	2.2086

^aAdditional cases due to neonatal deaths, stillbirths, and miscarriages due to congenital infection

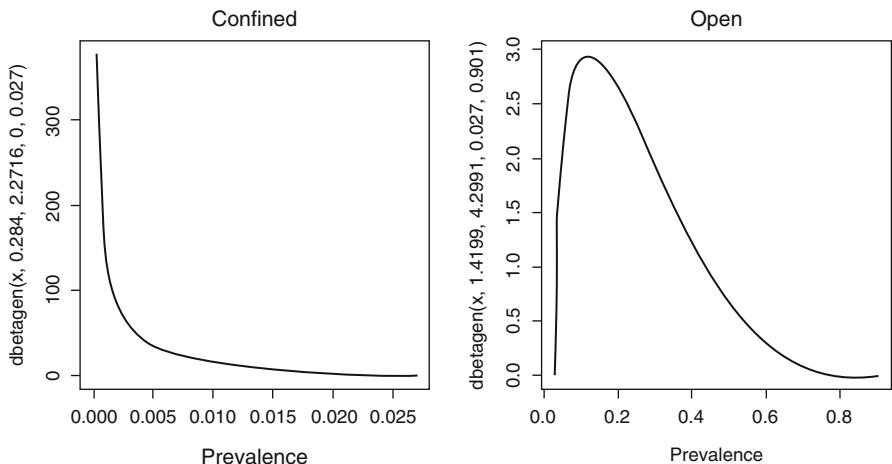


Fig. 7.2 Frequency distributions for *T. gondii*. Prevalence in confined versus open/free range systems

Table 7.4 Estimates of current pork-attributable human toxoplasmosis health outcomes per year

Outcome (equation)	Mean	5%	95%
Total cases (7.3)	37,027.08	21,128.75	56,698.22
Hospitalizations (7.4)	1904.72	962.80	3173.62
Deaths (7.5)	138.40	71.07	226.88
Congenital cases (7.6)	40.45	20.66	66.55
Total QALYs (7.7)	4036.72	2091.79	6587.93

is the excess risk of human toxoplasmosis associated with a production shift ΔC_{in} swine from total confinement to open/free range production, where H_k denotes the random variable measuring annual loss for health outcome measure k , for $k = \text{total cases, hospitalizations, deaths, or total QALYs}$. A conceptual diagram of the simulation model is depicted in Fig. 7.2 below.

Results

Health Outcome Distributions

Table 7.4 provides the simulated distributions for current levels of pork-attributable toxoplasmosis related health outcomes in the U.S. These estimates are very close to those obtained by Batz et al. (2012). Figure 7.3 is a histogram of the full distribution generated for current total QALYs lost due to toxoplasmosis.

Fig. 7.3 Distribution of current total QALYs lost per year due to pork attributable *T. gondii*

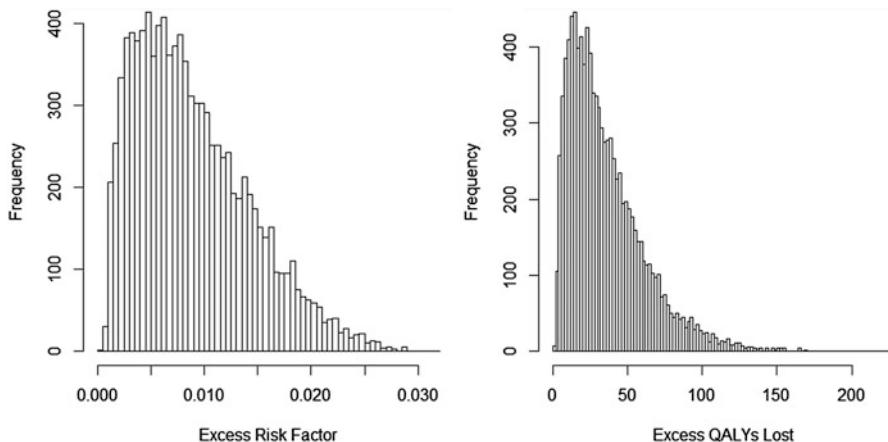
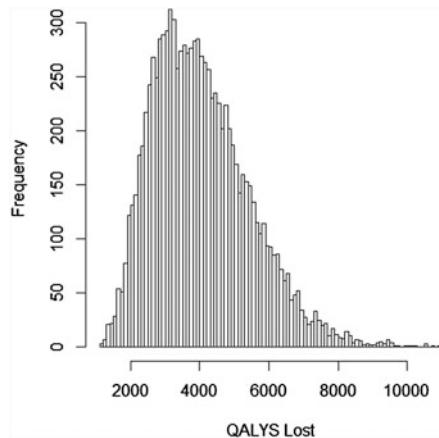


Fig. 7.4 Distributions of the excess risk factor (left panel) and QALYs lost per Year (right panel) at 0.001 production shift

Excess Risk Factor and QALYs Lost Due to Reduced Confinement

A shift of swine from total confinement operations by a fraction by $\Delta C = 0.001$ (approx. 65,000 pigs) increases the excess risk factor by an average of 0.0089 (95% CI = 0.002, 0.0192). Since the risk factor distribution (Eq. 7.2) is a composite of two beta distributions with long right-hand tails, it is also highly skewed as shown in Fig. 7.4 (left panel). The corresponding distribution of QALYs lost due to a shift of 0.001 of pigs from total confinement operations, shown in the right panel of Fig. 7.4 (and calculated via Eq. 7.8 with H_k = Eq. 7.7) has a mean of 35.98 quality-adjusted

Table 7.5 Excess QALYs lost (lower, mean, and upper estimates) per year versus production shift fraction

Shift fraction ΔC	Equivalent pigs	Mean QALYs lost/ year	5% Confidence level	95% Confidence level
0.0001	6493	3.60	0.68	8.69
0.0002	12,985	7.20	1.36	17.38
0.0003	19,478	10.79	2.04	26.08
0.0004	25,970	14.39	2.73	34.77
0.0005	32,463	17.99	3.41	43.46
0.0006	38,955	21.59	4.09	52.15
0.0007	45,448	25.18	4.77	60.84
0.0008	51,940	28.78	5.45	69.53
0.0009	58,433	32.38	6.13	78.23
0.0010	64,925	35.98	6.82	86.92

life-years (95% CI = 6.82, 86.92). This is equivalent to an expected loss of one human QALY for every 64,925/35.98 ≈ 1804 pigs moved from total confinement production. From the confidence limits for QALYs lost, as many as 9520 pigs, or as few as 747 pigs, shifted from total confinement operations, could cause the loss of approximately one human QALY. Table 7.5 shows the lower, mean, and upper estimates of the excess QALYs lost as the production shift fraction increases from 0 to 0.001.

Discussion and Conclusions

This chapter has used a simple causal model to estimate how human Toxoplasmosis rates would be affected by shifting swine production from total confinement systems to open/free range systems. It determines plausible expected values and ranges of health effects that reflect the inherent uncertainty in the input parameters. A probabilistic simulation model provided the mechanism to generate the desired results. The model predicts that every 1804 pigs shifted to open production (95% confidence interval 747–9520) cause an expected additional loss of a human quality-adjusted life year (QALY). As of 2011, the relatively small number of USDA certified organic swine (12,373) corresponded to an expected human health loss of about (12,373/1783) = 6.94 human QALYs lost per year. Shifting 0.1% of total swine (approx. 65,000 pigs) from total confinement to open/free range operations would cause a loss of approximately 36 human QALYs per year, including between 1 and 2 adult deaths per year (138.40 current annual deaths x 0.0089 avg. excess risk factor). The number of pigs certified as organic increased by 3286 pigs from 2010 to 2011 (USDA-ERS 2012), which is equivalent to an average estimated human QALY loss of 1.82.

The results above are sensitive to several of the estimated parameters. In particular, the pork attributable fraction (*proportion foodborne x proportion from pork*), with an estimated mean value of approximately 0.20 (0.41×0.50). Equations (7.3) and (7.8) imply that the mean excess risk values are proportional to the mean pork attributable fraction. Both of the subparameter distributions are the results of aggregating expert opinions, with obvious shortcomings relative to estimates derived from experimental evidence. However, by modeling the true fractions as relatively broad probability distributions, we hope to have captured this uncertainty in our output confidence intervals. Similarly, the mean excess risk values are proportional to the *difference* in prevalence rates between pigs raised in confined versus open/free range systems (see Eqs. 7.2 and 7.8). The prevalence values are based on aggregation of experimental results—better than expert opinion, but still subject to inaccuracies when applied to current-day, nationwide experiences. In this case, we also rely on having captured the uncertainty via probability distributions. But we also benefit from only needing to accurately estimate the difference in the two rates rather than their absolute values. Collecting additional data regarding the prevalence of *T. gondii* in open/free range swine could allow tighter estimates for the prevalence distribution (P_O in our study) and thus tighter estimates of the health impacts. Finally, the mean risk results we report are also proportional to the mean health outcome estimates of Scallan et al. (see Eqs. 7.3–7.5 and 7.8), which have their own limitations as discussed by the authors.

The scope of this analysis is limited to human health effects attributable to Toxoplasmosis from pork for two alternative pig production systems. We do not attempt to perform a risk or cost-benefit analysis of all aspects of confined versus open/free range production of pigs. However, it may be worthwhile to outline other impacts at a high level to demonstrate that the issue is complex and multidimensional. We have not found any studies providing verified quantifiable *human health benefits* from consuming pork from open/free range production that could offset the risks quantified in this research. Since open/free range production is often associated with antibiotic free production, many consumers may prefer it on that basis. However, there is currently no scientific evidence that pork from pigs produced using conventional methods, which could sometimes include subtherapeutic use of antibiotics as a feed additive, or use at higher levels to treat disease under prescribed conditions, poses a greater human health risk from antibiotic-resistant organisms (Cox and Popken 2014; Cox and Popken 2010; Cox et al. 2009). Nor should open/free range pigs be considered free from such organisms. Weese et al. (2011) performed a longitudinal study showing that piglets raised on an antimicrobial free farm in Canada, a country even more restrictive than the U.S. regarding antibiotic use in food animals, had MRSA prevalence rates as high as 65%. O'Brien et al. (2012) found no statistical difference in MRSA rates between meat from hogs raised conventionally and meat from hogs raised without antibiotics. Antibiotic free (ABF) and organic pig farms have also been reported to have higher prevalence rates than conventional farms of bacteria such as *Salmonella* (Gebreyes et al. 2008). Pigs raised on outdoor pastures may also be more subject to infection by parasites and certain other diseases, but perhaps less subject to respiratory diseases

(Jolie et al. 1998). In Denmark, the parasite problem is considered to be a threat to outdoor pig production (Roepstorff et al. 2011).

Still, consumers may perceive health benefits from eating open/free range pork, and may also have concerns regarding animal welfare (even though pigs raised indoors are less subject to temperature extremes and sunburn) and environmental impacts. Consumers may not be aware of new and old (re-emerging) human health threats posed by open/free range production practices. Kijlstra et al. (2009) examined policy issues related to increased use of organic livestock production in the Netherlands with reference to increased Toxoplasmosis in free range pigs, and higher dioxin levels in the eggs of free range hens. They identified the need for increased communication of risks to consumers to avoid even more significant problems in the future. This could be achieved in the U.S. by labeling requirements, increased use of USDA awareness bulletins and newsletters, and training directed at farm veterinarians on how to identify and mitigate high prevalence situations. Organic farm producers need to be willing to undertake reasonable risk mitigation steps, such as reducing swine contact with cats and rodents. Consumers, especially pregnant women, need to use caution in handling raw pork, and thoroughly cook all meat before eating.

Quantitative risk assessment alone cannot determine whether the trade-offs identified in this study are desirable from a public policy point of view, but it clarifies the approximate magnitude of the plausible human health harm that might realistically be expected to be caused by greater use of open/free range production systems.

References

- Batz MB, Hoffmann S, Morris JG Jr (2012) Ranking the disease burden of 14 pathogens in food sources in the United States using attribution data from outbreak investigations and expert elicitation. *J Food Prot* 75(7):1278–1291
- Cook AJC, Holliman R, Gilbert RE, Buffolano W, Zufferey J, Petersen E, Jenum PA, Foulon W, Semprini AE, Dunn DT (2000) Sources of toxoplasma infection in pregnant women: European multicentre case-control study. Commentary: Congenital toxoplasmosis—further thought for food. *BMJ* 321(7254):142–147
- Cox LA Jr, Popken DA (2014) Quantitative risk assessment of human MRSA risks from swine. *Risk Anal* 39(9):1639–1650
- Cox LA Jr, Popken DA (2010) Assessing potential human health hazards and benefits from subtherapeutic antibiotics in the United States: tetracyclines as a case study. *Risk Anal* 30(3):432–457
- Cox LA Jr, Popken DA, Mathers J (2009) Human health risk assessment of penicillin/aminopenicillin resistance in enterococci due to penicillin use in food animals. *Risk Anal* 29 (6):796–805
- Cressey P, Lake R (2005) Ranking food safety risks: development of NZFSA policy 2004–2005. Institute of Environmental Science and Research (online report). Available at http://www.foodsafety.govt.nz/elibrary/industry/Ranking_Food_Safety_Risks-Science_Research.pdf. Accessed 13 Feb 2013
- Davies PR (2011) Intensive swine production and pork safety. *Foodborne Pathog Dis* 8(2):189–201

- Davies PR, Morrow WEM, Gamble HR, Deen J, Patton S (1998) Prevalence of antibodies to *Toxoplasma gondii* and *Trichinella spiralis* in finishing swine raised in different production systems in North Carolina, USA. *Prev Vet Med* 36(1):67–76
- Davis R (2008) Teaching project simulation in Excel using PERT-Beta distributions. *INFORMS Trans Educ* 8(3):139–148
- Dubey JP (2009) Toxoplasmosis in pigs—the last 20 years. *Vet Parasitol* 164(2–4):89–103
- Dubey JP (2013) Swine toxoplasmosis. Veterinary Division - Animal Health Programs (website). Available at <http://www.ncagr.gov/vet/FactSheets/Toxoplasmosis.htm>. Accessed 25 Feb 2013
- Dubey JP, Leighty JC, Beal VC, Anderson WR, Andrews CD, Thulliez P (1991) National seroprevalence of *Toxoplasma gondii* in pigs. *J Parasitol* 77(4):517–521
- Dubey JP, Gamble HR, Hill D, Sreekumar C, Romand S, Thuilliez P (2002) High prevalence of viable *Toxoplasma gondii* infection in market weight pigs from a farm in Massachusetts. *J Parasitol* 88:1234–1238
- Dubey JP, Hill DE, Sundar N, Velmurugan GV, Bandini LA, Kwok OCH, Pierce V, Kelly K, Dulin M, Thulliez P, Iwueke C, Su C (2008) Endemic toxoplasmosis in pigs on a farm in Maryland: isolation and genetic characterization of *Toxoplasma gondii*. *J Parasitol* 94:36–41
- Dubey JP, Hill DE, Rozeboom DW, Rajendran C, Choudhary S, Ferreira LR, Kwok OCH, Su C (2012) High prevalence and genotypes of *Toxoplasma gondii* isolated from organic pigs in northern USA. *Vet Parasitol* 188(1–2):14–18
- Gebreyes WA, Bahnsen PB, Funk JA, McKean J, Patchanee P (2008) Seroprevalence of *Trichinella*, *Toxoplasma*, and *Salmonella* in antimicrobial-free and conventional swine production systems. *Foodborne Pathog Dis* 5(2):199–203
- van der Giessen J, Fonville M, Bouwknegt M, Langelaar M, Vollema A (2007) Seroprevalence of *Trichinella spiralis* and *Toxoplasma gondii* in pigs from different housing systems in The Netherlands. *Vet Parasitol* 148(3–4):371–374
- Guerina NG, Hsu H-W, Meissner HC, Maguire JH, Lynfield R, Stechenberg B, Abroms I, Pasternack MS, Hoff R, Eaton RB, Grady GF (1994) Neonatal serologic screening and early treatment for congenital *Toxoplasma gondii* infection. *N Engl J Med* 330(26):1858–1863
- Hamilton B, Sutton PD (2012) Recent trends in births and fertility rates through June 2012 (website). Available at http://www.cdc.gov/nchs/data/hestat/births_fertility_june_2012/births_june_2012.pdf. Accessed 3 Jan 2013
- Havelaar AH, Galindo AV, Kurowicka D, Cooke RM (2008) Attribution of foodborne pathogens using structured expert elicitation. *Foodborne Pathog Dis* 5(5):649–659
- Hays SM (1996) The cat/pig Toxoplasmosis connection. *Agric Res* 44(2):8–9
- Hill DE, Dubey JP (2013) *Toxoplasma gondii* prevalence in farm animals in the United States. *Int J Parasitol* 43(2):107–113
- Hill DE, Haley C, Wagner B, Gamble HR, Dubey JP (2010) Seroprevalence of and risk factors for *Toxoplasma gondii* in the US swine herd using sera collected during the National Animal Health Monitoring Survey (Swine 2006). *Zoonoses Public Health* 57(1):53–59
- Hill DE, Baroch J, Swafford SR, Fournet VM, Pyburn DG, Schmit BB, Gamble HR, Feidas H, Theodoropoulos G (2014) Surveillance of feral swine for *Trichinella* spp. and *Toxoplasma gondii* in the US and host-related factors associated with infection. *Vet Parasitol* 205:653–655
- Hoffmann S, Batz MB, Morris JG Jr (2012) Annual cost of illness and quality-adjusted life year losses in the United States due to 14 foodborne pathogens. *J Food Prot* 75(7):1292–1302
- Jolie R, Backstrom L, Pinckney R, Olson L (1998) Ascarid infection and respiratory health in feeder pigs raised on pasture or in confinement. *Swine Health Prod* 6(3):115–120
- Jones JL, Kruszon-Moran D, Wilson M, McQuillan G, Navin T, McAuley JB (2001) *Toxoplasma gondii* infection in the United States: seroprevalence and risk factors. *Am J Epidemiol* 154 (4):357–365
- Jones JL, Kruszon-Moran D, Sanders-Lewis K, Wilson M (2007) *Toxoplasma gondii* Infection in the United States, 1999–2004, Decline from the Prior Decade. *Am J Trop Med Hyg* 77 (3):405–410
- Kijlstra A, Eissen OA, Cornelissen J, Munniksma K, Eijck I, Kortbeek T (2004) *Toxoplasma gondii* infection in animal-friendly pig production systems. *Invest Ophthalmol Vis Sci* 45 (9):3165–3169

- Kijlstra A, Meerburg BB, Bos P (2009) Food safety in free-range and organic livestock systems: risk management and responsibility. *J Food Prod* 72(12):2629–2637
- McKean J, O'Conner A, Pyburn D, Beary J (2009) Survey of market swine to determine prevalence of Toxoplasma in meat juice samples from selected abattoirs. In: 8th international symposium: epidemiology and control of foodborne pathogens in pork, Quebec City, Canada
- Mead PS, Slutsker L, Dietz V, McCraig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV (1999) Food-related illness and death in the United States. *Emerg Infect Dis* 5(5):607–625
- O'Brien AM, Hanson BM, Farina SA, Wu JY, Simmering JE, Wardyn SE, Forshey BM, Kulick ME, Wallinga DB, Smith TC (2012) MRSA in conventional and alternative retail pork products. *PLoS One* 7(1):e30092
- Patton S, Zimmerman J, Roberts T, Faulkner C, Diderrick V, Assadi-Rad A, Davies P, Kliebenstein J (1996) Seroprevalence of *Toxoplasma gondii* in hogs in the National Animal Health Monitoring System (NAHMS). *J Eukaryot Microbiol* 43(5):121S
- Patton S, Faulkner C, Anderson A, Smedley K, Buch E (2002) *Toxoplasma gondii* Infection in sows and market-weight pigs in the United States and its potential impact on consumer demand for pork. National Pork Board Report NPB# 00-130 (online report)
- Roepstorff A, Mejer H, Nejsum P, Thamsborg S (2011) Helminth parasites in pigs: new challenges in pig production and current research highlights. *Vet Parasitol* 180(1–2):72–81
- Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM (2011) Foodborne illness acquired in the United States—major pathogens. *Emerg Infect Dis* 17(1): 7–15. Also available at: http://wwwnc.cdc.gov/eid/article/17/1/p1-1101_article.htm. Accessed 26 Feb 2014
- Tenter A, Heckereth A, Weiss L (2000) *Toxoplasma gondii*: from animals to humans. *Int J Parasitol* 30(12–13):1217–1258
- Thomas MK, Murray R, Flockhart L, Pintar K, Pollari F, Fazil A, Nesbitt A, Marshall B (2013) Foodborne Pathog Dis 10(7):639–648
- USDA-APHIS (2008) Biosecurity on U.S. Swine Sites. USDA-APHIS (online report). Available at http://www.aphis.usda.gov/animal_health/nahms/swine/downloads/swine2006/Swine2006_is_biosecurity.pdf. Accessed 2 Jan 2013
- USDA-APHIS-VS-CEAH (2011) Seroprevalence of *Trichinella* and *Toxoplasma* in US grower/finisher pigs, 2006. USDA-APHIS (online report). Available at http://www.aphis.usda.gov/animal_health/nahms/swine/downloads/swine2006/Swine2006_is_trich.pdf. Accessed 31 Jan 2013
- USDA-ERS (2012) Table 2. U.S. certified organic farmland acreage, livestock numbers, and farm operations. US Dept of Agriculture, Economic Research Service (online file). Available at http://www.ers.usda.gov/datafiles/Organic_Production/National_Tables/_Farmlandlivestockandfarm.xls. Accessed 23 Feb 2014
- USDA-NASS (2012) Meat animals production, disposition, and income: 2011 summary. USDA-NASS (online database). Available at <http://www.ers.usda.gov/data-products/agricultural-base-line-database.aspx>. Accessed 22 Feb 2014
- Vaillant V, de Valk H, Ancelle T, Colin P, Delmas MC, Dufour B, Pouillot R, Le Strat Y, Weinbreck P, Jouglard E, Desenclos JC (2005) Foodborne infections in France. *Foodborne Pathog Dis* 2(3):221–232
- Weese J, Zwambag A, Rosendal T, Reid-Smith R, Friendship R (2011) Longitudinal investigation of methicillin-resistant *Staphylococcus aureus* in piglets. *Zoonoses Public Health* 58(4):238–243

Chapter 8

How Well Can High-Throughput Screening Tests Results Predict Whether Chemicals Cause Cancer in Mice and Rats?



Introduction

Over the past half century, an enduring intellectual and technical challenge for risk analysts, statisticians, toxicologists, and experts in artificial intelligence, machine-learning and bioinformatics has been to predict *in vivo* biological responses to realistic exposures, with demonstrably useful accuracy and confidence, from *in vitro* and chemical structure data. The common goal of many applied research efforts has been to devise and validate algorithms that give trustworthy predictions of whether and by how much realistic exposures to chemicals change probabilities of adverse health responses. This chapter examines recent, promising results suggesting that high-throughput screening (HTS) assay data can be used to predict *in vivo* classifications of rodent carcinogenicity for certain pesticides. Anticipating the focus on evaluation analytics for assessing the performance of systems, policies, and interventions in Chaps. 9 and 10, it also undertakes an independent reanalysis of the underlying data to determine how well this encouraging claim can be replicated and supported when the same data are analyzed using slightly different methods.

In principle, every student of statistics or bioinformatics with access to relevant data is equipped to participate in the challenge of constructing useful predictive models. All that is required is to define one or more dependent variables indicating *in vivo* responses to chemical exposures in one or more test species of interest (e.g., rodents); to link data on these responses for a database of chemicals to results of one or more *in vitro* assays (e.g., for genotoxicity, gene mutations, or chromosomal damage in bacteria and in mammalian cell cultures) and/or chemical structure features, to be used as predictors; and then to apply one's favorite predictive analytic techniques to see how well they can predict *in vivo* responses from the selected predictors. Investigators with an interest in systems biology may use envisioned possible causal pathways and mechanisms or modes of action to guide or rationalize their selection of predictors, while others may prefer pure black-box statistical

methods that simply seek the most predictive patterns, whether or not they conform to any biological theory or model. Over the decades, many predictive analytics techniques have been tried, from clustering and regression modeling to expert systems (largely in the 1970s–1990s) to artificial neural networks to current machine-learning methods such as Random Forest, ensembles of Bayesian Networks, or support vector machines. For any predictive technique, important questions of training and test set design, model validation, sensitivity and specificity of predictions, and generalizability of results arise. Yet, the intrinsic interest and importance of the topic and the comparative ease of addressing it by applying predictive analytics software has generated a large literature, replete with comparisons among methods for various training and test data sets.

Despite these many efforts, the results of decades of predictive modeling remain, at best, very mixed. For example, in the 1990s, several quantitative structure-activity relationship (QSAR) computer programs were developed to screen inventories of chemicals for mutagenicity and carcinogenicity. Some of the most commonly used systems (Deductive Estimation of Risk from Existing Knowledge (DEREK), Toxicity Prediction by Computer-Assisted Technology (TOPKAT), and Multiple Computer Automated Structure Evaluation (MCASE)) were promoted as being valuable for predicting these endpoints (e.g., Patlewicz et al. 2003), and were used accordingly by regulators and companies to screen and prioritize chemicals for risk assessment. EPA continues to develop, endorse, and apply such models, claiming that they have “demonstrated reasonable predictive abilities” (EPA 2011). However, when they were applied to a set of chemicals of great practical interest—a panel of 394 marketed pharmaceuticals—all turned out to have less than 52% sensitivity for predicting positive Ames assays, and even worse accuracy for predicting other genotoxic assay results (Snyder et al. 2004): “20% of the 124 non-carcinogens were positive in at least one genotoxicity assay while two-thirds of the 77 rodent carcinogens lacked activity in the genotoxicity tests employed” (Guyton et al. 2009).

Similarly, even the most successful systems in an experiment that tested how well the best predictive algorithms could predict rodent carcinogenicity for 30 chemicals had only about 60–65% accuracy (Benigni and Zito 2004). Valerio et al. (2007) reported 97% sensitivity but only 53% specificity for software used by the Food and Drug Administration (FDA) to predict rodent carcinogenicity of naturally occurring molecules found in human diets (i.e., few false negatives but many false positives), and Valerio et al. (2010) found that two such software programs “both exhibited poor performance in predicting [rodent] carcinogens” when evaluated in an external validation study of 43 phytochemicals. Walmsley and Billinton (2011) note that even the reported high sensitivity for some current *in vitro* test batteries for predicting rodent carcinogenicity (e.g., about 9 of 10 rodent carcinogens correctly classified as such based on combinations of bacterial and mammalian cell tests) (March Kirkland et al. 2005) is less encouraging than it seems, insofar as the same test batteries also misclassify as many as 9 out of 10 non-carcinogens as being carcinogens. In other words, the predictive power of the *in vitro* test batteries is not much better than would be achieved by simply assuming that all chemicals are rodent carcinogens, thus creating excellent sensitivity (no false negatives) but poor

specificity (many false positives). The authors note that many potentially pharmaceutical compounds now classified as probable carcinogens based on genotoxicity results in bacterial and mammalian cells may not be carcinogens at all.

Finally, a long-standing literature notes that even a system that could accurately predict rodent carcinogenicity might have little value for predicting human carcinogenicity. For example, some have argued that EPA estimates far more chemicals as being carcinogenic in humans than do other authorities, such as the International Agency for Research on Cancer (IARC), due largely to over-reliance on animal data and to the limitation that “the true predictivity for human carcinogenicity of animal data is even poorer than is indicated by EPA figures alone” (Knight et al. 2006). Differences across species in gross anatomy (e.g., no Harderian, Zymbal, or preputial glands in humans), pharmacokinetics and pharmacodynamics, and species-specific modes of action (e.g., protein drop nephropathy in male rat kidneys) can all make rodent carcinogenicity of uncertain relevance to human carcinogenicity. This concern, though important, is beyond the scope of this chapter, which focuses on the narrower question of how well rodent carcinogenicity can be predicted from *in vitro* assay data and chemical structure (QSAR) information.

In light of this history of limited predictive performance, it is worth asking the following three methodological questions.

1. *How good (i.e., how predictively accurate, e.g., as indicated by rates of false positives and false negatives in external validation test sets) are current rodent carcinogen prediction systems?*
2. *Are there fundamental limitations on the ability of computer algorithms to predict accurately (with high sensitivity and specificity) the carcinogenicity in rodents of a broad range of chemicals?* For example, are there limits on the kinds of concepts or patterns that a computer algorithm can learn from data, and is there reason to believe that a concept such as “is a rodent carcinogen” can be learned accurately from data on chemicals and assay results? If not—if the desired classification rules are not learnable from examples, even with the help of guidance based on biological knowledge—then such a fundamental limitation cannot necessarily be removed by simply collecting more data, or by obtaining more knowledge. Rather, it may simply be impossible to learn accurate predictive classification rules or procedures from available data. To our knowledge, potential fundamental limitations on the learnability of accurate prediction rules for classifying chemicals as rodent carcinogens, or for predicting cancer dose-response relations based on chemical and *in vitro* data, have not previously been carefully studied. Appropriate methods for addressing fundamental limitations of what is learnable are available in the machine-learning literature (e.g., using the [Probably Approximately Correct](#) learning frameworks and alternatives), but these methods have not as yet had much impact on QSAR or systems biology research used in carcinogenicity prediction systems.
3. *What practical constraints limit predictive performance,* e.g., due to incomplete knowledge of relevant biological mechanisms, or limited sizes and diversity of training and test data sets? Unlike fundamental limitations, such practical

limitations might be removed by further scientific research. This chapter and its appendix touch briefly on practical aspects of training set design, classification of chemicals with respect to rodent carcinogenicity, external test set design and validation, and extrapolation of risk predictions from tested to untested chemicals.

The following sections focus mainly on the first of these questions, critically evaluating the performance of a predictive scoring system and of the underlying HTS assays for rodent carcinogenicity of pesticides studied by Kleinstreuer et al. (2013). Questions 2 and 3—fundamental limitations to predictability and practical issues in developing and testing predictive scores or classifications—are also touched on as needed to understand the strengths and limitations of this predictive model.

For as long as there has been interest in developing algorithms to screen chemicals for likely *in vivo* activities based on relatively inexpensive QSAR and *in vitro* assay results, there have also been claims of encouragingly accurate performance of the systems in current use. However, as discussed in the preceding references, these optimistic appraisals have generally not been followed by equally good performance on external validation tests sets, when predictions must be made in advance of knowing the correct answers (e.g., Benigni and Zito 2004). Claims of accurate prediction therefore deserve to be scrutinized. We do so using the recently published National Center for Computational Toxicology (NCCT) research article by (Kleinstreuer et al. 2013) as a case study. The article, promisingly entitled “*In vitro* perturbations of targets in cancer hallmark process predict rodent chemical carcinogenesis,” is an exciting piece of work. It applies contemporary knowledge of the biological “hallmarks of carcinogenesis” framework to inform selection and combination of high throughput screening (HTS) assay results that might indicate the activation of different causal pathways involved in causation of cancer. Kleinstreuer et al. conclude that “A simple scoring function... applied to an external test set of 33 compounds with carcinogenicity classifications from the EPA’s Office of Pesticide Programs... successfully ($p = 0.024$) differentiated between chemicals classified as ‘possible’/‘probable’/‘likely’ carcinogens and those designated as ‘not likely’ or with ‘evidence of noncarcinogenicity.’ This model represents a chemical carcinogenicity prioritization tool supporting targeted testing and functional validation of cancer pathways.” The following sections re-examine these encouraging findings in detail, and seeks to independently reproduce them. It concludes that, despite the promising and plausible research direction embodied in the proposed scoring approach, the claimed predictive accuracy appears to be an artifact of errors in classification of chemicals and in selection and use of statistical methods. It is not clear based on the data analyzed and associations reported between model predictions and chemical carcinogenicity classifications whether the system’s true performance is better than random guessing. However, a different analysis that examines how rodent carcinogenicity classification counts vary across chemicals with different predictive scores suggests that the simple scoring model does indeed have useful predictive power.

Case Study: Reassessing the Accuracy and Robustness of a Rodent Carcinogenicity Prediction System

Purpose, Scope, and Interpretation of the Original Study

The following abstract from the article by Kleinstreuer et al. (2013) succinctly expresses the motivation, ambitions, rationale, and hoped-for results from important current efforts to use biological knowledge to help improve prediction of the *in vivo* rodent carcinogenicity of chemicals from relatively inexpensive high-throughput screening (HTS) *in vitro* assay results. (See Guyton et al. 2009, for a similar approach.)

“Thousands of untested chemicals in the environment require efficient characterization of carcinogenic potential in humans. A proposed solution is rapid testing of chemicals using *in vitro* high-throughput screening (HTS) assays for targets in pathways linked to disease processes to build models for priority setting and further testing. We describe a model for predicting rodent carcinogenicity based on HTS data from 292 chemicals tested in 672 assays mapping to 455 genes. All data come from the EPA ToxCast project. The model was trained on a subset of 232 chemicals with *in vivo* rodent carcinogenicity data in the Toxicity Reference Database (ToxRefDB). Individual HTS assays strongly associated with rodent cancers in ToxRefDB were linked to genes, pathways, and hallmark processes documented to be involved in tumor biology and cancer progression. . . .A simple scoring function was generated to identify chemicals with significant *in vitro* evidence that was predictive of *in vivo* carcinogenicity in different rat tissues and organs. This scoring function was applied to an external test set of 33 compounds with carcinogenicity classifications from the EPA’s Office of Pesticide Programs and successfully ($p = 0.024$) differentiated between chemicals classified as ‘possible’/‘probable’/‘likely’ carcinogens and those designated as ‘not likely’ or with ‘evidence of noncarcinogenicity.’ This model represents a chemical carcinogenicity prioritization tool supporting targeted testing and functional validation of cancer pathways.”

The chemicals of primary interest in this case study are food crop pesticides that are believed to operate through non-genotoxic mechanisms to induce one or more of the currently recognized hallmarks of carcinogenesis, i.e., sustained proliferative signaling, evasion of growth suppression signals, evasion of immune detection and of destruction of compromised cells, acquisition of replicative immortality, tumor-promoting inflammation, active invasion and metastasis, induction of neoangiogenesis, increased genome instability and mutation, evasion of apoptosis, and deregulation of cellular energetics (Hanahan and Weinberg 2011).

Kleinstreuer et al. interpreted their work as testing the following hypothesis:

*H1: “Chemicals that perturb certain cancer-linked targets or processes in human *in vitro* HTS assays will have a significantly higher likelihood of being carcinogens, as evidenced by carcinogenicity in the 2-year chronic assays in rodents.”*

The corresponding null hypothesis is:

*H₀: Chemicals identified as perturbing relevant pathways based on the *in vitro* HTS assay results are no more likely than other chemicals to exhibit carcinogenicity in the 2-year chronic assays in rodents.*

They concluded that their data support hypothesis H₁ by allowing confident rejection of H₀. They also propose a prioritization method that scores the possible carcinogenic potentials of chemicals by counting the number of cancer-associated endpoints that are identified as “significantly perturbed” in assay screening.

Original Data, and Replication Process and Results

Kleinsteuer et al. (2013) began their analysis by searching for assays that predict rodent carcinogenicity based on finding significantly increased univariate odds ratios (ORs) for a chemical being classified as a rodent carcinogen if the assay is positive compared to if it is not. ORs were assessed in a training set of 292 chemicals for which both *in vitro* assay results and *in vivo* 2-year rodent chronic assay results were available. These 292 chemicals were ToxCast Phase I chemicals for which 2-year chronic cancer bioassay data are available from the EPA Toxicity Reference Database (ToxRefDB, <http://actor.epa.gov/toxrefdb/>). This database classifies each chemical as positive or negative for preneoplastic or neoplastic lesions in rats (232 chemicals) and mice (223 chemicals, 200 of which overlap with those for rats). Only the most common cancer endpoints were included for each species. The most common endpoints were defined as those that were positive for at least 20 chemicals. These endpoints were described as liver preneoplastic or neoplastic, lung preneoplastic, and spleen preneoplastic for mice; and as kidney preneoplastic, liver preneoplastic or neoplastic, testes preneoplastic or neoplastic, and thyroid preneoplastic or neoplastic for rats. The 292 chemicals used by Kleinsteuer et al. are listed in their Supplementary Table 1.

To independently replicate and validate this approach and quantify the predictivity of the HTS data for rodent carcinogenicity, we obtained data and software from the article publication site, the study authors, and the National Pesticide Information Center. Data files used in the study were obtained from the journal article publication site (<http://toxsci.oxfordjournals.org/content/131/1/40/suppl/DC1>). The files listed in the article as Supplementary Tables 1–6 mapped to the following file names (Table 8.1):

We converted the pdf files File007, File008, File010 to data files via a combination of OCR software and manual correction.

Upon request, the corresponding author, Dr. Richard Judson sent the software used in the study and the full set of input and output tables in electronic form. Table 8.2 lists several of these files which proved key to validating the results.

For their training set, as described in the article, Kleinsteuer et al. (2013) drew on a database of 309 unique chemicals, 292 of which were flagged for usage. Of the

Table 8.1 Data files from journal website

File as named in Kleinstreuer et al. (2013)	File as named at website	Contents
Supplementary Table 1	toxsci_12_0526_File007.pdf	Chemical codes/names and indicator flag of use in study
Supplementary Table 2	toxsci_12_0526_File008.pdf	Table of hits by chemical and in-vivo endpoint
Supplementary Table 3	toxsci_12_0526_File009.xlsx	AC50 values by chemical and assay
Supplementary Table 4	toxsci_12_0526_File010.pdf	Gene mappings and process counts by assay
Supplementary Table 5	toxsci_12_0526_File011.pdf	Figure 3 diagram
Supplementary Table 6	toxsci_12_0526_File012.xlsx	Research article database

Table 8.2 Additional key files sent by NCCT on request

File name	Contents
Hallmark2.R	R software used to generate odds ratios, identify significant variables, and compute predictive scores
ORforestData_assay.txt	Odds ratios, LCIs, UCIs, and 2×2 contingency table values for individual assay-endpoint combinations
ORforestData_assay_perms.txt	Odds ratios, LCIs, and UCIs for “permuted” chemicals, one set for each endpoint
CancerPred_all_ALL.txt	Cancer prediction scores for all 292 chemicals (similar to Table 1, but expanded to cover training chemicals)

292 chemicals used, 60 had no rat-related in-vivo endpoint data (232 remaining), while 69 had no mouse-related in-vivo endpoint data (223 remaining). The database provided AC50 data values for 664 unique assays for each of the 309 chemicals. There were 673 assay names in the gene mapping file. (Six of the 664 assay columns in the AC50 data file had names that were not listed in the gene mapping file. All AC50 values for these 6 were zero, however, as was also true of many other assays. This did not impact the results.)

For their test data set, Kleinstreuer et al. (2013) first identified 60 chemicals, listed in their Table 1 (“Summary of cancer hazard model for chemicals not included in the training set for rat endpoints”) that had not been used in constructing their risk prediction model, but that had *in vitro* assay results. From these, they selected as their final external validation test set a subset of 33 chemicals that had EPA Office of Pesticide Programs (OPP) human carcinogenicity classifications (shown in the last column of their Table 1). They note that “these ‘human’ classifications are in reality a summary of data from [largely] rodent studies and so are comparable with the data used in developing the model.” For purposes of testing and validating model predictions, Kleinstreuer et al. assigned any of the 60 test set chemicals with a classification containing the words “Likely”, “Probable”, or “Possible” for OPP’s assessment of carcinogenic potential a value of 1, and assigned those containing

“Not Likely” or “Evidence of noncarcinogenicity” a value of 0. They excluded all other chemicals, e.g., those with classifications such as “Not classifiable” or “Suggestive evidence”, or “Insufficient data.”

The data provided to us by NCCT did not include an electronic version of their Table 1, the external validation test data set. We therefore independently obtained a pdf document from the National Pesticide Information Center at http://npic.orst.edu/chemicals_evaluated.pdf, “Chemicals Evaluated for Carcinogenic Potential” (Nov 2012), published by the EPA Office of Pesticide Programs (OPP), that contains the cancer classification for most of the chemicals used in this study. We converted the pdf file to data via OCR software to support automated computations. (The URL in the article for the 2010 version of this document links to a website that states that the paper file can only be obtained by phone or mail.) We then manually checked each entry in Table 1 of Kleinstreuer et al. (2013) to verify the accuracy of the provided classifications of carcinogenic potency.

Appendix A presents details of our chemical-by-chemical review and replication effort. Although we were able to independently confirm most (58 out of 60) of the carcinogenicity classifications in Kleinstreuer et al.’s Table 1 as matching those provided by OPP, we found discrepancies for 2 out of 60 chemicals. For these chemicals, the classifications reported by Kleinstreuer et al. and attributed to OPP differ from those in the EPA/OPP published data that we retrieved. These two cases are as follows.

- Methylene bis(thiocyanate) [CASRN 6317–18-6] (MITC) is shown in Table 1 of Kleinstreuer et al. with an EPA carcinogenic potential classification of “Likely to be carcinogenic—based on metam sodium data”. However the EPA/OPP document states: “There are insufficient data to characterize the cancer risk of MITC,” with a report date of 2009. The computed cancer hazard score for that MITC is relatively high, at 16. It is unclear whether the stated rationale, “...based on metam sodium data” was inserted by Kleinstreuer et al. Although MITC is a breakdown product of metam sodium, it is not formed quickly and the use of the metam sodium risk information as a surrogate for it is scientifically questionable (www.epa.gov/opprrd1/REDs/3082red.pdf). Indeed, separate EPA documentation explicitly states that “it is not appropriate to quantify MITC cancer potential using the metam sodium cancer slope factor. . . .” www.epa.gov/pesticides/chem-search/cleared_reviews/csr_PC-068103_13-May-04_a.pdf. In any case, the classification of MITC in Table 1 of Kleinstreuer et al. does not match that provided by OPP.
- Etridiazole [also called Terrazole, CASRN 2593–15-9] is shown in Table 1 of the article as having “No data.” However, the EPA/OPP document described above classifies it as “Group B—Probable Human Carcinogen”, with a report date of 1999. In more detail, “Etridiazole was classified by the Agency’s Health Effects Division Cancer Peer Review Committee (CPRC) as a Probable Human Carcinogen. This classification is based on the following factors: (i) occurrence of multiple tumor types in male and female rats (tumor sites noted were the liver, bile duct, mammary gland, thyroid, and testes) including the induction of a rare

bile duct tumor (cholangiocarcinoma), and (ii) non-neoplastic lesions observed in similar target organs that lend support to the association of etridiazole exposure with the induction of tumors; increased absolute and relative liver weight (males), hepatocytomegaly (males); clear, basophilic, and eosinophilic cellular alterations (males and females); cholangiectasis (females); centrilobular pigmentation (females); spongiosishepatitis of the liver (males); and testicular interstitial cell hyperplasia (males) and (iii) positive mutagenicity data. The carcinogenicity study in mice was determined to be unacceptable and not adequate for assessment of the carcinogenic potential of etridiazole in this species” <http://www.epa.gov/oppsrrd1/REDs/0009red.pdf>. The cancer hazard score for this chemical computed by Kleinstreuer et al.’s model was zero, so the fact that it was misclassified as having “No data” prevented an important discrepancy between this prediction and the multiple tumor types observed in rats from being taken into account.

We obtained matching classifications for all of the remaining 58 chemicals in Table 1 of Kleinstreuer et al. (or had a similar conclusion of “no data” where that was shown). However, the discrepancies for MITC and etridiazole have a significant impact on the study results, as discussed later.

Appendix A to this chapter provides additional details of our findings for all 60 chemicals, including several cases in which current knowledge might lead to changes in the EPA-OPP classifications used by Kleinstreuer et al. However, as our main goal is simply to verify whether the claimed accuracy can be confirmed using the same data and methods as the original authors as far as possible, we do not attempt to update or correct their Table 1 except for MITC and etridiazole, where it appears that the stated methodology was not followed correctly.

Original Methods, and Replication Process and Results

In addition to independently reproducing the data in the training data set and external validation data set (with the exceptions just discussed), we also sought to reproduce the methodology used by Kleinstreuer et al. (2013) as far as possible, based on the documentation provided and the additional data files and software obtained from the authors. This section describes how we replicated these methods.

Replicating Selection of Predictors Based on Significant Assay-Endpoint Combinations

Kleinstreuer et al. (2013) determined which chemical/assay combinations would be considered “significant” for predicting carcinogenic potential using odds ratio values, as well as their 95% confidence intervals, from two sources: univariate odds ratios, and average “permuted” odds ratios, as described below and in paragraph 3 of Materials and Methods in the article. In several places where the article’s

discussion was insufficient for us to reproduce results, the R software code provided by the authors gave the needed clarification.

Kleinstreuer et al. first computed an odds ratio for each assay/endpoint combination (664 x 11) using the vector of AC50 values for the assays (with non-zeros converted to 1) and the 0/1 vector of endpoint values. In accordance with the study procedures, any chemical rows corresponding to an “NA” endpoint result and any assay column with fewer than ten hits over the 292 chemicals used in the training set were eliminated. We used software that we developed independently in the Python programming language to check whether we could reproduce the study results.

Replicating the Univariate Odds Ratios Calculations

For each assay-endpoint combination, we fed the two binary vectors into a Python function that computed contingency table values n_{00} , n_{01} , n_{10} , and n_{11} corresponding to the counts of vector elements falling into each cell n_{xy} (with the convention that $x = \text{assay}$, $y = \text{endpoint}$). For instances where all four cell counts were non-zero, the function returned the odds ratio:

$$\text{OR} = n_{00} * n_{11} / n_{01} * n_{10},$$

and the standard error:

$$\text{SE} = \sqrt{1/n_{00} + 1/n_{11} + 1/n_{01} + 1/n_{10}}$$

In the case of any zero cell elements, $\text{OR} = 1$ and $\text{SE} = 0$ was returned (the computation was indeterminate and flagged for non-use). We computed the upper and lower 95% confidence intervals using the following standard asymptotic formulas:

$$\begin{aligned}\text{LCI}_{95} &= \exp(\log(\text{OR}) - 1.96 * \text{SE}) \\ \text{UCI}_{95} &= \exp(\log(\text{OR}) + 1.96 * \text{SE})\end{aligned}$$

These computations, verified as the same used by Kleinstruer et al. via the software code, correspond to the “Wald” method for calculating odds ratios based on a normal approximation to the log odds ratios. Other methods are available, including the exact Fisher (conditional maximum likelihood), midpoint (median unbiased), and small sample adjusted methods. Such alternative methods are appropriate when one or more cells have small values, as happens often in this study. These alternative methods are readily available in the R language epitools module (<http://cran.r-project.org/web/packages/epitools/index.html>). Experimentation showed that they often provide wider confidence intervals than the Wald method, potentially producing fewer false positives (in which an artificially narrow confidence interval wrongly excludes the null hypothesis of no significant relation, $\text{OR} = 1$). We verified that our software was computing the same odds ratios and confidence intervals by checking against the intermediate output files provided by Dr. Judson.

Table 8.3 Comparison of computed versus NCCT-provided endpoint CI values

End point	Computed LCI	NCCT LCI	Computed UCI	NCCT UCI
MLiver2Pre	0.326	0.326	2.65	2.65
MLiver3Neo	0.279	0.274	2.80	2.80
MLung2Pre	0.310	0.296	3.88	3.88
MSpleen2Pre	0.316	0.304	3.80	3.80
RKidney2Pre	0.291	0.291	3.23	3.23
RLiver2Pre	0.302	0.302	2.65	2.65
RLiver3Neo	0.314	0.314	3.60	3.60
RTestes2Pre	0.303	0.303	3.25	3.25
RTestes3Neo	0.304	0.310	3.53	3.53
RTroid2Pre	0.303	0.303	2.94	2.92
RTroid3Neo	0.299	0.299	3.12	3.03

Replicating the Permuted Odds Ratios Calculations

Kleinsteuer et al. (2013) also describe odds ratios, one per endpoint, determined by “permuting the endpoints and calculating the OR values for all assays”. The 95% confidence intervals for each endpoint are said to be calculated from the “OR distribution across all assays”. The precise meaning of these statements, described below, was made apparent by studying the provided R software code.

A single “permutation” involves taking a random permutation of the 0/1 endpoint vector, and computing the odds ratios for the permuted vector versus each of the 664 chemical assays. To obtain the result for a single endpoint, 10,000 permutations are performed, and the odds ratio vectors of results from each are concatenated into a single long vector (any indeterminate elements are removed). The vector is then sorted and the elements at 0.025 and 0.975 of the vector length are extracted. These values determine the 95% confidence intervals associated with that endpoint.

We developed the same procedure in Python using the same inputs. Table 8.3, produced using our Python code, provides results very similar (but not identical) to those of the original authors.

The small differences observed above are plausibly due to differences in random number generation, which resulted in different permutation sets; they did not affect the selection of significant assay-endpoint combinations.

Replicating the Significance Test for Predictor Variable Selection

We followed the same significance test used by Kleinsteuer et al. (2013): “Assay-endpoint pairs were considered significant if the CI for the pair did not include 1.0 (i.e., an OR of ‘no evidence of association’), and if the point estimate of the OR was outside of the 95% permutation test-derived CI for the endpoint.” Applying their additional filter of including only cases with three or more true positives ($n_{11} \geq 3$),

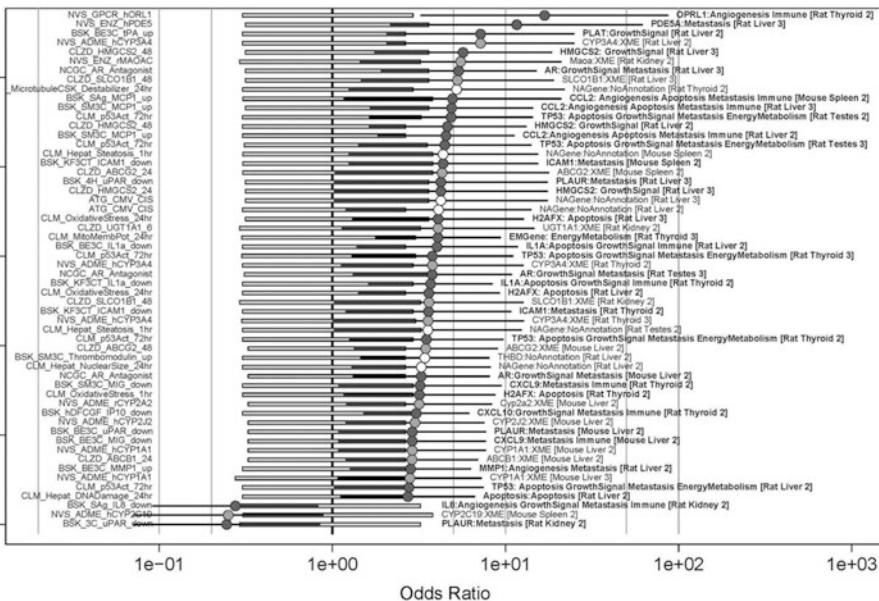


Fig. 8.1 Forest plot showing the mean OR and CIs for each significant association between *in vitro* assay and *in vivo* endpoint. Only associations with three or more true positives are shown. The colored circles give the point estimate of the OR and whiskers give the 95% CI. The gray bars indicate the endpoint-specific permutation test 95% CI. The linkage to types of processes is indicated by the color of the OR circle: dark gray is cancer hallmark-related, light gray is XME-related, and white is other. The assay name is listed at the far left. The associated gene, gene-related process, species, cancer type, and cancer severity level (2 = preneoplastic lesions, 3 = neoplastic lesions) are indicated to the right. A darker line indicates overlap of the assay-specific and the endpoint CIs (from Kleinstreuer et al. 2013)

we replicated the data underlying their Fig. 1 (odds ratio forest graph), reproduced as our Fig. 8.1. Our Table 8.4 summarizes the computed values underlying Fig. 8.1.

Our replication identified two significant assay-endpoint combinations that were not included in the Kleinstreuer et al. article or in the results data files that they provided. These two combinations are excerpted below:

Key	Assay	GeneSymbol	EndPoint	LCI	OR	UCI	PermLCI	PermUCLCI	TruePos
ATG_CMV_CIS_RLiver3Neo	ATG_CMV_CIS	NAGene	RLiver3Neo	1.030702	4.229167	17.35307	0.313793	3.596491	3
ATG_CMV_CIS_RLiver2Pre	ATG_CMV_CIS	NAGene	RLiver2Pre	1.202019	4.104	14.0121	0.301818	2.650794	6

They were not included in their Fig. 1. However, in their file of all endpoints computations, ORforestData_assay.txt, similar combinations are coded as follows:

Endpoint	Assay	OR	LCI	UCI	TP	FP	FN	TN
#CHR_Rat_Liver_2_PreneoplasticLesion	ATG_CMV_CIS	4.1	1.2	14	6	5	50	171
#CHR_Rat_Liver_3_NeoplasticLesion	ATG_CMV_CIS	4.23	1.03	17.4	3	8	18	203

Note that the OR related values are in agreement between the files. The permutation CIs for the endpoints (theirs not shown here) are also in agreement. It is the “#” symbol in the file that seems to have caused these combinations to be excluded from their calculations. We conclude that there was probably miscoding or corruption in the data files that mistakenly caused the “#” designation. Fortunately, the practical impacts are minor: oxyfluorfen (CASRN 42874-03-3) shown in Table 1 of the article should have a total score of 11 versus the 9 shown, and the Liver 2 and Liver 3 columns in that row should have 1 added to them, making them 3 and 2 respectively. This does not have a significant impact on the results and will be discussed further below.

Overall, we were able to reproduce most of the variable-selection process used by Kleinstreuer et al. (2013). However, we do not endorse this approach to variable selection based on odds ratios. We recommend instead using Bayesian Model Averaging, cross-validation, or other model ensemble methods to help overcome model selection, multiple comparison, and over-fitting biases that can inflate false-positive rates and reduce generalization accuracy. These sources of avoidable bias and error do not appear to have been adequately controlled in the OR-based selection procedure applied by Kleinstreuer et al. (2013).

Replicating the Cancer Hazard Scoring

The identified “significant” assay-endpoint combinations were used as predictor variables for calculating a cancer hazard score for each chemical. The score for each chemical is defined simply as a count of how many assays were activated, i.e., had non-zero AC50 values, that coincide with a significant (rat-related) assay-endpoint pair. Table 1 of Kleinstreuer et al. (2013) lists the resulting total scores, along with a breakdown by endpoint type for each of the 60 chemicals. Using our Python software, we reproduced all of the scores as shown, with the exception of the two table values just discussed.

Replicating the Comparison of the Model-Predicted Cancer Hazard Scores to EPA’S Binary Cancer Classifications

The key research question that the previous steps are intended to provide data to answer is: How well do the computed cancer hazard scores predict the externally derived 0/1 cancer potential classification scores provided by EPA-OPP? As previously discussed, the hazard scores were generated by applying the scoring procedure (counting the number of relevant activated assays), developed from the training set of 232 chemicals with (rat and mouse) in-vivo endpoint information, to a test set of 60 chemicals without in-vivo endpoint information (of which 33 had externally provided cancer classification data). To compare the predictive scores (counts) to the

Table 8.4 Replication of significant assay endpoint pairs and OR values

Assay	GeneSymbol	EndPoint	LCI	OR	UCI	PermLCI	PermUCL	TruePositives
NVS_GPCR_lMorL1	OPRL1	RTroid2Pre	3.25	16.76	86.55	0.30	2.94	6
NVS_ENZ_hpDE5	PDE5A	RLiver3Neo	2.17	11.56	61.46	0.31	3.60	3
BSK_BE3C_tPA_up	PLAT	RLiver2Pre	2.07	7.17	24.82	0.30	2.65	8
NVS_AdME_hCYP3A4	CYP3A4	RLiver2Pre	2.07	7.17	24.82	0.30	2.65	8
CLZD_HMGCS2_48	HMGCS2	RLiver3Neo	1.76	5.68	18.37	0.31	3.60	5
NVS_ENZ_rMAOAC	Maoa	RKidney2Pre	1.45	5.50	20.88	0.29	3.23	4
NCGC_AR_Antagonist	AR	RLiver3Neo	1.92	5.36	14.98	0.31	3.60	7
CLZD_SLC01B1_48	SLC01B1	RLiver3Neo	1.47	5.28	18.95	0.31	3.60	4
CLM_MicrotubuleCSK_Destabilizer_24hr	NAGene	RTroid2Pre	1.25	5.22	21.85	0.30	2.94	4
BSK_SAg_MCP1_up	CCL2	MSpleen2Pre	1.17	4.94	20.86	0.32	3.80	3
BSK_SM3C_MCP1_up	CCL2	RLiver3Neo	1.66	4.88	14.29	0.31	3.60	6
CLM_p53act_72hr	TP53	RTtestes2Pre	1.64	4.85	14.32	0.30	3.25	6
BSK_3C_Vis_down	Cytotox	MLung2Pre	1.76	4.85	13.34	0.31	3.88	13
CLZD_HMGCS2_48	HMGCS2	RLiver2Pre	1.64	4.62	13.07	0.30	2.65	9
BSK_SM3C_MCP1_up	CCL2	RLiver2Pre	1.84	4.53	11.16	0.30	2.65	12
CLM_p53act_72hr	TP53	RTtestes3Neo	1.42	4.46	14.00	0.30	3.53	5
BSK_KF3CT_ICAMI_down	ICAMI	MSpleen2Pre	1.25	4.36	15.27	0.32	3.80	4
CLM_Hepat_Steatosis_lhr	NAGene	MSpleen2Pre	1.25	4.36	15.27	0.32	3.80	4
CLM_CellLoss_72hr	Cytotox	RTroid3Neo	1.59	4.30	11.65	0.30	3.12	26
CLZD_ABCG2_24	ABCG2	MSpleen2Pre	1.04	4.30	17.73	0.32	3.80	3
BSK_4H_uPAR_down	PLAUR	RLiver3Neo	1.03	4.23	17.35	0.31	3.60	3
ATG_CMV_CIS	NAGene	RLiver3Neo	1.03	4.23	17.35	0.31	3.60	3
CLZD_HMGCS2_24	HMGCS2	RLiver3Neo	1.03	4.23	17.35	0.31	3.60	3
ATG_CMV_CIS	NAGene	RLiver2Pre	1.20	4.10	14.01	0.30	2.65	6
CLZD_UGT1A1_6	UGT1A1	RKidney2Pre	1.14	4.08	14.58	0.29	3.23	4
CLM_OxidativeStress_24hr	H2AFX	RLiver3Neo	1.31	4.08	12.68	0.31	3.60	5

CLM_CellLoss_72hr	Cytotox	RTroid2Pre	1.72	4.07	9.66	0.30	2.94	^~
CLM_MitoMembPFot_24hr	EMGene	RTroid3Neo	1.78	4.07	9.32	0.30	3.12	12
BSK_BE3C_Il1a_down	IL1A	RLiver2Pre	1.39	4.02	11.66	0.30	2.65	8
CLM_p53Act_72hr	TF53	RTroid3Neo	1.30	3.78	10.97	0.30	3.12	6
NVS_AdME_hCYP3A4	CYP3A4	RTroid2Pre	1.13	3.78	12.58	0.30	2.94	5
NCGC_AR_Antagonist	AR	RTestes3Neo	1.32	3.77	10.77	0.30	3.53	6
CLM_Hepat_CellLoss_lhr	Cytotox	RTroid2Pre	1.42	3.75	9.90	0.30	2.94	8
BSK_KF3CT_Il1a_down	IL1A	RTroid2Pre	1.62	3.69	8.37	0.30	2.94	12
CLZD_SLC01B1_48	SLC01B1	RKidney2Pre	1.03	3.61	12.63	0.29	3.23	4
CLM_OxidativeStress_24hr	H2AFX	RLiver2Pre	1.42	3.61	9.20	0.30	2.65	10
BSK_KF3CT_ICAM1_down	ICAM1	RTroid2Pre	1.20	3.59	10.74	0.30	2.94	6
NVS_AdME_hCYP3A4	CYP3A4	RTroid3Neo	1.01	3.57	12.68	0.30	3.12	4
CLM_Hepat_Steatosis_lhr	NAGene	RTestes2Pre	1.03	3.56	12.32	0.30	3.25	4
CLM_p53Act_72hr	TF53	RTroid2Pre	1.26	3.49	9.66	0.30	2.94	7
CLZD_ABCG2_48	ABCG2	MLiver2Pre	1.34	3.46	8.96	0.33	2.65	14
BSK_SM3C_TVirobomodulin_up	THBD	RLiver2Pre	1.46	3.42	8.02	0.30	2.65	12
NCGC_AR_Antagonist	AR	MLiver2Pre	1.32	3.26	8.06	0.33	2.65	15
CLM_Hepat_NuclearSize_24hr	NAGene	RLiver2Pre	1.30	3.26	8.15	0.30	2.65	10
BSK_SM3C_MIG_down	CXCL9	RTroid2Pre	1.09	3.21	9.42	0.30	2.94	6
CLM_CellLoss_lhr	Cytotox	RTroid3Neo	1.13	3.21	9.10	0.30	3.12	6
CLM_OxidativeStress_lhr	H2AFX	RTroid2Pre	1.17	3.18	8.68	0.30	2.94	7
NVS_AdME_fCYP2A2	Cyp2a2	MLiver2Pre	1.21	3.17	8.29	0.33	2.65	13
BSK_hDFCGF_IP10_down	CXCL10	RTroid2Pre	1.52	3.05	6.13	0.30	2.94	23
NVS_AdME_hCYP2J2	CYP2J2	MLiver2Pre	1.20	3.00	7.50	0.33	2.65	14
NVS_AdME_hCYP1A1	CYP1A1	MLiver2Pre	1.09	2.89	7.65	0.33	2.65	12
BSK_BE3C_MIG_down	CXCL9	MLiver2Pre	1.09	2.89	7.65	0.33	2.65	12
BSK_BE3C_uPAR_down	PLAUR	MLiver2Pre	1.09	2.89	7.65	0.33	2.65	12

(continued)

Table 8.4 (continued)

Assay	GeneSymbol	EndPoint	LCI	OR	UCI	PermLCI	PermUCL	TruePositives
CLZD_ABCB1_24	ABCB1	MLiver2Pre	1.20	2.88	6.90	0.33	2.65	15
BSK_BE3C_MMPI_up	MMP1	RLiver2Pre	1.27	2.83	6.27	0.30	2.65	13
CLM_p53Act_72hr	TP53	RLiver2Pre	1.03	2.77	7.40	0.30	2.65	8
CLM_Hepat_DNADamage_24hr	Apoptosis	RLiver2Pre	1.12	2.73	6.62	0.30	2.65	10
BSK_SAg_IL8_down	IL8	RKidney2Pre	0.09	0.27	0.82	0.29	3.23	4
NVS_ADME_hCYP2C19	CYP2C19	MSpleen2Pre	0.07	0.25	0.88	0.32	3.80	3
BSK_3C_uPAR_down	PLAUR	RKidney2Pre	0.07	0.25	0.84	0.29	3.23	3

externally provided EPA-OPP binary classifications, Kleinstreuer et al. (2013) performed a Mann-Whitney test (also known as Mann-Whitney-Wilcoxon rank sum test or MWW) to assess the statistical significance of the association between the 33 cancer hazard scores and the binary cancer classifications. The objective of MWW is to test the null hypothesis that two populations have the same distribution of scores against the alternative hypothesis that one tends to have larger values than the other. The MWW test was originally devised for continuous variables (here we have binary and integer variables), but in practice has been applied to ordered categorical data as well. The authors reported a significant correlation (value not provided) with a significance level of 0.024 from their MWW test. Since this is well under the conventional 0.05 level, they concluded that their methodology is significantly predictive in the external validation test set: “We have demonstrated an approach to identify and test molecular pathways or processes that, when perturbed by a chemical, raise the likelihood that the chemical will be a carcinogen. . . . A simple scoring function built from these associated genes was significantly predictive of cancer hazard classifications for an external test set.”

We performed an independent analysis to check this conclusion. First, we noted that the MWW test is not applicable when there are ties in the ranks of the values, of which there are many. It cannot be relied upon to determine a correct significance value under these conditions. (It is not clear what implementation of MWW Kleinstreuer et al. used, but the R function “*wilcox.test*” provides a warning message indicating that an exact p-value cannot be computed when the data contains tied values. We therefore applied a different test—the Kendall tau-b correlation test—that adjusts for tied values. We used the standard implementation available in the Python *scipy* library (<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.kendalltau.html>). We first performed the Kendall tau-b test (with adjustments for ties) on the data as reported in Table 1 of Kleinstreuer et al. (2013), without corrections. This test provided a correlation coefficient of 0.30 and a barely significant p-value of 0.0453, larger than their reported p value of 0.024 but still slightly less than 0.05. (For comparison, we were able to derive 0/1 classification data for 176 of the 232 in the training data set using the EPA/OPP document, and found that test provided a correlation of 0.20 with a *p* value of 0.00188, that is, a highly significant correlation.)

We next performed the same test on the *corrected* data with methylene bis (thiocyanate) (MITC) removed from the external validation test set, etridiazole added, and the minor score correction made for oxyfluorfen. This test provided a correlation coefficient of 0.20 and a *p* value of 0.09 that is *not* significant at the conventional 5% significance level (although it is at the 10% significance level). If both MITC and etridiazole are included, the Kendall tau-b test yields a correlation coefficient of 0.25 and a *p* value of 0.09. Thus, Kleinstreuer et al.’s rejection of the null hypothesis of no statistically difference association between scores and binary external classifications at the conventional 5% significance level appears to depend crucially on the mistaken classifications of two chemicals and use of the MWW test without needed corrections for ties.

An Alternative Comparison of the Model-Predicted Cancer Hazard Scores to EPA'S Binary Cancer Classifications

That a statistical test for association between predictive scores and the binary classifications that they are used to predict is not significant at the 5% level does not mean that there is no valid or useful predictive relation between the scores and the classification. It only means that association may not be the best way to describe that relation. Figure 8.2 shows a direct plot of the fraction of chemicals classified by EPA as possible, probable, or likely rodent carcinogens against the predictive score (i.e., the count of the number of HTS assays considered positive, as previously discussed). Among 25 test set chemicals with scores less than 7, fewer than half (48%) are classified as rodent carcinogens. Among the remaining eight chemicals, having scores of 7 or more, 100% are classified as rodent carcinogens. Thus, the score provides useful predictive information, although the relation is not smoothly increasing, and hence ordinal association measures may not be the best way to characterize it.

In summary, rather than testing the significance of measures of ordinal association, an alternative, simple characterization of the predictive relation between HTS-derived scores and EPA classifications of the rodent carcinogenicity of these pesticides is that the pesticides with more than six “hits” (positive assays) are very likely to be classified as rodent carcinogens. This high-scoring fraction comprised 8/33, or just under 25%, of the test data set.

Predicting Cancer in Rodents, Not Humans

Following Kleinstreuer et al., we have focused so far on predicting cancer hazard in rodents, not in humans. In developing their prediction model, Kleinstreuer et al. extracted information from ToxRefDB for chemicals with entries corresponding to preneoplastic and carcinogenic pathologies in mouse and rat. For mice, the data extracted corresponded to effects in mice classified as “liver preneoplastic”, “liver neoplastic”, “lung preneoplastic”, and “spleen preneoplastic” and in rats “kidney preneoplastic”, “liver preneoplastic”, “liver neoplastic”, “testes preneoplastic”, “testes neoplastic”, “thyroid preneoplastic”, and “thyroid neoplastic” (Kleinstreuer et al. 2013, Supplemental Table 1). However, mouse liver tumorigenesis often lacks relevance to human cancer risk, can occur at high background rates, or may have dose-response thresholds and modes of action not relevant to human cancer (Maronpot 2009). Likewise, thyroid lesions in rats, both preneoplastic and neoplastic, often result from mechanisms of little to no concern for human health due to well-recognized species differences in sensitivity and mechanisms which alter thyroid hormone homeostasis in the rat but not in humans (Hill et al. 1998). Similarly, testicular preneoplastic and neoplastic lesions in rats are not be indicative of substantive cancer hazard to humans (Cook et al. 1999). Thus, the relevance of the predictive relation in Fig. 8.2 for predicting human carcinogenicity remains an open question.

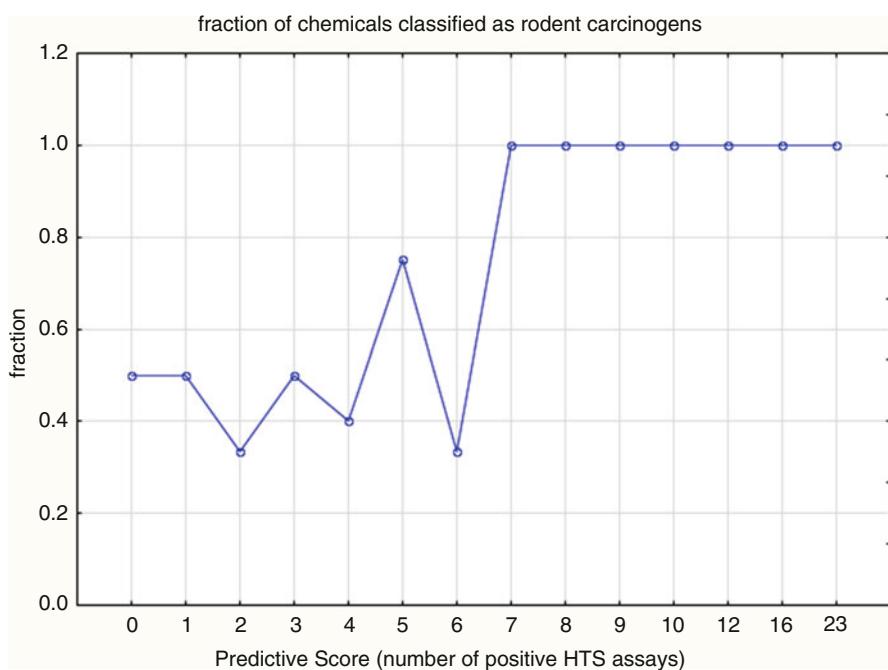


Fig. 8.2 The fraction of chemicals classified as rodent carcinogens is 100% if and only if the predictive score (count of positive HTS assays) exceeds 6

In determining potential cancer risk to humans, EPA seeks to integrate all of the relevant and reliable information, and to reach conclusions on potential hazards and risks to humans based on the knowledge of nature and incidence of the pathological responses, species specificity and sensitivity, and dose response. The predictive relation in Fig. 8.2 includes substances as positive even when EPA concludes that the level of evidence does not indicate a “likely” or “sufficient” or “probable” cancer risk to humans. Appendix A examines carcinogen classifications from the standpoint of possible human relevance. It concludes that, of the rodent carcinogenicity classifications shown by Kleinstreuer et al. (2013) for 154 chemicals, 104 might be changed if the hazard classification were to be made more relevant to humans.

Discussion and Conclusions

Although we were able to replicate most of the data and results reported by Kleinstreuer et al. (2013), their key conclusions on predictive power of HTS data proved to be very sensitive to the uncertain classifications of two chemicals and to the choice of statistical methods. The carcinogenic potential classifications for 2 of

60 chemicals differed from those in EPA/OPP published data (http://npic.orst.edu/chemicals_evaluated.pdf), and more recent data suggest reclassifying additional chemicals if the goal is to identify potential human carcinogens (Appendix A). Moreover, the Mann-Whitney-Wilcoxon rank sum test used was not appropriate for the data due to many tied ranks. Correcting these two chemical classifications and applying a test (Kendall tau-b) that correctly adjusts for ties, we found that the HTS-based cancer hazard scores no longer significantly predict *in vivo* cancer results. A change of classification to a single chemical can change the original study conclusion, which is therefore not robust. This outcome suggests a clear need for more robust predictions (e.g., based on application of current model ensemble machine learning methods) and highlights the potential value of using multiple subsets of training data to achieve predictive models and quantitative conclusions that are less sensitive to minor changes in chemical classifications and statistical methods.

A plausible reaction to the replication and reanalysis results presented here is that, even if the Kleinstreuer et al. results are not statistically significant at a conventional 0.05 significance level, the fact that they are statistically significant at the 0.10 level is still encouraging. In the reconstructed model we developed, we observed that chemicals with a score of 7 or more were all classified as rodent carcinogens (in the test set), but each chemical with a score of 0–6 had one or more chemicals that are not rodent carcinogens. Thus, the simple scoring system proposed by Kleinstreuer et al. appears to have genuine predictive power, although its predictions are strong only for scores of at least 7 (about 25% of the chemicals examined).

The generalizability of these results to other data sets (e.g., outside the EPA-OPP data sets used in this study) remains unknown, and the probable error rates for applications of the procedure to new chemicals have not been characterized. For developing models to predict human cancer hazard from HTS data and data extracted from ToxRefDB, or other similar databases of chemically-induced rodent histopathological responses, we posit that the overall weight of evidence determinations for human risks are key, and assay-endpoint combinations should not include rodent pathologies known to be of little or no relevance to humans.

The recent work by NCCT was presented as showing that “A simple scoring function built from these associated genes was significantly predictive of cancer hazard classifications for an external test set.” Figure 8.2 supports this description for the pesticides studied, especially the roughly 25% of them with score greater than 6. For the remaining chemicals, it is not clear that the scoring system can predict cancer hazard classifications well, consistent with the poor performance of previous systems in predicting rodent carcinogens in external validation studies (e.g., Valerio et al. 2010; Walmsley and Billinton 2011; Benigni and Zito 2004; Snyder et al. 2004). We do not interpret these limitations as undermining the exciting program of research described by Kleinstreuer et al. (2013), but as showing that more (straight-forward) work needs to continue to be done to allow the results and predictive performance of these models be rigorously documented, thoroughly evaluated, and compared to the results of previous approaches across a wide range of chemicals.

References

- Benigni R, Zito R (2004) The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat Res* 566(1):49–63
- Cook et al (1999) Rodent Leydig cell tumorigenesis: a review of the physiology, pathology, mechanisms, and relevance to humans. *Crit Rev Toxicol* 29:169–261
- EPA (2011) <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=238403>
- Guyton KZ, Kyle AD, Aubrecht J, Cogliano VJ, Eastmond DA, Jackson M, Keshava N, Sandy MS, Sonawane B, Zhang L, Waters MD, Smith MT (2009) Improving prediction of chemical carcinogenicity by considering multiple mechanisms and applying toxicogenomic approaches. *Mutat Res* 681(2–3):230–240
- Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* 144(5):646–674
- Hill et al (1998) Assessment of thyroid follicular cell tumors. EPA/630/R-97/002
- March Kirkland D, Aardema M, Henderson L, Müller L (2005) Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mutat Res* 584(1–2):1–256. Erratum in: *Mutat Res.* 2005 Dec 7;588(1):70
- Maronpot RR (2009) Biological basis to differential susceptibility to hepatocarcinogenesis among mouse strains. *Toxicol Pathol* 22:11–33
- Kleinsteuer NC, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Paul KB, Reif DM, Crofton KM, Hamilton K, Hunter R, Shah I, Judson RS (2013) In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. *Toxicol Sci* 131(1):40–55
- Knight A, Bailey J, Balcombe J (2006) Animal carcinogenicity studies: 1. Poor human predictivity. *Altern Lab Anim* 34(1):19–27
- Patlewicz G, Rodford R, Walker JD (2003) Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity. *Environ Toxicol Chem* 22(8):1885–1893
- Snyder RD, Pearl GS, Mandakas G, Choy WN, Goodsaid F, Rosenblum IY (2004) Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ Mol Mutagen* 43(3):143–158. Erratum in: *Environ Mol Mutagen.* 2006 Apr;47(3):225
- Valerio LG Jr, Arvidson KB, Chanderban RF, Contrera JF (2007) Prediction of rodent carcinogenic potential of naturally occurring chemicals in the human diet using high-throughput QSAR predictive modeling. *Toxicol Appl Pharmacol* 222(1):1–16
- Valerio LG Jr, Arvidson KB, Busta E, Minnier BL, Kruhlak NL, Benz RD (2010) Testing computational toxicology models with phytochemicals. *Mol Nutr Food Res* 54(2):186–194. <https://doi.org/10.1002/mnfr.200900259>
- Walmsley RM, Billinton N (2011) How accurate is in vitro prediction of carcinogenicity? *Br J Pharmacol* 162(6):1250–1258

Chapter 9

Mechanistic Causality: Biological Mechanisms of Dose-Response Thresholds for Inflammation-Mediated Diseases Caused by Asbestos Fibers and Mineral Particles



Introduction

As explained in Chap. 2, mechanistic causal models of how effects propagate through a system typically require more detailed information to build and validate than other forms of causal analysis, including predictive and attributive causal modeling. Substantial applied and computational mathematical research, modeling, and algorithm development is sometimes needed to describe with useful accuracy how a system evolves over time. On the other hand, mathematical analysis can also reveal robust qualitative properties of a system's dynamic response to inputs. For example, many complex feedback control networks exhibit the qualitative property of *bistability*, in which a sufficiently long and intense stimulus or exogenous input causes the system to shift from its normal state to a new one with different properties that then becomes the new stable state of the system. Such stimulus-driven switches in behaviors occur frequently in biological regulatory networks and in other (e.g., socioeconomic) systems with positive feedback loops. This chapter considers the implications of recent advances in molecular biological understanding of the causal mechanisms of inflammation-mediated diseases for quantitative dose-response modeling. It focuses on the dynamic behavior of the NLRP3 (nucleotide-binding oligomerization domain-, leucine-rich repeat- and pyrin domain-containing) inflammasome, a signaling complex that is activated in response to sufficiently large exposures to potentially injurious agents including *Staphylococcus aureus* or *Listeria monocytogenes* bacteria, influenza and other viruses, radiation, asbestos fibers, and respirable crystalline silica (RCS) and that has been implicated in a host of inflammation-mediated diseases including asbestosis, fibrosis, mesothelioma, lung cancer, heart disease, gout, arthritis, and diabetes. Given this large and diverse array of agents and diseases for which NLRP3 provides a key to pathological responses, we will focus on how mineral particles and fibers such as asbestos can

activate the NLRP3 inflammasome and on the consequences for the shape of the dose-response relationship for inflammation-mediated responses to exposure.

Sufficiently high and prolonged inhalation exposures to some respirable elongated mineral particles (REMPs), notably including amphibole asbestos fibers, can increase risk of inflammation-mediated diseases including malignant mesothelioma, pleural diseases, fibrosis, and lung cancer. Although the molecular mechanisms of pathogenesis are still being elucidated, it is now clear that chronic inflammation sustained by ongoing activation of the NLRP3 inflammasome plays a crucial causal role, enabling immune cells to produce the potent proinflammatory cytokines IL-1 β and IL-18. This insight, which has been developed in detail largely over the past decade, harmonizes with previous understanding that had identified upregulation of reactive oxygen species (ROS) as playing a central role in creating and maintaining a pro-inflammatory environment in these diseases and others, such as COPD. It is now understood that ROS (in particular, mitochondrial ROS) contributes to NLRP3 activation via a well-elucidated mechanism involving oxidation of reduced thioredoxin and association of thioredoxin-interacting protein with NLRP3, although the precise roles of ROS in the two-step priming and activation of NLRP3 are still being clarified. Lysosomal destabilization, efflux of cytosolic potassium ions and influx of calcium ions, signals from damaged mitochondria, both translational and post-translational controls, and prion-like polymerization have increasingly clear roles in regulating NLRP3 activation.

As the molecular biology of inflammation-mediated responses to REMPs exposure becomes clearer, a practical question looms: *What do these mechanisms imply for the shape of the dose-response function relating exposure concentrations and durations for EMPs of different shapes, sizes, and surface chemistries to the risk of pathological responses?* For example, does increasing understanding of how REMPs affect the NLRP3 inflammasome have any clear implications for the existence of dose-response thresholds or threshold-like nonlinearities? How much knowledge must be accumulated before useful answers can be given to such questions? We propose that the partial understanding of NLRP3-mediated REMPs effects available today is already sufficient to show that threshold-like dose-response nonlinearities should be expected. Biomathematical analysis of regulatory mechanisms and networks provides general conditions that lead to such thresholds; these include (a) Cooperativity in the assembly of supramolecular signaling complexes such as the inflammasome and apoptosome, leading to a characteristic all-or-nothing response; (b) Positive feedback loops in regulatory networks, leading to bistability in the network response; (c) Overwhelming or suppression of defensive barriers for maintaining homeostasis, such as IL-1 β -mediated suppression of antioxidant defenses; and (d) Damage thresholds below which responses are controlled and above which they are not, as in lysosome destabilization-induced activation of NLRP3. Each of these general classes of mechanisms for generating exposure-response thresholds is already known to hold for NLRP3 activation in response to stimuli such as REMPs exposures. Moreover, some of them (such as bistability induced by positive feedback loops) are robust qualitative features of dose-response, insensitive to further details of the underlying biochemical mechanisms. It is

therefore timely to start considering the implications of these advances in biological understanding for human health risk assessment with dose-response thresholds.

Understanding of the toxicological mechanisms of health risks caused by inhaling mineral particles, including respirable crystalline silica (RCS), elongated mineral particles (EMPs), and asbestos fibers, has been greatly advanced over the past decade by the discovery and gradual elucidation of the functioning of inflammasomes that coordinate inflammatory responses to inhalation exposures. It is plausible that the NLRP3 inflammasome plays a decisive role in initiation and progression of various inflammation-mediated diseases caused by such inhalation exposures, including silicosis, fibrosis, lung cancer, and malignant mesothelioma (Sayan and Mossman 2016), as well as in diseases of other organs and organ systems, such as gout, acute myocardial infarction and inflammatory bowel disease (Veltman et al. 2017). However, these substantial advances in biological insights have yet to be matched by corresponding advances in understanding their implications for quantitative dose-response modeling, health risk assessment, and uncertainty characterization. This chapter examines the biomathematical implications for dose-response relations and health risks of what is now known about exposure-induced NLRP3 assembly and activation dynamics and subsequent disruptions of normal homeostasis resulting in chronic inflammation and increased risks of inflammation-mediated diseases. Crucial questions for dose-response modeling include the following:

1. Is there an exposure concentration threshold below which adverse responses of interest do not occur?
2. For any exposure concentration, is there an exposure duration threshold before which adverse responses of interest do not occur?
3. If such thresholds exist, how large are they, and on what factors do they depend?
4. What is the shape of the concentration-duration-risk relationship for exposures above the exposure concentrations and duration thresholds (if any)? How do physiochemical properties of inhaled materials and pharmacokinetic and biochemical parameters for exposed individuals affect the answer?
5. How sure can we be about the answers to the preceding questions, and what information would most help to reduce remaining uncertainties?

To address these questions, the following sections examine biomathematical implications for dose-response relationships of currently known biochemistry and toxicology of NLRP3-mediated inflammation and disease processes.

Biological Background: NLRP3 Inflammasome Responses to Mineral Particles

Excellent reviews of the biology of the NLRP3 inflammasome and its roles in various inflammation- and immune system-mediated diseases are already available. Sayan and Mossman (2016) discuss NLRP3 inflammasome priming, activation, and

signaling specifically for mineral particles. Figure 9.1 depicts a common contemporary view of the causal cascades that can trigger priming, assembly, and activation of this inflammasome in various cells, including monocytes, macrophages, dendritic cells, and human mesothelial cells (Thompson et al. 2017) in response to a variety of environmental triggers. In brief, exposures producing pathogen-associated molecular patterns (PAMPs) and danger-associated molecular patterns (DAMPs) trigger a signaling cascade (“Signal 1”) via phosphorylation of Toll-like receptors (TLRs) that stimulate NF- κ B-mediated upregulation of nuclear transcription of the genes for NLRP3, proIL-1 β , and proIL-18. These genes are translated into corresponding proteins that are released to the cytoplasm as inactive building blocks, thus “priming” the inflammasome by making these components available in the cytosol for assembly and activation. Priming also deubiquitinates the NLRP3 protein (via pathways that can involve mitochondrial ROS or ATP signaling), allowing it to be activated.

In macrophages and dendritic cells, the inflammasome is assembled and activated upon receipt of a second signal, denoted by “Signal 2” in Fig. 9.1, that triggers oligomerization of its inactive NLRP3, apoptosis-associated speck-like protein (ASC), which is linearly ubiquitinated and phosphorylated in the process, and procaspase-1 components (Guo et al. 2015; Bednash and Mallampalli 2016).

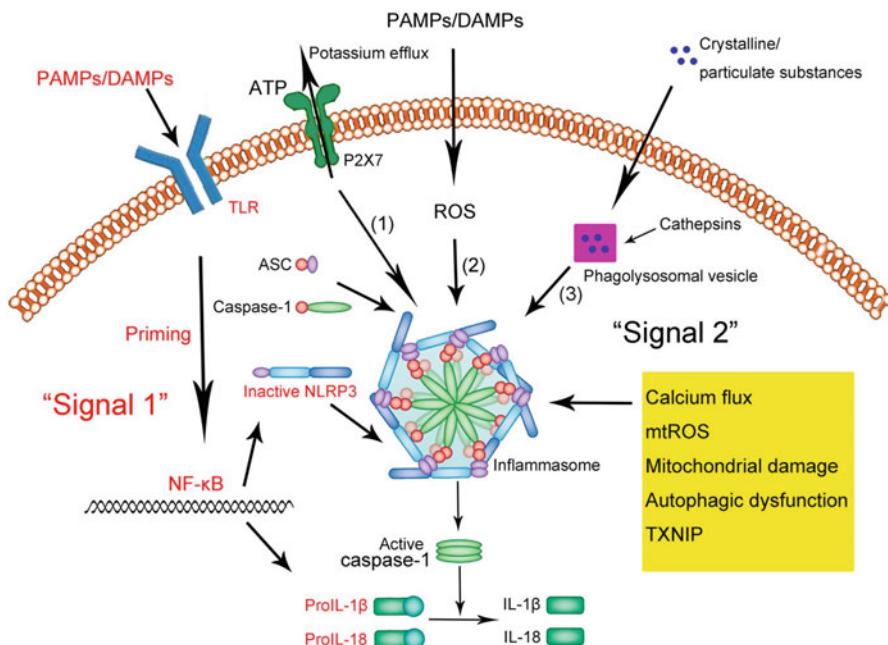


Fig. 9.1 NLRP3 inflammasomes within a cell’s cytosol use active caspase-1 to cleave ProIL-1 β and ProIL-18, forming mature inflammatory cytokines IL-1 β and IL-18, in response to exposures to mineral particles and other stimuli. Source: Shao et al. (2015). This picture is provided under the terms of the Creative Commons Attribution License (CC BY)

(In monocytes, no second signal is needed (*ibid.*).) These components (the “mers” or units in the oligomer) come together via oligomerization to form a fully assembled and active inflammasome. The Signal 2 that triggers this process typically involves calcium ion fluxes and K⁺ efflux (via a P2X7-dependent pore); PAMP- and DAMP-associated production of reactive oxygen species (ROS); mitochondrial damage and production of mitochondrial ROS (mtROS); or, most importantly for crystalline silica and asbestos fiber-induced diseases, frustrated phagocytosis of mineral particles leading to lysosomal membrane rupture and release of cathepsin B and other pro-inflammatory contents (Guo et al. 2015; Sayan and Mossman 2016). In a positive feedback loop, cytoplasmic ROS and mitochondrial ROS stimulate translocation of thioredoxin-interacting protein (TXNIP) from the nucleus into the cytoplasm and mitochondria, where it binds to and inhibits the activity of the antioxidant thioredoxins TRX1 and TRX2, respectively, thus further elevating of cytoplasmic and mitochondrial ROS (Harjith et al. 2014).

Elevated ROS and molecules from ruptured lysosomes activate the primed NLRP3 inflammasomes in the cytoplasm, causing them to cleave the protective tails from ProIL-1 β and ProIL-18 (using active caspase-1) to form mature inflammatory cytokines IL-1 β and IL-18. These potent pro-inflammatory cytokines, in turn, act as signals stimulating and coordinating other inflammatory events in other cells, eventually leading to pyroptosis (i.e., inflammatory cell death) of the host cell, recruitment of activated macrophages and neutrophils, and chronic unresolved inflammation in the lung or other target tissues. NLRP3 protein nucleates growth of prion-like filaments of ASC, from which pro-caspase-1 filaments subsequently grow and become activated via autoproteolysis, generating active caspase-1 (Guo et al. 2015). Repeated and widespread cycles of epithelial cell injury and tissue damage, partial repair (eventually leading to fibrosis and scarring in the alveolar epithelium), and stimulated cell division and proliferation of progenitor cells can increase the risks of inflammation-mediated diseases and pathologies, including fibrosis, silicosis, asbestosis, lung cancer, and malignant mesothelioma.

We will refine this basic description of NLRP3 inflammasome biology later to emphasize the roles of several other positive feedback loops, but the version just described and depicted in Fig. 9.1 suffices to understand the main steps where thresholds might arise in dose-response modeling. They are priming, assembly, activation, and signaling by the activated inflammasomes within and between cells.

Thresholds in NLRP3 Priming: Receptor-Mediated Signal Transduction and Critical Mass of NLRP3 Protein Required for Activation

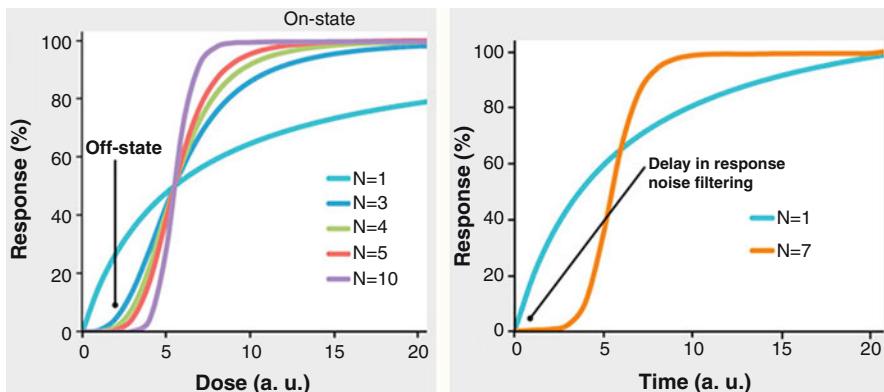
Receptor-mediated responses such as priming of the inflammasome typically exhibit thresholds or threshold-like nonlinearities if a sufficiently large (threshold) fraction of receptors must be simultaneously bound by a ligand to trigger a signaling cascade

(Andersen et al. 2014). Lower ligand concentrations almost certainly fail to trigger a response, with probability near 1; concentrations that are higher trigger a response with probability close to 1; and the interval of concentrations that gives intermediate probabilities of triggering a response tends to be very short, corresponding to a “sharp transition” threshold in response probability (Cox 2006; Wu 2013). For example, for Toll-like receptor 4 (TLR4) ligand-activated signaling, Gottschalk et al. (2016) identify distinct thresholds for NF- κ B and MAPK signaling activation in both mouse and human macrophages. Such thresholds protect the cell against responding to low levels of exposures (e.g., from endogenous bacteria) while allowing high levels to trigger inflammatory signaling.

The NLRP3 inflammasome cannot be assembled and activated until a sufficient quantity of the NLRP3 protein has accumulated in a cell’s cytosol (Bednash and Mallampalli 2016). Accumulation of NLRP3 protein depends on a dynamic balance between its production, which is triggered by NF- κ B signaling to the nucleus (Fig. 9.1), and its inactivation or destruction. NLRP3 protein can be temporarily inactivated by ubiquitination and then activated relatively rapidly via deubiquitinating enzymes (Py et al. 2013). Fully assembled NLRP3 inflammasomes are also removed by autophagy in autophagosomes formed in response to, and partly co-localized with, the inflammasomes (Shi et al. 2012). As NLRP3 protein accumulates, the threshold for signal 2 to trigger assembly and activation of the inflammasome decreases (Bednash and Mallampalli 2016). Once a tipping point is reached in which the rate of production of active (deubiquitinated) protein has exceeded the rate of inactivation and removal for long enough for active NLRP3 protein levels to accumulate to a critical level, even low levels of signal 2 will result in NLRP3 inflammasome activation. Conversely, if the signals that stimulate NLRP3 protein production and deubiquitination are insufficiently strong and protracted to enable its accumulation to overwhelm the inactivation and degradation processes, then assembly and activation do not occur: priming fails to lead to activated NLRP3 inflammasomes. Thus, effective priming does not occur if exposure concentration and duration do not generate a large enough supply of deubiquitinated NLRP3 protein so that assembly and activation of NLRP3 inflammasomes can proceed (Bednash and Mallampalli 2016). Exposures cannot cause NLRP3 inflammasome-mediated responses unless they are sufficiently high and sustained for such effective priming to be completed.

Thresholds for NLRP3 Assembly: Cooperativity in Oligomerization Kinetics

If exposure is sufficient for priming to occur and if further exposure generates signal 2 while the primed state lasts, then NLRP3 inflammasomes will begin to assemble via energetically favorable oligomerization. Several types of signaling complexes with multiple domains assembled into functional units (oligomers), including the



Source: Wu (2013) (Reproduced with permission from Elsevier)

Fig. 9.2 Cooperativity in oligomerization processes leads to sharp transitions in responses. N = Hill coefficient measuring cooperativity; a.u. = arbitrary units. Source: Wu (2013) (Reproduced with permission from Elsevier)

apoptosome and the inflammasome, are formed via oligomerization exhibiting kinetic cooperativity: attachment of further units becomes progressively easier following initial nucleation and increasing availability of binding sites for additional units as the oligomerized array expands (Bagci et al. 2006; Wu 2013). Wu (2013) emphasizes the importance of replacing a traditional view of signal transduction and “signalsome” assembly as a cascade of events (Fig. 9.1) with a view in which large spatial arrays of oligomer components undergo these cascades side by side, in parallel, with the parallel cascades facilitating each other, leading to sharp transitions in the responses of cells to concentrations of signals that stimulate assembly of the oligomers. Figure 9.2 shows the mathematical implications (from the Hill equation in biochemistry or the equivalent Langmuir adsorption isotherm in surface chemistry) of increasing cooperativity in attachment of units to binding sites during parallel assembly of oligomers.

In the left panel, the “Dose” on the horizontal axis refers to concentration of a ligand (e.g., deubiquitinated NLRP3 protein) in arbitrary units (a.u.), scaled so that a dose of 5 is defined as the level that elicits 50% of the maximum response. Thus, all dose-response curves must pass through this point, which is fixed by definition. The “Response” on the vertical axis shows the fraction of the maximum response (e.g., maximum production of fully assembled oligomers) achieved. As cooperativity in oligomerization increases (indicated by increasing values of the Hill coefficient, N), the dose-response curves become steeper and increasingly threshold-like and the minimum concentration needed to elicit a significant positive response increases while the minimum dose needed to achieve maximum response decreases. The right panel shows that higher cooperativity also induces an increased time delay before the response departs significantly from zero, as initialization (successful nucleation) of the oligomerization process takes longer. These patterns are predicted to hold for

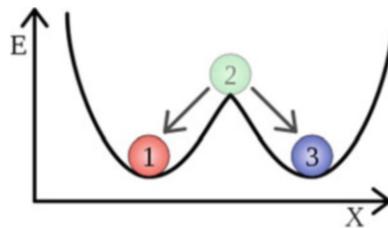
assembly of signaling complexes—higher-order assemblies of oligomers consisting of intracellular adapters, signaling enzymes and their substrates—that can be well approximated by the Hill equation; the NLRP3 inflammasome is only one of many examples (Qiao and Wu 2015). The threshold-like nonlinearities in concentration-response curves (left side of Fig. 9.2) and in the time needed to respond significantly (right side of Fig. 9.2) protect macrophages and other cells against continually responding to low-level stimuli.

Thresholds in NLRP3 Activation: Lysosome Disruption and ROS

Signal 2 is generated by several inter-linked processes involving mitochondrial damage, release of mitochondrial ROS (mtROS), increase in lysosomal membrane permeabilization (LMP), disruption of the lysosome, and release of lysosomal hydrolases, including cathepsins, to the cytosol (Fig. 9.1); these events can trigger a cell death pathway (Boya and Kroemer 2008; Repnik et al. 2014). Moreover, human monocytes sequester iron and use iron ions to activate the NLRP3 inflammasome (Nakamura et al. 2016). Studies of carbon nanotubes suggest that fiber geometry is important. Stiff fibers beyond a critical length poke against the inner leaflet of the soft lysosomal membrane, causing lysosomal permeabilization (Zhu et al. 2016). Despite this complexity, it is now clear that accumulation of protonated lysosomotropic agents above a concentration threshold triggers detergent-like disruption of the lysosomal membrane and that accumulation of iron in the lysosome—possibly accelerated by phagocytosis of mineral particles with iron ions available to participate in Fenton reactions—catalyzes ROS-induced disruption of its membrane (Boya and Kroemer 2008; Schilling 2016). Rupture of the lysosome membrane activates a MAPK signaling pathway that, in turn, contributes to activation of the NLRP3 inflammasome via oligomerization of its ASC component (Okada et al. 2014). Rupture of a membrane triggered by accumulation of destabilizing contents past a critical level is inherently a threshold-like response. In addition, LMP, ROS, and the NLRP3 inflammasome participate in several positive feedback loops. As discussed next, such loops can create threshold-like responses for the set of variables involved, even if none of them by itself has such a threshold-like response.

Thresholds in NLRP3 Signaling: Positive Feedback Loops and Bistability

A bistable dynamical system is one with two stable equilibrium states. Figure 9.3 illustrates bistability visually: the position of a ball, indicated by X , has stable local equilibria at both position 1 and position 3. Pushing the ball rightward from position



Source: <https://en.wikipedia.org/wiki/Bistability>

Fig. 9.3 Concept of a bistable system. A ball can be in stable equilibrium (a local minimum of the energy, E) in positions 1 or 3. Position 2 is an unstable equilibrium. For the NLRP3 inflammasome, the “position” variable X can be reinterpreted as ROS. Source: <https://en.wikipedia.org/wiki/Bistability>

1 will eventually cause it to cross a threshold (position 2), leaving the basin of attraction for position 1 (to which it would otherwise have returned when the force was removed) and entering the basin of attraction for position 3, to which it will now spontaneously move in the absence of further exposure to the driving force. This is a visual metaphor for switch-like transition thresholds in bistable systems: sufficient exposure to an outside force that exogenously increases a variable eventually drives the system past a threshold and causes it to enter a new stable equilibrium with permanently increased levels of the variable.

In biological systems and chemical reaction networks, bistability commonly arises from positive feedback loops among variables, and switch-like behavior generated by such loops is a common motif found in a wide variety of regulatory networks (e.g., Siegal-Gaskins et al. 2011; Chakravarty and Barik 2017). That a system can have multiple equilibrium states is easily illustrated. Consider a system with two variables, x and y , with each increasing the level of the other, as follows:

$$\frac{dy}{dt} = x - y$$

$$\frac{dx}{dt} = y - 2xy$$

This pair of coupled ordinary differential equations (ODEs) states that y is formed at a rate proportional to the level of x and is removed at a rate proportional to its own level, where both rates are measured in units of y per unit time formed or removed. Likewise, x is formed at a rate proportional to y and is removed at a rate proportional to the product of x and y . In steady-state equilibrium, if one exists, $\frac{dy}{dt} = \frac{dx}{dt} = 0$, and so $y = x$ (from the first equation) and the second equation then implies $y = 2y^2$. The system has two possible equilibria: $x = y = 0$ and $x = y = 0.5$. Bistable systems have the additional feature that each of the two equilibrium states is locally stable, meaning that the system will return to it after small perturbations.

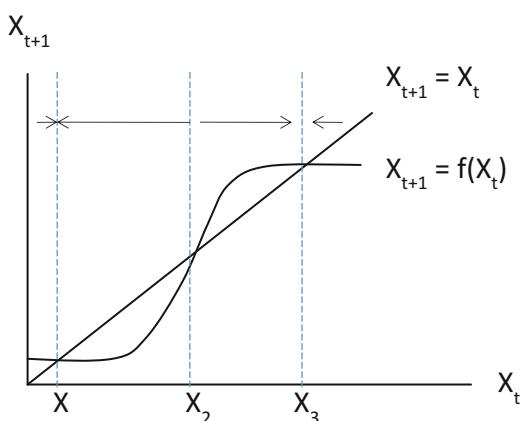
How bistability emerges from positive feedback loops is easy to see for a single feedback loop; Fig. 9.4 sketches the main idea. Write the loop as a chain $X \rightarrow Y \rightarrow \dots \rightarrow Z \rightarrow X$ with the same first and last variable. Now, imagine fixing the value of the first variable at some value and letting each variable in turn adjust to the new equilibrium value determined by its predecessor’s value (assuming that such an equilibrium value exists). A new value of X is determined by this equilibration

process at the right end of the chain. From this perspective, each value of X supplied at the start of the chain determines a new value at the end of it. This relation can be described as $X_{t+1} = f(X_t)$ where t indexes iterations and f is a function mapping the starting value of X to its ending value by traversing the loop once, as just described. Equilibrium values of X are fixed points of this function, i.e., values such that $X = f(X)$; starting from such a point, the iterative mapping leaves the value of unchanged. In biological systems, it is common for f to have a sigmoid shape, as illustrated in Fig. 9.4. In this diagram, the horizontal axis consists of possible starting values for X ; the vertical axis represents the next value of X , determined from the starting value by traversing the feedback loop once; the sigmoid function represents the feedback loop mapping f , and the line $X_{t+1} = X_t$ is the locus of possible equilibrium points. The left end of the function f is usually flat because one or more elements in the loop, possibly being stabilized by their own negative feedback networks, respond very little to small perturbations in low levels of X .

The right end of f is flat because one or more variables is saturated: it has reached a maximum possible value, and the loop as a whole cannot be driven to higher levels of all its variables once any of them is saturated. Between these two regions, in which the next value of X is insensitive to the current value, is a region where the function f is increasing. Between X_1 and X_2 , f lies below the equilibrium line, meaning that the next value of X is less than its starting value, and the system moves leftward, back toward the stable equilibrium X_1 . (The subscripts on X_1 , X_2 and X_3 denote specific values, not successive iterations.) Between X_2 and X_3 , f lies above the equilibrium line, and the system moves rightward, toward the stable equilibrium X_3 . The tipping point, or threshold, separating the basins of attraction for these two equilibria is the unstable equilibrium point X_2 .

Interpreatively, the bistability model in Fig. 9.4 shows that there is a normal healthy equilibrium state, corresponding to X_1 , in which variables in the feedback loop are at their unperturbed levels. This is a locally stable state, maintained by homeostatic mechanisms, and hence is restored following small perturbations, such as after an exposure that is insufficiently high and prolonged to push the system past

Fig. 9.4 A bistable system has two stable equilibria (values X_1 and X_3) separated by an unstable one (X_2)



X_2 . However, exposures large enough to increase X past X_2 induce a transition to a new, pathological state, X_3 , which is also stable. For example, if X represents ROS and is part of one or more positive feedback loops, then one or more high-ROS disease states may be reachable via sufficiently large exposure-related increases in ROS that will then remain stable, corresponding to a chronic inflammatory state. Unresolved chronic inflammation with persistently elevated levels of ROS, inflammasome activation, macrophage and neutrophil infiltration and activation, and proinflammatory cytokines can thus be viewed as a stable pathological equilibrium, analogous to X_3 in Fig. 9.4 (Cox 2011). We shall add inflammasomes to this picture momentarily.

Many aspects of inflammation, including EMP-induced inflammation, are now known to involve positive feedback loops. Bistability has been noted for mitochondrial ROS production (Pereira et al. 2016); IL-1 β upregulation in response to TNF α in a model of inflammatory responses to influenza (Jin et al. 2014); cytokine network signaling and immune cell responses in cancer (Li and Levine 2017); and bacterial infections (Malka et al. 2010). Both asbestos and erionite fibers prime and activate the NLRP3 inflammasome in human mesothelial cells via an autocrine feedback loop that passes through the IL-1R receptor (Hillegass et al. 2013). Table 9.1 summarizes multiple positive feedback loops identified for NLRP3 inflammasome activation, ROS generation, and signaling. In these diagrams, an arrow between two quantities means that an increase in the one at the arrow's tail increases the one at its head. Each chain starts and ends with the same variable, indicating a positive feedback loop.

Loops 1 and 2 are discussed in more detail by Boya and Kroemer (2008). Harijith et al. (2014) provide details of loops 3–5 and Cox (2011) discusses the neutrophil and macrophage loops summarized in loop 6.

The individual feedback loops in Table 9.1 clearly intersect with each other, forming a network of overlapping positive feedback loops. A fuller picture of the

Table 9.1 Positive feedback loops for production of reactive oxygen species (ROS) and activation of the NLRP3 inflammasome

Loop	Description of positive feedback loop
1	ROS → LMP → lysosomal enzymes released → phospholipase A2 activated → mitochondrial outer membrane permeabilization (MOMP) → ROS
2	LMP → lysosomal enzymes released → phospholipase A2 activated → LMP
3	ROS → thioredoxin-interacting protein (TXNIP) translocated from nucleus to cytoplasm and mitochondria → antioxidant thioredoxins inhibited → ROS (cytoplasmic and mitochondrial)
4	ROS → NLRP3 activation → IL-1 β , IL-18 → inflammatory cytokines → cell damage → pyroptosis → release of cell contents → ROS
5	ROS → NLRP3 activation → DNA damage → pyroptosis → release of cell contents → ROS
6	ROS → decrease in antioxidants → secretion of matrix metalloproteinase 12 (MMP-12) and neutrophil elastase (NE) → neutrophil and macrophage recruitment and activation → NE, ROS

cytokine signaling network and immune cell population responses (such as recruitment and activation of macrophages and neutrophils, cross-talk with T cells, activation and eventual depletion or overwhelming of additional antioxidant defenses, and so forth) would be far more complex than these few feedback loops and would include negative feedback loops that attempt to maintain healthy homeostasis (Cox 2011). However, the existence and prominence of positive feedback loops is consonant with the existence of observed bistable (or multistable) dose-response relationships in which sustained high exposures eventually overcome stabilizing loops or reservoirs (e.g., reducing active antioxidant production and depleting antioxidant pools) and shift the immune network from a normal low-ROS healthy state to a pathological high-ROS chronic inflammation state, accompanied by increases in the variables in loops 1–6 that increase, and are increased by, ROS. This includes priming and activation of NLRP3 inflammasomes in loops 4 and 5.

From Cells to Tissues: Percolation Thresholds for Spread of Inflammation

Activation of NLRP3 inflammasomes, secretion of inflammatory cytokines, induction of a high-ROS state, and apoptosis or pyroptosis are all normal parts of acute inflammation, e.g., in response to bacterial infections. They do not normally cause the previously mentioned diseases associated with chronic inflammation. What determines whether inflammation resolves itself, subsiding back into homeostasis, or progresses to pathological chronic inflammation and increased risks of diseases? Bistability is likely an important part of the answer. As suggested by the ball metaphor in Fig. 9.3, exposure must be large enough for long enough to move the system from its normal homeostatic basin of attraction past the unstable equilibrium threshold and into the pathological basin of attraction in order to create chronic inflammation as a new equilibrium. The positive feedback loops in Table 9.1 add detail to this metaphor by specifying variables such as ROS, events such as NLRP3 activation, processes such as pyroptosis, and conditions such as LMP that are involved in the shift between basins of attraction. However, they are still rather abstract, leaving aside details such as which specific cell populations (e.g., monocytes, macrophages, dendritic cells, alveolar epithelial cells, mesothelial cells, and so forth) are involved in each change, and how strongly. Further information about specific exposures and responses can be added within the framework when it is available. For example, the recent discovery that crocidolite asbestos fibers oxidize Thioredoxin-1 (Trx1), releasing TXN1 and activating inflammasomes in human mesothelial cells (Thompson et al. 2014) shows that feedback loop 3 in Table 9.1, and other loops that it activates, are relevant for crocidolite asbestos exposures. However, the feedback loops and mechanisms of NLRP3 priming, assembly, activation, and signaling discussed so far have focused primarily on events within individual cells and their compartments, especially the nucleus, cytosol,

mitochondria, and lysosome. It is useful to also consider how inflammation spreads between cells.

Intracellular spread of inflammation and ROS activation is facilitated by the release of NLRP3 inflammasomes into extracellular spaces following pyroptosis. The released inflammasomes continue to broadcast inflammatory cytokines and act as danger signals to neighboring cells, which may then mount their own inflammasome-mediated defense if the signaling is strong enough and lasts long enough (Baroja-Mazo et al. 2014). Inflammation also spreads among neighboring cells via prion-like propagation mediated by release and subsequent uptake of ASC specks (Franklin et al. 2014). In extracellular space, these specks continue to promote IL-1 β maturation, which increases inflammatory cytokine signaling to nearby cells. In addition, when phagocytized by macrophages, the previously released ASC specks induce lysosomal damage, nucleation of ASC oligomerization and fiber growth, and IL-1 β activation in the macrophages, initiating a new round of inflammatory responses in these new cells (Franklin et al. 2014). For the population of cells, this leads to a further positive feedback loop:

pyroptosis of inflamed macrophage → NLRP3 inflammasome and ASC speck release into extracellular space → NLRP3 inflammasome and ASC signaling, IL-1 β activation in extracellular space → uptake of ASC speck by new macrophage → inflammation of new macrophage → pyroptosis of inflamed macrophage

This loop has an important spatial component, as it takes place within a volume of tissue where cell populations, recruited immune cells, and NLRP3 inflammasomes and ASC specks in extracellular space interact. Moreover, there is considerable heterogeneity in the responses of cells within the same local volume of tissue. Being surrounded by other cells undergoing inflammation and pyroptosis increases the probability that a cell will also succumb to them, but not all cells respond the same, suggesting that the spread of inflammation can best be regarded as a spatial stochastic process, somewhat analogous to the spread of a forest fire through a population of heterogeneous trees. For such a spatial stochastic process, the question of how widely an initially localized inflammation will spread through the affected tissue (analogous to how far a forest fire will spread if an edge or area is initially ignited) can be addressed using stochastic percolation models (Guisoni et al. 2011; Squires et al. 2013). Such models consider the conditional probability that a cell will become inflamed (or, in our context, undergo pyroptosis) if its neighbors do. A common finding is that there is a *percolation threshold* for this conditional probability: above the percolation threshold, the inflammation spreads throughout the available area or volume where these strong dependencies hold, and below it, the inflammation is self-limiting and spreads only a finite distance before dying out. We conjecture that percolation models, thresholds, and phase transitions will prove to be useful for describing the spread of NLRP3-mediated inflammation through affected tissue, but this is currently only a conjecture.

Discussion and Conclusions

The previous sections have outlined several different general mechanisms that can create exposure-response relationships with thresholds or threshold-like nonlinearities, meaning S-shaped exposure concentration-response or duration-response functions with a sharp transition from the low to the high levels. Table 9.2 summarizes these mechanisms. Most of them are now known to apply to the assembly, priming, activation, and signaling of the NLRP3 inflammasome, although percolation thresholds for the spread of inflammation are currently speculative. Bistability based on positive feedback loop motifs is plausible both within EMP-exposed target cells (specifically including monocytes, macrophages, and mesothelial cells) and also in the wider cytokine signaling and immune cell population response and cross-talk causal networks to which they belong (see Table 9.1). As summarized in Table 9.3,

Table 9.2 Summary of five general mechanisms for exposure-response thresholds

Threshold mechanism	Main idea	Examples for NLRP3 inflammasome
Receptor-mediated signal transduction	Probability of activation switches from near 0 to near 1 as concentration crosses a threshold	<ul style="list-style-type: none"> NF-κB- signaling to -regulate NLRP3 transcription
Critical mass or concentration is required to trigger an event	Response occurs only if and when accumulated amount or concentration in a compartment reaches a critical point	<ul style="list-style-type: none"> Rupture of lysosome by accumulated intra-lysosomal ROS or other agents NLRP3 priming and assembly of NLRP3 inflammasome require a threshold level of deubiquitinated NLRP3 protein to accumulate
Cooperativity in oligomerization	High cooperativity (Hill coefficient) in oligomerization creates threshold-like concentration-response and a time delay for response (see Fig. 9.2)	<ul style="list-style-type: none"> Exposure concentration threshold for triggering assembly of NLRP3 inflammasome via oligomerization Exposure duration threshold for triggering assembly of NLRP3 inflammasome via oligomerization
Bistability in networks with positive feedback loops	Sufficiently high and prolonged exposures shift move system with positive feedback loops to a new, stable chronic inflammation equilibrium (see Figs. 9.3 and 9.4)	<ul style="list-style-type: none"> Positive feedback loops for ROS and NLRP3 inflammasome (see Table 9.1)
Percolation thresholds in spatial stochastic propagation processes	Spread of inflammation throughout a volume of tissue requires that the probability that a cell will become inflamed if its neighbors are must exceed a critical percolation threshold value	<ul style="list-style-type: none"> Prion-like transmission of inflammation among cells via ASC specks released upon pyroptosis has been identified, but percolation thresholds and phase transitions for tissue inflammation are currently only conjectured here

Table 9.3 Summary of proposed partial answers to risk analysis questions

Risk analysis question	Currently proposed partial answers
Q1: Is there an exposure concentration threshold below which adverse responses do not occur?	A1: Yes. Within a cell, NLRP3-mediated responses do not occur unless exposure concentrations are sufficient to trigger production of active NLRP3 protein (via NF-κB- signaling and deubiquitination), priming, assembly via oligomerization (see left side of Fig. 9.2), and activation via Signal 2 (typically involving MOMP, ROS, LMP, and lysosome disruption). All of these are threshold-like processes. In addition, bistability (or multi-stability) of cytokine and inflammation networks, and perhaps percolation thresholds for spread of inflammation, create thresholds for the inflammatory responses of multiple cells in a tissue
Q2: For any exposure concentration, is there an exposure duration threshold before which adverse responses of interest do not occur?	A2: Yes: If exposure durations are too brief, ASC polymerization and NLRP3 oligomerization will not be completed (see right side of Fig. 9.2)
Q3: If such thresholds exist, how large are they, and on what factors do they depend?	A3: Exposure concentration and duration thresholds needed to cause persistent NLRP3 activation and chronic inflammation are probably the same as or similar to those needed to create a persistent high-ROS inflammatory state. Both probably involve the same overlapping major bistable positive feedback loops (Table 9.1). Thresholds for Signal 2 to activate the NLRP3 inflammasome in a cell decrease with accumulation of active NLRP3 protein in its cytosol
Q4: What is the shape of the concentration-duration-risk relationship above the thresholds? How do physiochemical properties of EMPs and pharmacokinetic and biochemical parameters for exposed individuals affect the answer?	A4: The main shape of the dose-response relationship is probably switch-like, approximating all-or-nothing activation of various NLRP3 inflammasome activation and high-ROS loops. However, damage and disease from activated loops may progress over decades. Physiochemical properties (greater fiber length, stiffness, and availability of iron on the surface of EMPs) can reduce time to lysosome rupture and Signal 2 generation
Q5: How sure can we be about the answers to the preceding questions, and what information would most help to reduce remaining uncertainties?	A5: Oligomerization-based assembly of NLRP3, induction of high levels of mitochondrial and cytosolic ROS, disruption of the lysosome to generate Signal 2, existence of autocrine and other positive feedback loops, and prion-like spread of inflammation are all well established. Percolation thresholds are speculative

current knowledge strongly suggests that NLRP3 inflammasome-mediated responses to EMP and other (e.g., bacterial) exposures have exposure concentration and duration thresholds below which they do not occur. This is because relevant events do not occur unless exposure concentration is kept sufficiently high for sufficiently long to deplete or overwhelm protective resources such as antioxidant pools. Events with threshold-like response characteristics probably include assembly, priming, and activation of the NLRP3 inflammasome, involving initiation and completion of oligomerization and activation of ACS and NLRP3 proteins and accumulation of intralysosomal, mitochondrial, and cytosolic levels of key factors such as ROS to levels needed to generate System 2; activation of feedback loops; and transition from the low-ROS to the high-ROS state.

Much biochemical detail can and should be added to the framework outlined in Tables 9.1, 9.2 and 9.3. Such detail is needed to quantify the timing and speed of feedback loop activations in response to exposures for different substances (e.g., crystalline silica, asbestos fibers, other EMPs with different physiochemical properties) in different target organs and tissues with different biochemical parameters, e.g., reflecting genetic polymorphisms, prior smoking histories, and interindividual variability in biochemistry. More detail is also needed to quantify the subsequent time course of disease initiation and progression following induction of chronic inflammation. But it is already realistic to anticipate that detailed models will have threshold-like responses to exposure concentrations and durations, and to start considering implications for public and occupational health protection and risk analysis of exposure thresholds for inflammation-mediated diseases.

References

- Andersen ME, Preston RJ, Maier A, Willis AM, Patterson J (2014) Dose-response approaches for nuclear receptor-mediated modes of action for liver carcinogenicity: results of a workshop. *Crit Rev Toxicol* 44(1):50–63. <https://doi.org/10.3109/10408444.2013.835785>
- Bagci EZ, Vodovotz Y, Billiar TR, Ermentrout GB, Bahar I (2006) Bistability in apoptosis: roles of bax, bcl-2, and mitochondrial permeability transition pores. *Biophys J* 90(5):1546–1559
- Baroja-Mazo A, Martín-Sánchez F, Gomez AI, Martínez CM, Amores-Iniesta J, Compan V, Barberà-Cremades M, Yagüe J, Ruiz-Ortiz E, Antón J, Buján S, Couillin I, Brough D, Arostegui JI, Pelegrín P (2014) The NLRP3 inflammasome is released as a particulate danger signal that amplifies the inflammatory response. *Nat Immunol* 15(8):738–748. <https://doi.org/10.1038/ni.2919>
- Bednash JS, Mallampalli RK (2016) Regulation of inflammasomes by ubiquitination. *Cell Mol Immunol* 13(6):722–728. <https://doi.org/10.1038/cmi.2016.15>
- Boya P, Kroemer G (2008) Lysosomal membrane permeabilization in cell death. *Oncogene* 27 (50):6434–6451. <https://doi.org/10.1038/onc.2008.310>
- Chakravarty S, Barik D (2017) Steady state statistical correlations predict bistability in reaction motifs. *Mol Biosyst* 13(4):775–784. <https://doi.org/10.1039/c7mb00052a>
- Cox LA Jr (2006) Universality of J-shaped and U-shaped dose-response relations as emergent properties of stochastic transition systems. *Dose Response* 3(3):353–368. <https://doi.org/10.2203/dose-response.0003.03.006>

- Cox LA Jr (2011) A causal model of chronic obstructive pulmonary disease (COPD) risk. *Risk Analysis* 31(1):38–62
- Franklin BS, Bossaller L, De Nardo D, Ratter JM, Stutz A, Engels G, Brenker C, Nordhoff M, Mirandola SR, Al-Amoudi A, Mangan MS, Zimmer S, Monks BG, Fricke M, Schmidt RE, Espenik T, Jones B, Jarnicki AG, Hansbro PM, Bustos P, Marshak-Rothstein A, Hornemann S, Aguzzi A, Kastenmüller W, Latz E (2014) The adaptor ASC has extracellular and ‘prionoid’ activities that propagate inflammation. *Nat Immunol* 15(8):727–737. <https://doi.org/10.1038/ni.2913>
- Gottschalk RA, Martins AJ, Angermann BR, Dutta B, Ng CE, Uderhardt S, Tsang JS, Fraser ID, Meier-Schellersheim M, Germain RN (2016) Distinct NF-κB and MAPK activation thresholds uncouple steady-state microbe sensing from anti-pathogen inflammatory responses. *Cell Syst* 2 (6):378–390. <https://doi.org/10.1016/j.cels.2016.04.016>
- Guisoni N, Loscar ES, Albano EV (2011) Phase diagram and critical behavior of a forest-fire model in a gradient of immunity. *Phys Rev E Stat Nonlin Soft Matter Phys.* 83(1 Pt 1):011125
- Guo H, Callaway JB, Ting JP (2015) Inflammasomes: mechanism of action, role in disease, and therapeutics. *Nat Med* 21(7):677–687. <https://doi.org/10.1038/nm.3893>
- Harijith A, Ebenezer DL, Natarajan V (2014) Reactive oxygen species at the crossroads of inflammasome and inflammation. *Front Physiol* 5:352. <https://doi.org/10.3389/fphys.2014.003>
- Hillegass JM, Miller JM, MacPherson MB, Westbom CM, Sayan M, Thompson JK, Macura SL, Perkins TN, Beuschel SL, Alexeeva V, Pass HI, Steele C, Mossman BT, Shukla A (2013) Asbestos and erionite prime and activate the NLRP3 inflammasome that stimulates autocrine cytokine release in human mesothelial cells. Part Fibre Toxicol 10:39. <https://doi.org/10.1186/1743-8977-10-39>
- Jin S, Li Y, Pan R, Zou X (2014) Characterizing and controlling the inflammatory network during influenza A virus infection. *Sci Rep* 4:3799. <https://doi.org/10.1038/srep03799>
- Li X, Levine H (2017) Bistability of the cytokine-immune cell network in a cancer microenvironment. *Converg Sci Phys Oncol* 3(2):28. <http://iopscience.iop.org/article/10.1088/2057-1739/aa6c07>
- Malka R, Shochat E, Rom-Kedar V (2010) Bistability and bacterial infections. *PLoS One* 5(5): e10010. <https://doi.org/10.1371/journal.pone.0010010>
- Nakamura K, Kawakami T, Yamamoto N, Tomizawa M, Fujiwara T, Ishii T, Harigae H, Ogasawara K (2016) Activation of the NLRP3 inflammasome by cellular labile iron. *Exp Hematol* 44(2):116–124. <https://doi.org/10.1016/j.exphem.2015.11.002>
- Okada M, Matsuzawa A, Yoshimura A, Ichijo H (2014) The lysosome rupture-activated TAK1-JNK pathway regulates NLRP3 inflammasome activation. *J Biol Chem* 289(47):32926–32936. <https://doi.org/10.1074/jbc.M114.579961>
- Pereira EJ, Smolko CM, Janes KA (2016) Computational models of reactive oxygen species as metabolic byproducts and signal-transduction modulators. *Front Pharmacol* 7:457
- Py BF, Kim MS, Vakifahmetoglu-Norberg H, Yuan J (2013) Deubiquitination of NLRP3 by BRCC3 critically regulates inflammasome activity. *Mol Cell* 49(2):331–338. <https://doi.org/10.1016/j.molcel.2012.11.009>
- Qiao Q, Wu H (2015) Supramolecular organizing centers (SMOCs) as signaling machines in innate immune activation. *Sci China Life Sci* 58(11):1067–1072. <https://doi.org/10.1007/s11427-015-4951-z>
- Repnik U, Hafner Česen M, Turk B (2014) Lysosomal membrane permeabilization in cell death: concepts and challenges. *Mitochondrion* 19 Pt A:49–57. <https://doi.org/10.1016/j.mito.2014.06.006>
- Sayan M, Mossman BT (2016) The NLRP3 inflammasome in pathogenic particle and fibre-associated lung inflammation and diseases. Part Fibre Toxicol 13(1):51. <https://doi.org/10.1186/s12989-016-0162-4>
- Schilling JD (2016) Dousing fire with gasoline: interplay between lysosome damage and the NLRP3 inflammasome. Focus on “NLRP3 inflammasome signaling is activated by low-level lysosome disruption but inhibited by extensive lysosome disruption: roles for K⁺ efflux and Ca²⁺ influx”. *Am J Physiol Cell Physiol* 311(1):C81–C82. <https://doi.org/10.1152/ajpcell.00174.2016>

- Shao B-Z, Xu Z-Q, Han B-Z, Su D-F, Liu C (2015) NLRP3 inflammasome and its inhibitors: a review. *Front Pharmacol* 6:262. <https://doi.org/10.3389/fphar.2015.00262>
- Shi CS, Shenderov K, Huang NN, Kabat J, Abu-Asab M, Fitzgerald KA, Sher A, Kehrl JH (2012) Activation of autophagy by inflammatory signals limits IL-1 β production by targeting ubiquitinated inflammasomes for destruction. *Nat Immunol* 13(3):255–263
- Siegel-Gaskins D, Mejia-Guerra MK, Smith GD, Grotewold E (2011) Emergence of switch-like behavior in a large family of simple biochemical networks. *PLoS Comput Biol* 7(5):e1002039. <https://doi.org/10.1371/journal.pcbi.1002039>
- Squires S, Sytwu K, Alcalá D, Antonsen TM, Ott E, Girvan M (2013) Weakly explosive percolation in directed networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 87(5):052127
- Thompson JK, Westbom CM, MacPherson MB, Mossman BT, Heintz NH, Spiess P, Shukla A (2014) Asbestos modulates thioredoxin-thioredoxin interacting protein interaction to regulate inflammasome activation. *Part Fibre Toxicol* 11:24. <https://doi.org/10.1186/1743-8977-11-24>
- Thompson JK, MacPherson MB, Beuschel SL, Shukla A (2017) Asbestos-induced mesothelial to fibroblastic transition is modulated by the inflammasome. *Am J Pathol* 187(3):665–678. <https://doi.org/10.1016/j.ajpath.2016.11.008>
- Veltman D, Laeremans T, Passante E, Huber HJ (2017) Signal transduction analysis of the NLRP3-inflammasome pathway after cellular damage and its paracrine regulation. *J Theor Biol* 415:125–136. <https://doi.org/10.1016/j.jtbi.2016.12.016>
- Wu H (2013) Higher-order assemblies in a new paradigm of signal transduction. *Cell* 153(2):287–292. <https://doi.org/10.1016/j.cell.2013.03.013>
- Zhu W, von dem Bussche A, Yi X, Qiu Y, Wang Z, Weston P, Hurt RH, Kane AB, Gao H (2016) Nanomechanical mechanism for lipid bilayer damage induced by carbon nanotubes confined in intracellular vesicles. *Proc Natl Acad Sci U S A* 113(44):12374–12379

Part IV

Evaluation Analytics

Chapter 10

Evaluation Analytics for Public Health: Has Reducing Air Pollution Reduced Death Rates in the United States?



Introduction: Using Data from Natural Experiments to Understand Causality

An aim of applied science in general, and of epidemiology in particular, is to draw sound causal inferences from observations. For public health policy analysts and epidemiologists, this includes drawing inferences about whether historical changes in exposures have actually caused the consequences predicted for, or attributed to, them. The example of the Dublin coal-burning ban introduced in Chap. 1 suggests that accurate evaluation of the effect of interventions is not always easy, even when data are plentiful. Students are taught to develop hypotheses about causal relations, devise testable implications of these causal hypotheses, carry out the tests, and objectively report and learn from the results to refute or refine the initial hypotheses. For at least the past two decades, however, epidemiologists and commentators on scientific methods and results have raised concerns that current practices too often lead to false-positive findings and to mistaken attributions of causality to mere statistical associations (Lehrer 2012; Sarewitz 2012; Ottenbacher 1998; Imberger et al. 2011). Formal training in epidemiology may be a mixed blessing in addressing these concerns. As discussed in Chap. 2, concepts such as “attributable risk,” “population attributable fraction,” “burden of disease,” “etologic fraction,” and even “probability of causation” are solidly based on relative risks and related measures of statistical association; they do not necessarily reveal anything about predictive, manipulative, structural, or explanatory (mechanistic) causation (e.g., Cox 2013; Greenland and Brumback 2002). Limitations of human judgment and inference, such as confirmation bias (finding what we expect to find), motivated reasoning (concluding what it pays us to conclude), and overconfidence (mistakenly believing that our own beliefs are more accurate than they really are), do not spare health effects investigators. Experts in the health effects of particular compounds are not always also experts in causal analysis, and published causal conclusions are

often unwarranted, as reviewed in Chap. 2, with a pronounced bias toward finding “significant” effects where none actually exists (false positives) (Lehrer 2012; Sarewitz 2012; Ioannidis 2005; The Economist 2013).

This chapter applies methods of causal hypothesis-testing (Granger causality tests and conditional independence tests) to evaluation analytics, i.e., assessment of the effects caused by past changes, for the important practical problem of assessing improvements in public health risks caused by past reductions in air pollution concentrations. To do so, we take advantage of the fact that between years 2000 and 2010, air pollutant levels in counties throughout the United States changed significantly, with fine particulate matter (PM2.5) declining over 30% in some counties, and ozone (O₃) exhibiting large variations from year to year. This history provides an opportunity to compare county-level changes in average annual ambient pollutant levels to corresponding changes in all-cause and cardiovascular disease (CVD) mortality rates over the course of a decade. This chapter examines data from these “natural experiments” of changing pollutant levels for 483 counties in the 15 most populated U.S. states using quantitative methods for causal hypothesis testing, such as conditional independence and Granger causality tests. We shall see that the hypothesis of a significant statistical association between air pollution and mortality rates is well supported, but that the hypothesis of a predictive causal relation between them is not. For example, no significant positive associations are found between *changes* in PM2.5 or O₃ levels and corresponding changes in disease mortality rates between 2000 and 2010, nor for shorter time intervals of 1–3 years.

Dominici et al. (2014) noted that “[A]nalyses of observational data have had a large impact on air-quality regulations and on the supporting analyses of their accompanying benefits, [but] associational approaches to inferring causal relations can be highly sensitive to the choice of the statistical model and set of available covariates that are used to adjust for confounding. . . . There is a growing consensus. . . that the associational or regression approach to inferring causal relations—on the basis of adjustment with observable confounders—is unreliable in many settings.” They demonstrate that the choice of regression model can result in either statistically significant positive or statistically significant negative associations between air pollutant levels and mortality rates. This implies that implicit modeling choices can greatly affect—or even determine—the results presented to decision-makers and the public. Table 10.1 provides some examples of important policy-relevant conclusions and doubts about their validity from the recent air pollution health effects literature.

To overcome this difficulty, Dominici et al. (2014) proposed the use of *quasi-experiments* (QEs), or natural experiments, in which outcomes are compared between a treatment and control group are compared, but without random assignment or other determination of the treatment status by the researcher. As an example, they cite the Dublin coal-burning ban study discussed in Chap. 1, reporting significantly lower mortality rates in the 6 years following a ban on coal-burning in Dublin County, Ireland compared to the 6 years prior to the ban (Clancy et al. 2002). This proposal to use QEs to better assess causal relations between pollution levels and health effects has been hailed as “a paradigm-shifting solution” (Harvard Law Today 2014).

Table 10.1 Some conflicting claims about health effects known to be caused by air pollution

Pro (causal interpretation or claim)	Con (counter-interpretation or claim)
“Epidemiological evidence is used to quantitatively relate PM _{2.5} exposure to risk of early death. We find that UK combustion emissions cause ~13,000 premature deaths in the UK per year, while an additional ~6000 deaths in the UK are caused by non-UK European Union (EU) combustion emissions” (Yim and Barrett 2012)	“[A]lthough this sort of study can provide useful projections, its results are only estimates. In particular, although particulate matter has been associated with premature mortality in other studies, a definitive cause-and-effect link has not yet been demonstrated” (NHS 2012)
“[A]bout 80,000 premature mortalities [per year] would be avoided by lowering PM _{2.5} levels to 5 µg/m ³ nationwide” in the U.S. 2005 levels of PM _{2.5} caused about 130,000 premature mortalities per year among people over age 29, with a simulation-based 95% confidence interval of 51,000–200,000 (Fann et al. 2012)	“Analysis assumes a causal relationship between PM exposure and premature mortality based on strong epidemiological evidence... However, epidemiological evidence alone cannot establish this causal link” (EPA 2011, Tables 6–11) Significant negative associations have also been reported between PM _{2.5} (Krstić 2011) and short-term mortality and morbidity rates, as well as between levels of some other pollutants (e.g., NO ₂ (Kelly et al. 2011) and ozone (Powell et al. 2012)) and short-term mortality and morbidity rates
“Some of the data on the impact of improved air quality on children’s health are provided, including... the reduction in the rates of childhood asthma events during the 1996 Summer Olympics in Atlanta, Georgia, due to a reduction in local motor vehicle traffic” (Buka et al. 2006). “During the Olympic Games, the number of asthma acute care events decreased 41.6% (4.23 vs. 2.47 daily events) in the Georgia Medicaid claims file,” coincident with significant reductions in ozone and other pollutants (Friedman et al. 2001)	“In their primary analyses, which were adjusted for seasonal trends in air pollutant concentrations and health outcomes during the years before and after the Olympic Games, the investigators did not find significant reductions in the number of emergency department visits for respiratory or cardiovascular health outcomes in adults or children.” In fact, “relative risk estimates for the longer time series were actually suggestive of increased ED [emergency department] visits during the Olympic Games” (Health Effects Institute 2010)
“An association between elevated PM ₁₀ levels and hospital admissions for pneumonia, pleurisy, bronchitis, and asthma was observed. During months when 24-h PM ₁₀ levels exceeded 150 µg/m ³ , average admissions for children nearly tripled; in adults, the increase in admissions was 44%” (Pope 1989)	“Respiratory syncytial virus (RSV) activity was the single explanatory factor that consistently accounted for a statistically significant portion of the observed variations of pediatric respiratory hospitalizations. No coherent evidence of residual statistical associations between PM ₁₀ levels and hospitalizations was found for any age group or respiratory illness” (Lamm et al. 1994)
“Reductions in respiratory and cardiovascular death rates in Dublin suggest that control of particulate air pollution could substantially diminish daily death....Our findings suggest that control of particulate air pollution in Dublin led to an immediate reduction in cardiovascular and respiratory deaths” (Clancy et al. 2002)	Mortality rates were already declining long before the ban, and occurred in areas not affected by it. “Serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black-smoke exposure” (Wittmaack 2007). “Thus, a causal link between the decline in mortality and the

(continued)

Table 10.1 (continued)

Pro (causal interpretation or claim)	Con (counter-interpretation or claim)
"The results could not be more clear, reducing particulate air pollution reduces the number of respiratory and cardiovascular related deaths immediately" (Harvard School of Public Health 2002)	ban of coal sales cannot be established" (Pelucchi et al. 2009). "In contrast to the earlier study, there appeared to be no reductions in total mortality or in mortality from other causes, including cardiovascular disease, that could be attributed to any of the bans. That is, after correcting for background trends, similar reductions were seen in ban and non-ban areas" (HEI 2013)

Source: Adapted from Cox (2013)

Yet, since QEs were first introduced in social statistics in the 1960s, expert practitioners have recognized that "in many quasi-experiments, one is most often left with the question: 'Are there alternative explanations for the apparent causal association?'" (Harris et al. 2006). Such alternative explanations, or threats to the internal validity of causal inferences for the studied populations, are discussed in Chap. 1. They must be refuted before valid causal inferences can be drawn from QEs (Campbell and Stanley 1966; MacIur 1991; Rothman and Greenland 2005). For example, to be valid, the conclusion that a ban on coal-burning *caused* an immediate reduction in all-cause and cardiovascular mortality (Harvard School of Public Health 2002) would have had to refute plausible alternative explanations. Including a relevant historical or contemporaneous control group (using a pretest-posttest design or a nonequivalent control group design, respectively, in QE terminology) would have allowed the elimination of non-causal explanations, such as that (a) mortality rates were already declining before the ban, and continued to do so without significant change during and afterward for reasons unrelated to the ban (the "History" threat to internal validity, in QE terminology); or (b) mortality rates declined at the same rate in areas not affected by the ban as in areas affected by it. For the Dublin study, both possibilities (a) and (b) proved to be true, so that no valid conclusions about the impact of the ban on all-cause or cardiovascular mortality rates can be drawn (Wittmaack 2007; Pelucchi et al. 2009). Indeed, upon reanalysis using relevant control groups, no effect of the ban on these outcomes could be detected (Health Effects Institute 2013). Yet, as Dominici et al., rightly note, natural experiments occur frequently and if properly analyzed, can provide crucial policy-relevant insights into causality (or lack thereof) in observed exposure-response relations. In the U.S. for example, geographic heterogeneity in the rates at which pollutant levels have declined in different regions has created many natural experiments for assessing the effects of these changes on public health over time.

To take advantage of these natural experiments, the following sections compare changes in PM2.5 and O3 levels from 2000 to 2010 to corresponding changes in all-cause and CVD age-specific mortality rates over the same interval, for hundreds of counties in the 15 largest states in the U.S. Treating county as the unit of observation, as in the Dublin study and many others where individual-level exposure

data are not available, invites application of longitudinal designs and methods in which each county's history of pollution levels and mortality rates serves as its own control group for purposes of determining how subsequent changes in pollution are associated with subsequent changes in mortality rates (Campbell and Stanley 1966). Using repeated observations on the same counties over time also allows the effects of unmeasured (and possibly unknown) confounders to be largely controlled for as changes in pollutant levels and mortality rates are calculated—the basic strategy of panel data analysis (Angrist and Pischke 2009). The goal of our analysis is to understand the extent to which historical associations between pollutant levels and mortality rates reflect a clear causal relation, rather than merely coincident trends, the effect of confounders, or modeling choices.

Table 10.2 lists several quantitative methods for causal hypothesis testing, modeling, and analysis that have been extensively developed and applied over the

Table 10.2 Some formal methods for modeling and testing causal hypotheses

Method and references	Basic idea	Appropriate study design
Quasi-experimental design and analysis (Campbell and Stanley 1966)	Can control group comparisons refute alternative (non-causal) explanations for observed associations between hypothesized causes and effects, e.g., coincident trends and regression to the mean? If so, this strengthens causal interpretation	Observational data on subjects exposed and not exposed to interventions that change the hypothesized cause(s) of effects
Conditional independence tests (Freedman 2004; Friedman and Goldszmidt 1998)	Is hypothesized effect (e.g., cardiovascular disease (CVD) mortality rate) statistically independent of hypothesized cause (e.g., PM2.5 concentration), given (i.e., conditioned on) the values of other variables, such as education and income? If so, this undermines causal interpretation	Cross-sectional data; Can also be applied to multi-period data (e.g., in dynamic Bayesian networks)
Panel data analysis (Angrist and Pischke 2009; Stebbings 1978)	Are changes in exposures followed by changes in the effects that they are hypothesized to help cause? If not, this undermines causal interpretation; if so, this strengthens causal interpretation. Example: Are reductions in PM2.5 levels followed (but not preceded) by corresponding changes in CVD mortality rates?	Panel data study: Collect a sequence of observations on same subjects or units of observation (e.g., counties) over time
Granger causality test (Eichler and Didelez 2010)	Does the history of the hypothesized cause improve ability to predict the future of the hypothesized effect? If so, this strengthens causal interpretation; otherwise, it undermines causal interpretation Example: Can CVD mortality rates be predicted better from time series histories of PM2.5 levels and mortality rates than from the time series history of mortality rates alone?	Time series data on hypothesized causes and effects

(continued)

Table 10.2 (continued)

Method and references	Basic idea	Appropriate study design
Intervention analysis and change point analysis (Helfenstein 1991; Gilmour et al. 2006)	<p>Does the best-fitting model of the observed data change significantly at or following the time of an intervention? If so, this strengthens causal interpretation</p> <p>Do the quantitative changes in hypothesized causes predict and explain the subsequently observed quantitative changes in hypothesized effects? If so, this strengthens causal interpretation</p> <p>Example: Do mortality rates fall faster in counties where pollutant levels fall faster than in other counties?</p>	<p>Time series observations on hypothesized effects, and knowledge of timing of intervention(s)</p> <p>Quantitative time series data for hypothesized causes and effects</p>
Counterfactual and potential outcome models (Moore et al. 2012)	<p>Do exposed individuals have significantly different response probabilities than they would have had if they had not been exposed?</p> <p>Example: Do people have lower mortality risk after historical exposure reductions than they would have had otherwise?</p>	Cross-sectional and/or longitudinal data, with selection biases and feedback among variables allowed
Causal network, path analysis, and structural equations models of change propagation (Hack et al. 2010)	<p>Do changes in exposures (or other causes) create a cascade of changes through a network of causal mechanisms (represented by equations), resulting in changes in the effect variables?</p> <p>Example: Do relatively large variations in daily levels of fine particulate matter (PM2.5) air pollution create corresponding variations in markers of oxidative stress in the lungs?</p>	Observations of variables in a dynamic system out of equilibrium
Negative controls (for exposures or for effects) (Lipsitch et al. 2010)	<p>Do exposures predict health effects better than they predict effects that cannot be caused by exposures?</p> <p>Example: Do pollutant levels predict cardiovascular mortality rates better than they explain car accident mortality rates? If not, this weakens causal interpretation of the CVD associations</p>	Observational studies

Source: Adapted from Cox (2013)

past six decades (Cox 2013). Chapter 2 provides much more detail on several of these methods. Various advantages of these techniques, as compared to qualitative causal criteria (Rothman and Greenland 2005) such as the traditional Hill considerations and other weight-of-evidence and associational methods, are well explained and illustrated in the references for Table 10.2 (e.g., Greenland and Brumback 2002), along with their limitations (e.g., Freedman 2004). Prominent among these advantages is the development of empirically testable implications of causal hypotheses, such as conditional independence implications, timing implications, information-

theoretic implications, and exogeneity implications, with conditional probability distributions of some variables being determined by the values of others (see Chap. 2). These testable implications capture the inherent asymmetry inherent in the notion of causation, unlike correlations or other symmetric measures of association. They can be tested statistically using publically available standard computer codes, such as those in R and Python/NumPy. This enables different investigators, perhaps with very different prior beliefs, to reach the same conclusions from the same data. This points the way toward greater objectivity and definitiveness in determining via such tests the extent to which data do or do not support causal hypotheses, based on their testable implications.

Other reasons why modern methods of quantitative causal analysis should be (and increasingly are) included among current approaches in the epidemiologist's tool kit are discussed in modern epidemiology textbooks and monographs (e.g., Hernan and Robbins 2018) and in the references to Table 10.2. The purpose of this chapter is not to further review these methods, but to apply those that are most useful to the air pollution and mortality rate records in the United States.

Data and Methods

Cause-specific mortality rates, by county and age group, were downloaded from the Centers for Disease Control and Prevention (CDC) Wonder "Compressed Mortality, 1999–2010" database (CDC 2014). To create a geographically diverse sample, mortality rates were extracted at the county level for the 15 largest states in the U.S. (California, Texas, New York, Florida, Illinois, Pennsylvania, Ohio, Georgia, Michigan, North Carolina, New Jersey, Virginia, Washington, Massachusetts, Arizona) representing approximately 65% of the total U.S. population. We extracted mortality rates (per hundred thousand person-years) for all causes of death, and then created three disease subcategories: (1) diseases of the circulatory system (International Classification of Diseases, 10th revision codes [ICD-10]I00-I99), (2) all external causes of death (ICD-10 codes V01-Y89) and (3) total disease-related mortalities (all causes of death excluding external causes). The dependent variables shown in subsequent tables thus included the following:

- *CVRatePer100K*—Mortality rate (per 100,000 people per year) due to all heart/circulatory diseases
- *ExtRatePer100K*—Mortality rate due to external causes (used as a negative control). (To investigate whether the methods used can detect causal known relationships, we also used a positive control in which a known causal effect was simulated, as discussed later for Table 10.7.)
- *ACRatePer100K*—Mortality rate due to all disease-related (non-external) causes

Most of our analyses were restricted to ages 65+ years, as they have the highest CVD mortality rates. Age was categorized as 65–74 years, 75–84 years, and 85+ years.

County-level air quality data for PM2.5 (daily 24-h mean) and O₃ (daily maximum 8-h moving average) were downloaded from the U.S. Environmental Protection Agency Air Quality System (AQS) for all monitors located in each county ($n = 483$) of the 15 states listed above (EPA 2014). Data were obtained for the years 2000–2010. The two pollutant measures were summarized as county-level annual averages in our analyses.

The mortality and air quality data were merged by state/county and year. The resulting merged data file contained data for 483 distinct counties from 2000 to 2010, although not all counties collected both ozone and PM2.5 data for all years. These merged data files are freely available from the authors upon request.

Statistical Analysis Methods

The methods in Table 10.2 that are most useful for the air pollution and mortality rate data sets just described include conditional independence tests, longitudinal comparisons of changes in death rates and changes in pollution levels, Granger causality tests, and negative controls comparing presumably non-causal associations between longitudinal changes in accident and other “external” (non-disease) death rates and changes in pollutant levels to associations between changes in disease mortality rates and changes in pollutant levels. These are described in the following paragraphs. All statistical computations were carried out using the *Statistica 12.5* statistical computing environment, with the exception of the Granger causality tests, described below. Other methods in Table 10.2, such as change-point analysis and intervention analysis for an intervention that occurs at a single point in time (e.g., closing a steel mill or banning coal-burning in Dublin) are less relevant for these data, since both changes in PM2.5 and changes in mortality rates occurred gradually over a decade, rather than abruptly from before to after some intervention.

Association-Based Methods: Correlation and Regression

Although not methods of causal analysis, association-based methods such as correlation and regression analysis are widely used in air pollution health effects research (Dominici et al. 2014). We used these methods also to test whether applying them in this data set produced similar results to past studies. Intuitively, the absence of any association might be interpreted to suggest that causation is unlikely (Hill 1965). We used Pearson product-moment linear correlation coefficients and linear regression coefficients as measures of linear association, since past research suggests an approximately linear association of PM2.5 and O₃ with mortality (e.g., Lepeule et al. 2012).

Conditional Independence Tests

If a statistically significant association between exposure and response variables is found, e.g., based on linear correlation and regression tests, then an important screening test for potential causation is the *conditional independence* test: does a significant association remain even after conditioning on potential confounders, such as age or year? For example, if a significant association between PM2.5 and CVD mortality were hypothesized to be due to confounding by year (because both PM2.5 and CVD mortality rates both declined with time, even if one did not cause the other), then one could condition on year (i.e., holding it fixed at a given value, such as 2010), and test whether the conditional association vanishes within the subset of records with that value (e.g., with Year = 2010).

To avoid biasing results by manual selection of variables to condition on, we relied on automated backward stepwise variable selection in our multiple regression models. This is a standard—but deservedly controversial—technique. We do not advocate it for general use, as it over-fits models to data, producing excess false positives in simple settings. We therefore have used it only as a readily available automated approach that may be more familiar and easily available than alternatives such as Bayesian Model Averaging; but we have also verified the main conclusions using multiple disjoint random samples of the data (20% cross-validation), to guard against the defects of backward stepwise selection. The backward stepwise selection procedure uses successive F tests to determine whether dropping individual variables (e.g., O₃ concentration) from the set of potential explanatory variables significantly decreases the ability of the model to predict values of the dependent variable (e.g., CVD mortality risk). If not, i.e., if the F test indicates that the dependent variable is conditionally independent of a potential explanatory variable (such as O₃), given the values of other variables in the model, then that variable is automatically dropped from the final set of explanatory variables. Despite its flaws, use of this technique reduces subjectivity in choosing explanatory variables. We used the default settings in *Statistica* (e.g., *p* values of 0.05 to define significant associations).

Correlations Among Changes over Time

Perhaps the most important screening test we use for potential causality is examining whether changes in an exposure help predict and explain changes in a response. A frequent confusion in epidemiology is to interpret the *slope* of a concentration-response relation as indicating the future *change* in response (e.g., mortality rates) that would be caused by a unit change in future exposure concentration. This is incorrect, since many concentration-response associations are not entirely causal (e.g., due to confounders or modeling biases). Rather than using slopes of cross-sectional regression lines as proxies for causal impacts, we directly tested whether there were significant positive correlations and regression coefficients between longitudinal changes in county-specific PM2.5 and O₃ levels from 2000 to 2010,

and corresponding longitudinal changes in county-specific and age-specific mortality rates; and whether counties with more rapid declines in PM2.5 and O₃ had more rapid declines in mortality than those with slower declines, or where concentrations increased.

Granger Causality and Negative and Positive Controls

A more general approach than studying associations between changes in exposure concentrations and changes in mortality rates over a single time interval is to use time series analysis to test whether past values of exposure help to predict present and future mortality rates more accurately than they can be predicted from past mortality rates alone. This is the basic idea of the *Granger causality test* (Eichler and Didelez 2010). If the future of a mortality rate time series is conditionally independent of the past and present exposure time series, given the past and present mortality rate series, so that knowing exposure does not improve ability to predict future mortality rates, then exposure is not a Granger-cause of mortality. The Granger causality test produces a p-value for the null hypothesis that one time series does not improve prediction of another compared to using lagged values of the dependent variable itself.

We performed the Granger tests, using the *grangercausalitytests* function in the Python *statsmodels* module, for each county and age category combination described above, with the restriction that the combination must have at least ten consecutive annual values available for analysis. We tested lags of 1–3 years, as many previous studies suggest that reductions in PM2.5 and other pollutants lead to almost immediate reductions in mortality rates, e.g., within as little as a few days, and certainly well within a year or 2 (e.g., Friedman et al. 2001; Clancy et al. 2002; Lepeule et al. 2012; Yang et al. 2013). The Python Granger function *grangercausalitytests* provides *p* values for each of four separate test statistics (two based on the *F* distribution and two on the chi-square distribution), all of which yield closely similar results. We evaluated the proportion of counties for which these tests produced a *p*-value of 0.05 or less; random variation alone could explain this occurring in about 5% of counties. Significantly higher levels would be suggestive of a Granger causality effect.

In addition to formal test statistics, we also compared the statistical association between changes in exposures and changes in disease-related mortality rates, on the one hand, to the association between changes in exposures and changes in non-disease-related (external-cause) mortality rates, on the other. The external-cause mortality rates include deaths due to accidents and assaults, changes in which are presumably not caused by changes in pollution levels. Such negative controls test whether hypothesized causal associations are stronger than those presumed to be non-causal (Lipsitch et al. 2010). As discussed further later (c.f. discussion of Table 10.7), we also simulated the effects of a positive causal relation between changes in pollution levels and changes in mortality rates. This simulation-based analysis served as a type of positive control to test whether sample sizes are large

enough and whether the statistical methods we applied are powerful enough to detect such genuine causal effects if they are present. Finally, we briefly examined the geographic pattern of results to determine whether findings appeared to hold consistently in different parts of the United States.

Results

Descriptive Analytics

Figure 10.1 shows trends in average pollution levels, population, and mortality rates for all counties from 2000 to 2010. For each time series, values are normalized by dividing by the value in 2000, so that all time series values in 2000 are defined as 1. PM2.5 and CVD mortality rates declined most steeply over this interval (two lowest curves), while population levels and external-cause mortality rates (e.g., from accidents) increased, perhaps reflecting a longer-lived, aging population.

Figure 10.2 shows how the age-specific mortality curve, plotting annual deaths per capita vs. age, has shifted downward over time. (The horizontal positions for the rates have been spread out to allow easy visualization of trends. Vertical bars indicate 95% confidence intervals for the mean mortality rates but are very narrow

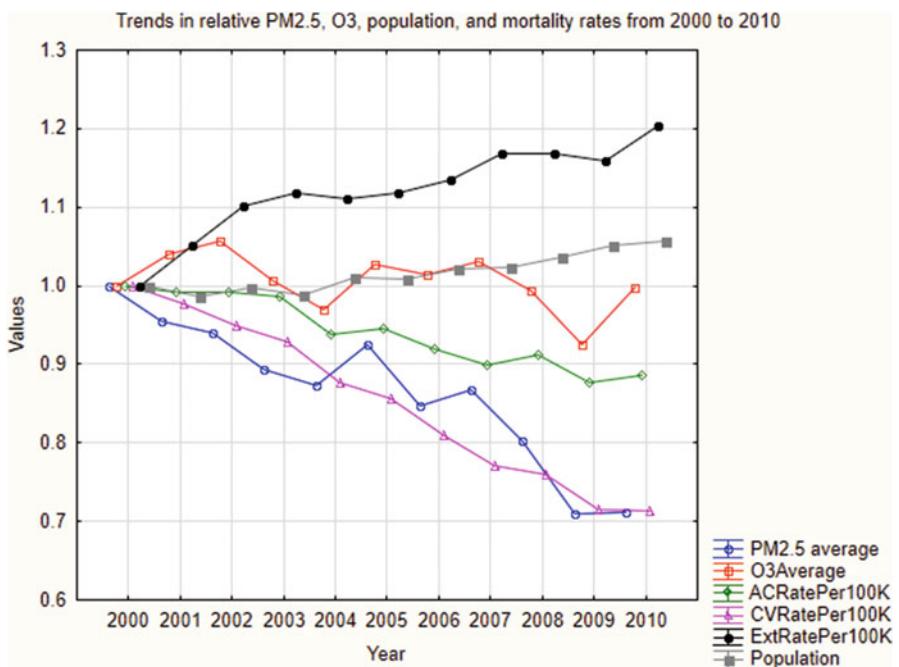


Fig. 10.1 Trends in relative values of pollutants, mortality rates, and population, 2000–2010

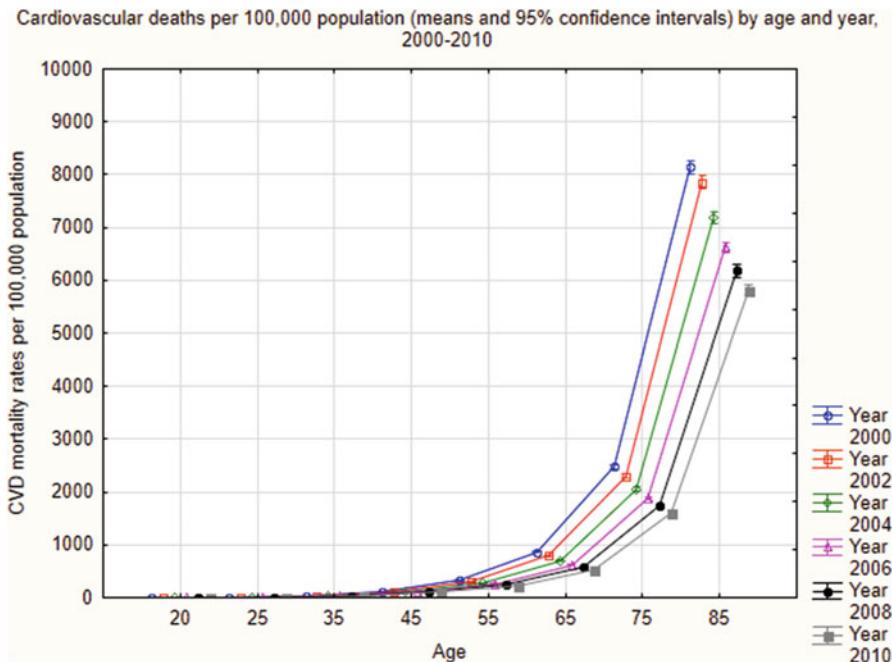


Fig. 10.2 Declines of age-specific cardiovascular disease (CVD) mortality rates over time (top curve is for year 2000, bottom curve is for year 2010)

due to the large sample sizes.) Clearly, age-specific mortality rates have declined for all age groups, but most for the older age groups.

Figure 10.3 shows analogous curves for age groups 55–64, 65–74, 75–84, and 85 or older, abbreviated 55, 65, 75, and 85, respectively, for different average PM2.5 levels in 2000 (left) and 2010 (right). At all PM2.5 levels, age-specific mortality rates declined conspicuously from 2000 to 2010. In both years, mortality rates in the oldest age categories were higher at PM2.5 levels of $12 \mu\text{g}/\text{m}^3$ than at $3 \mu\text{g}/\text{m}^3$, suggesting a possible persistent positive association between PM2.5 concentrations and elderly mortality rates.

There was substantial geographic heterogeneity in both PM2.5 values and CVD mortality rates among the counties in this study, allowing the relation between them to be studied with considerable statistical power despite the smoothing effects of using county-level data (Savitz 2012). PM2.5 average levels ranged from below 2 to above $20 \mu\text{g}$ per cubic meter, and cardiovascular deaths per 100,000 people per year ranged from close to zero (for younger age groups) to over 10,000 deaths per 100,000 person-years (for the oldest age group in early years). Even for a single age group (e.g., 75–84 year-olds) and a single year (2010), there is a greater than fivefold variation in CVD mortality rates and a more than eightfold variation in average PM2.5 levels among counties, as shown in Fig. 10.4.

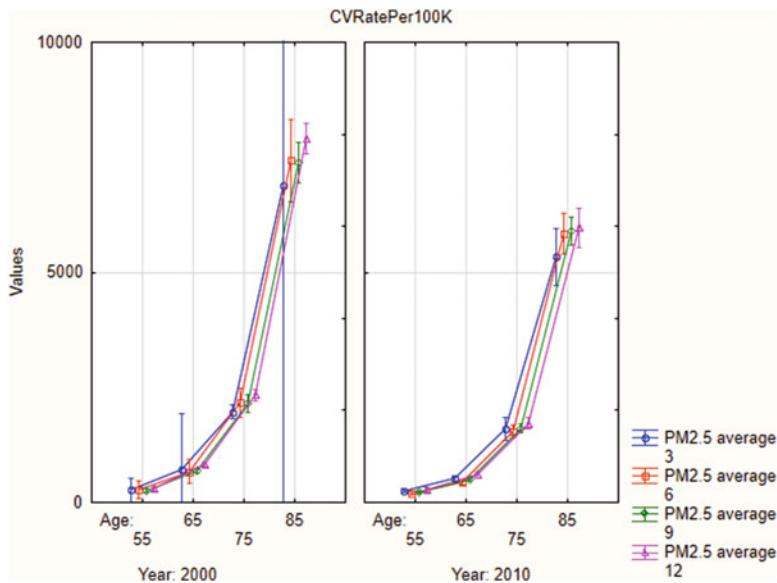


Fig. 10.3 Decline of older age-specific mortality rates over time (left panel is for year 2000, right panel is for year 2010) for counties with different average PM2.5 levels

Bivariate histogram for CVD mortality rate (per 100,000 people per year) and PM2.5 (average micrograms per cubic meter) for people aged 75–84 in 2010

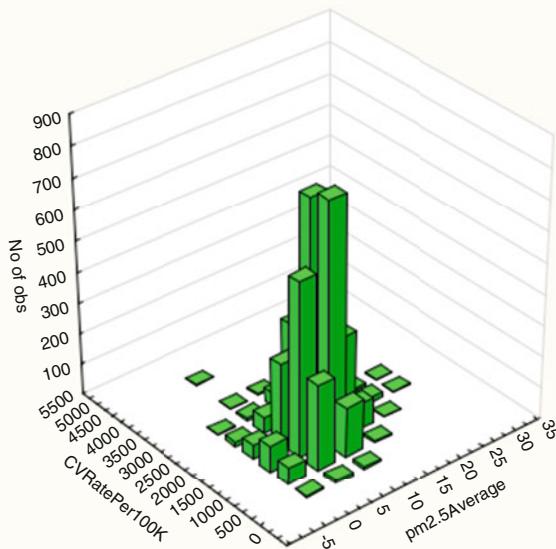


Fig. 10.4 There is substantial geographic heterogeneity in PM2.5 levels and CVD mortality rates even within a single age group and year (here, 75–84 year olds in 2010)

Results on Statistical Associations Between Pollutant Levels and Mortality Rates

Table 10.3 shows the Pearson correlation coefficients between PM2.5 and O3 levels, county population sizes, and all-cause, cardiovascular, and external-cause (non-disease) mortality rates, holding year and age fixed at 2010 and 75–84 years, respectively. Similar correlations hold for other years. All off-diagonal correlation coefficients in Table 10.3 are statistically significant from zero ($p < 0.05$) except for the -0.09 correlation between PM2.5 levels and non-disease mortality rates (ExtRatePer100k). Specifically, Table 10.3 shows the following significant associations:

- PM2.5 and O3 concentrations are positively associated with each other (correlation $r = 0.28$)
- Both PM2.5 and O3 concentrations are positively associated with both all-cause and cardiovascular mortality rates.
- O3 is also positively associated with non-disease mortality rates but PM2.5 is not. (All positive correlations in Table 10.3 are significant, but the -0.09 numbers are not.)
- Population size of a county is positively associated with PM2.5 and is negatively associated with O3 and with all mortality rates.
- All mortality rates (disease-related and non-disease-related) are positively associated with each other, but negatively associated with population size.

The associations in Table 10.3 may or may not be causal, but they are not explained by coincident historical trends (since the year is held fixed at 2010) nor by confounding by age category, since the age category is also held fixed at 75–84. Whether confounding by education, income, temperature, or other variables might account for some of these associations—for example, if mortality rates and PM2.5 are both elevated on cold days or in colder regions; or if lower-income families tend to live in more polluted areas and also to have higher age-specific mortality rates irrespective of location—cannot be determined from the exposure and mortality rate data alone.

In multiple linear regression modeling of the association between explanatory variables and elderly (75–84 years-old) CVD mortality rate using automated backward stepwise variable selection via F tests, only the regression coefficient between PM2.5 and CVD mortality rate, but not O3 and CVD mortality risk, remains significant. Thus, there is a positive association between PM2.5 levels and CVD mortality rates among the elderly that is not explained by coincident historical trends, nor by confounding by age or population or O3; but the correlations between O3 and CVD mortality rates, and between O3 and all-disease mortality rate, vanish after conditioning (via multiple linear regression) on PM2.5 and population size for all disease-related mortalities. In short, PM2.5, but not O3, passes this conditional independence test for being a potential causal driver of elderly mortality rates. Similarly, for all age categories and years, PM2.5 average levels but not O3 levels help to predict CVD mortality rates.

Table 10.3 Pearson correlations between pairs of exposure and response variables for elderly (75–84 year-old) people in 2010

Variable	Correlations between county-specific average PM2.5 and O3 concentrations and mortality rates for 75–84 year olds in 2010						
	Means	PM2.5 average	O3 average	Population	ACRatePer100K	CVRatePer100K	ExtRatePer100K
PM2.5 average	9.16	1.00	0.28	0.14	0.17	0.22	-0.09
O3 average	0.04	0.28	1.00	-0.20	0.30	0.14	0.20
Population	157,83.16	0.14	-0.20	1.00	-0.33	-0.15	-0.34
ACRatePer100K	4,855.06	0.17	0.30	-0.33	1.00	0.72	0.38
CVRatePer100K	1,614.13	0.22	0.14	-0.15	0.72	1.00	0.19
ExtRatePer100K	137.28	-0.09	0.20	-0.34	0.38	0.19	1.00

Table 10.4 County-specific average PM2.5 concentration is significantly positively associated with county-specific CVD mortality rates across all age categories and years

	Regression summary for dependent variable: CVRatePer100K R = 0.78 R ² = 0.605 adjusted R ² = 0.605						
	N = 21,613	b*	Std. err. of b*	b	Std. err. of b	t(21608)	p-value
Intercept				11,4927.7	7043.87	16.3160	0.000000
Year	-0.08	0.0046		-60.3	3.51	-17.2049	0.000000
Age	0.81	0.0048		120.0	0.70	170.4133	0.000000
pm25Average	0.04	0.0047		33.6	3.54	9.4979	0.000000
Population	0.08	0.0048		0.0	0.00	17.4577	0.000000

Table 10.4 shows the results of a multiple linear regression with backward stepwise variable selection; results were also confirmed in multiple disjoint random samples (20% cross-validation samples). The b^* column contains standardized regression coefficients (scaling each variable in terms of standard deviations) and the b column contains the unstandardized regression coefficients. As expected, *Year* is negatively associated with CVD mortality risk, and *Age* is positively associated with CVD mortality risk. *Age* is quantitatively by far the most important predictor of risk. PM2.5 average concentration makes the smallest, but still highly statistically significant ($p < 0.000001$), contribution to predicting CVD values. Population (specific to each county and age group) is also a significant predictor of CVD risk. Results for all-disease-related mortality (AC) risks are similar, with the standardized regression coefficient for PM2.5 increasing to 0.06, with the exception that both ozone (O3) and population size are significantly negatively associated with AC mortality rates (standardized regression coefficients of -0.12 for Population and -0.02 for O3). Interpretively, the coefficient for PM2.5 in Table 10.4 ($b = 33.6$) indicates that CVD mortality risk increases by 33.6 deaths per 100,000 person-years for each microgram per cubic meter increase of PM2.5 in air, assuming other variables are held constant. The mean CVD mortality rate averaged over all age categories and years is 1931.6 deaths per 100,000 person-years, so a change in PM2.5 of 10 $\mu\text{g}/\text{m}^3$ corresponds to a change in CVD mortality rate of approximately $(10 \mu\text{g}/\text{m}^3) * (33.6 \text{ deaths per } 100,000 \text{ person-years per } \mu\text{g}/\text{m}^3) / (1931.6 \text{ deaths per } 100,000 \text{ person-years}) = 336/1931.6 = 17.4\%$. This slope factor could be described as a 17.4% increase in mortality per 10 $\mu\text{g}/\text{m}^3$ increase in PM2.5 concentration.

Results on Correlations Between Changes in Variables over Time

Tables 10.5 and 10.6 show correlations between *changes* in all-cause mortality, CVD mortality, and non-disease mortality, respectively (the columns) and different possible predictors (the rows), for all counties included in the study. Table 10.5 presents results for the 75–84 year-old group, and Table 10.6 repeats the analysis for all age groups.

Table 10.5 Pearson correlations between changes in variables from 2000 to 2010 for elderly (75–84 year-old) people

Variable	Correlates of changes in mortality rates from 2000 to 2010 for 75–85 year-olds		
	delta AC mortality	delta CVD mortality	delta external rate
delta PM2.5	-0.07	-0.08	0.04
delta O3	0.03	0.03	0.06
delta population	-0.59	-0.56	-0.44
delta AC mortality	1.00	0.99	0.81
delta CVD mortality	0.99	1.00	0.79
delta External rate	0.81	0.79	1.00
PM2.5 average	-0.10	-0.12	-0.10
O3 average	-0.04	-0.04	-0.18
Population 2010	0.07	0.07	0.06
ACRatePer100K	-0.05	-0.07	-0.07
CV RatePer100K	0.06	0.06	-0.02
ExtRatePer100K	-0.16	-0.17	0.14

Table 10.6 Pearson correlations between changes in variables from 2000 to 2010 for all age groups

Variable	Correlates of changes in mortality rates from 2000 to 2010		
	delta AC mortality	delta CVD mortality	delta External rate
delta PM2.5	-0.00	-0.01	0.06
delta O3	0.06	0.08	0.02
delta population	-0.15	-0.15	-0.17
delta AC mortality	1.00	0.95	0.69
delta CVD mortality	0.95	1.00	0.59
delta External rate	0.69	0.59	1.00
Age 2000	-0.03	-0.14	0.24
PM2.5 average 2000	-0.02	-0.04	-0.04
O3 Average 2000	-0.05	-0.07	-0.05
ACRatePer100K 2000	0.17	-0.01	0.40
CV RatePer100K 2000	0.23	0.02	0.40
ExtRatePer100K 2000	0.21	-0.03	0.56

For the 75–85 year old age category, changes in all-cause (AC) and CVD mortality rates are significantly positively correlated with each other, as expected, and with changes in external-cause mortality rates. They are significantly negatively correlated with increases in population. Neither is significantly correlated with changes in PM2.5 or changes in O3. For all age groups, changes in PM2.5 are significantly but weakly positively correlated with changes in external-cause mortality rates. Changes in O3 are significantly positively correlated both with changes in AC mortality rates and with changes in CVD mortality rates. Increases in population are significantly correlated with reductions in all mortality rates.

In multivariate analysis using multiple linear regression, changes in both AC (all-cause) and CVD mortality rates are conditionally independent of changes in both PM2.5 and O₃, given changes in population size, changes in external-cause mortality rates, and age in 2010. These three explanatory variables are automatically selected by backward stepwise variable selection, while changes in PM2.5 and O₃ are dropped, as they provide no additional information useful for predicting the AC or CVD mortality rates. Thus, by this criterion, changes in PM2.5 and O₃ levels do not help to predict or explain changes in CVD or AC mortality rates, undermining a causal interpretation of the positive associations between them in the cross-sectional analysis in Table 10.3.

Other, perhaps unexpected, correlations between changes in variables in Table 10.6 include a strong positive correlation (0.59) between changes in external-cause mortality rates and changes in CVD mortality rates; and positive correlations between baseline levels of mortality rates and changes in their levels. Thus, relatively high-risk areas in 2000 tended to become more risky by 2010. As expected, older age categories saw relatively large reductions in disease mortality rates (but increases in non-disease mortality rates).

Granger Causality Test and Control Results

Granger tests using standard time series regression models with maximum lags of 1, 2, or 3 years show that, for all age categories tested (65–74, 75–84, and 85 or older) and for all mortality outcomes considered (CVD, all-disease, and external-cause mortality rates), both PM2.5 and O₃ histories are not useful for predicting mortality rates in most (over 90%) of the counties. PM2.5 and O₃ have predictive coefficients for CVD and all-disease mortality rates that are significantly different from zero in only a small minority of counties (7% for AC mortality, 6% for CVD mortality, and 7% for external-cause mortality, which was used as a negative control), roughly consistent with, though slightly higher than, the 5% false-positive error rate that might occur by chance due to the 5% significance level used in the tests. (For 483 counties and a true false-positive rate of 5%, there is about a 26% probability that the sample proportion of false positives would exceed 6% or be less than 4% by chance.) Perhaps more importantly, the negative control (external-cause mortalities) also shows that O₃ and PM2.5 histories on time scales of several years are not Granger-causes of CVD or all disease-related deaths any more than they are of external-cause deaths. For example, the age group and lag with the highest fraction of Granger-positive associations between PM2.5 and CVD rate is the 85+ age group with a lag of 1 year: this fraction is 11%. But the corresponding fraction for Granger-positive associations between PM2.5 and external-cause mortalities is greater, at 14%. Thus, the Granger tests do not support a conclusion of a genuine causal effect, i.e., positive results clearly above what might occur by chance and what is found for the negative controls.

Table 10.7 Fractions of counties with positive Granger causality tests for PM2.5 and all-cause (AC), cardiovascular disease (CVD), and external-cause mortality rates, for different age groups and lags (1–3 years)

Age/Lag	AC mortality rate	CVD rate	External rate
65	0.06	0.06	0.07
1	0.09	0.08	0.10
2	0.04	0.05	0.10
3	0.05	0.05	0.02
75	0.08	0.06	0.06
1	0.10	0.08	0.05
2	0.08	0.06	0.08
3	0.04	0.05	0.06
85	0.08	0.06	0.08
1	0.15	0.11	0.14
2	0.06	0.03	0.09
3	0.04	0.03	0.01
Overall	0.07	0.06	0.07

Results averaged over all three lags are shown in bold

Given the well-known limitations of *p*-values and significance testing, it may also be useful to consider that, if pollutant levels were detectable causal drivers of increased mortality rates at recent historical levels, then this causal relation should have been visible in a large majority of counties. The fractions in Table 10.7 might all be expected to exceed 50% in the presence of clear Granger-causality, i.e., most counties should have shown evidence of a Granger-positive association between PM2.5 and mortality rates caused by them. Intuitively, as suggested by Fig. 10.1, although pollutant levels declined substantially in most counties from 2000 to 2010, declines in CVD and AC mortality rates did not appear to proceed more quickly when PM2.5 declined quickly than when it did not, or than when it increased. The Granger test results confirm this suggestion at the level of individual counties and for time lags of 1–3 years.

Positive Controls: Does Absence of Evidence Constitute Evidence of Absence?

Might the absence of a significant association between county-specific changes in PM2.5 levels and changes in mortality rates between 2000 and 2010, shown in Tables 10.5 and 10.6 and in corresponding multiple linear regression models, be due to limited statistical power to detect changes in the presence of substantial heterogeneity and variability in the data? To check the statistical power of these methods, we modified the observed data by adding a known “signal”—a 2.6% decrease in CVD mortality rate per $\mu\text{g}/\text{m}^3$ decrease in PM2.5 concentration, based on the slope estimate of Lepeule et al. (2012). We then tested whether this known signal is detectable through the noise in the data using the methods we have applied.

Table 10.8 shows the results of multiple linear regression applied to the artificial data set with a simulated known causal impact of exposure. The simulated effect of

Table 10.8 Multiple linear regression detects PM2.5 effects on mortality rates of the sizes predicted from previously published regression slope coefficients (Lepeule et al. 2012)

Regression for CVD mortality rate with simulated effect of P M2.5 R = 0.76465 R ² = 0.58469 adjusted R ² = 0.5838 F(3,1421) = 666.84 p < 0.0000 Std. error of estimate: 1178.0						
N = 1425	b*	Std. err. of b*	b	Std. err. of b	t(1421)	p-value
Intercept			541.5	74.8	7.2	0.000
delta PM2.5	0.04	0.017	37.9	16.5	2.3	0.022
CVRatePer100K_2000	-0.75	0.017	-0.5	0.0	-43.9	0.000
delta population	-0.16	0.017	-0.0	0.0	-9.2	0.000

changes in PM2.5 on changes in CVD mortality rates, based on the 2.6% slope coefficient for change in mortality rate per $\mu\text{g}/\text{m}^3$ change in PM2.5) was successfully detected. (All predictors remain significant using backward stepwise variable selection.) This suggests that an effect of this size would probably have been detected in the real data if it had been present. This type of positive control gives some reassurance that the substantial variability and heterogeneity in county-level time series data would not hide causal effects of the sizes that have sometimes been estimated from standard associational (regression-based) models by assuming that slope coefficients are causal, if such causal effects were actually present.

Finally, we briefly examined the geographic distribution of associations. Previous investigators have reported that chronic exposure to PM2.5 is associated with mortality in the eastern and central regions of the United States, but not in the western region (Zeger et al. 2008). In our data set, for the main elderly population (75–84 year-olds) in 2010, PM2.5 was statistically significantly positively correlated with CVD mortality in Florida and overall in pooled data from counties in all states. It was statistically negatively correlated with all-disease (AC) mortality rate in Arizona and statistically positively correlated with AC mortality rate in Florida and overall. Otherwise, state-specific correlations in 2010 were not individually statistically significant at the conventional 0.05, significance level, and were a mix of non-significant positive and negative correlations with no obvious geographic distribution.

Discussion and Conclusions: Caveats for Causal Interpretations of Regression Coefficients

The epidemiological and risk assessment literatures on human health effects of air pollution contain dozens of studies that attribute reductions in mortality risks to reductions in air pollution levels, and that estimate the slope of the concentration-response association between exposures to pollutants and corresponding mortality rates (e.g., Pope 1989; Clancy et al. 2002; Lepeule et al. 2012; Cesaroni et al. 2013; Fann et al. 2012; Dai et al. 2014). The work reported here contributes a new data set

to this literature. It supports previous findings of positive PM2.5-mortality associations, based on PM2.5 (and O₃) and age-specific mortality data, based on county-level data from the 15 largest U.S. states over the years from 2000 to 2010. Confirming earlier studies such as Lepeule et al. (2012), we found a statistically significant positive association between PM2.5 (and also O₃) concentrations and both all-disease related and CVD mortality rates, as well as a significant positive association between O₃ and external-cause mortalities, which we used as a negative control (Tables 10.3 and 10.4).

However, such associations between historical levels of exposure and response variables do not necessarily describe predictive or manipulative *causal* relations. In our examination of historical changes in pollutant levels and mortality rates (Tables 10.5 and 10.6 and multiple regression models and Granger causality tests), actual *changes* in PM2.5 and O₃ levels over time did not significantly help to predict or explain corresponding observed *changes* in all-disease or CVD mortality rates over time. This argues against facile causal interpretations of the significant statistical associations between pollution levels and mortality rates. Such causal interpretations of slope coefficients are commonly made in air pollution health effects (and other) epidemiology. For example, the study of Lepeule et al., updating the important Harvard Six Cities Study, offers the important causal interpretation that “These results [i.e., that each 10 µg/m³ increase in PM2.5 was associated with a 26% increase in cardiovascular mortality risk] suggest that further public policy efforts that reduce fine particulate matter air pollution are likely to have continuing public health benefits.” But, as emphasized in Chap. 2, such policy-relevant causal conclusions are unwarranted if the exposure-response association discussed is not a causal relation, and if the changes referred to are only the hypothetical ones implied by a slope coefficient, rather than actual changes in the levels of exposure and mortality time series.

Study Limitations

The study and conclusions in this chapter have several limitations. Although our analysis of county-level data does not provide evidence that the roughly 30% reduction in PM2.5 levels from 2000 to 2010 (Fig. 10.1) caused any detectable effect on disease-related mortality rates, it remains possible that such an effect was present that is too small to detect. For example, if each 10 µg/m³ change in PM2.5 concentration causes only a 1.03% change in CVD mortality rate, as estimated by Dai et al. (2014), then the power of our data set would not be great enough to distinguish this from zero. In addition, like many other studies, our analysis lacked individual-level exposure data. Our basic units of observation are death counts, by cause, within age categories, years, and counties; finer resolution would require a different data set. Age and death are available at the individual level, making this a semi-individual design (Künzli and Tager 1997), rather than a purely ecological design; but other individual covariates are not available. On the other hand, the fact that we follow the same counties over multiple years contributes one of the strengths

of a panel study design: the effects of fixed (or slowly changing) possible confounders or effect modifiers, such as differences in income or education or regional climate, cancel out when changes (deltas) in mortality rates are calculated for the same locations in successive years. In addition, our study substantially meets several criteria for useful ecological studies (Savitz 2012): marked variation across geographic units (counties); unlikely confounding (due to the longitudinal panel design, in which counties serve as their own controls for purposes of subtracting out fixed effects of confounders when computing changes over time); opportunities to include negative controls (external-cause mortalities); and simulated positive controls (via simulation of postulated causal impacts).

A remaining question is, if the significant associations between PM2.5 and O₃ on the one hand and CVD and all-cause mortality on the other are not due to a causal relationship between pollutant exposure and disease, then what does explain them? Our analyses have ruled out coincident trends (since the associations hold even within single years) and chance (since the correlation and regression coefficients reported are statistically significant), as well as fixed confounders (due to the panel design) as plausible explanations. Possible confounders that might co-vary with exposure levels over time, and thus offer explanations, range from co-pollutants to lagged daily temperatures (e.g., if very hot or very cold areas have higher levels of PM2.5, perhaps due in part to coal-fired power plants that power air conditioning or heating, and independently have higher mortality rates). Attaching more variables to the county-specific mortality rate and pollution level data, such as daily temperature (high and low), could potentially help to answer this question. But at present, the answer is unknown.

Finally, by focusing on changes in annual average pollutant levels and mortality rates at the individual county level, we have foregone opportunities to model or “adjust” for effects of seasonality, more granular spatial variations, and measured or latent confounders. As Dominici et al. (2014) suggest, it is not uncommon for different regression models based on different modeling choices and assumptions to produce very different answers. For example, regression coefficients that are significantly positive in one model may be significantly negative in another, depending on which variables and interaction terms are included. By using several different approaches (conditional independence tests, Granger tests, positive and negative controls, automated variable selection) as well as relatively simple measures of association (correlations and linear regression coefficients, fractions of counties with Granger-positive associations) computed using standard, widely available software for all tests), we have sought conclusions that are more robust and objective by minimizing opportunities for manual intervention to shape the results.

Comparisons to Conclusions from Other Studies

The coefficient for PM2.5 in Table 10.4 ($b = 33.6$), corresponding to a 17% increase in mortality per 10 µg/m³ increase in PM2.5 concentration, is well within the range of other recent association-based estimates based on regression relations. For

example, Dai et al. (2014), in a study of 75 U.S. cities between 2000 and 2006, reported a 1.03% (95% CI: 0.65%, 1.41%) increase in CVD mortality with each $10 \mu\text{g}/\text{m}^3$ increase in PM2.5, averaged over a 2-day period. In their update of the Harvard Six Cities Study, Lepeule et al. (2012) estimated a 26% (95% CI: 14%, 40%) increase in CVD mortality for each $10 \mu\text{g}/\text{m}^3$ increase in PM2.5, averaged over the three prior years. Thus, our value of 17.4% falls between these two estimates, and is within the 95% CI of the Lepeule study. Some other recent studies have not detected clearly significant associations between PM2.5 levels and most CVD or all-cause mortality rates (Beelen et al. 2014), or found no association between local trends in mortality and local trends in yearly average PM2.5 after adjusting for national trends and local differences (Greven et al. 2011). For the U.S. county data set we have analyzed, our main conclusions are that (a) There are statistically significant associations between PM2.5 and both all-disease and CVD mortality risks; but (b) There is no clear evidence of a causal relation between PM2.5 and O₃ concentration levels and mortality rates. These results differ both from studies that do not find clear associations, and also from some authoritative opinions, including views in an Expert Elicitation Study for the U.S. EPA (2006), that statistically significant exposure-response associations between PM2.5 and CVD mortality are probably causal.

While our results do not support some previous expert judgment-based assessments of causality, this is consistent with studies showing that firmly expressed opinions of key experts about air pollution health effects associations being causal (e.g., Harvard School of Public Health 2002) have later proved to be unwarranted (e.g., HEI 2013). The practice of applying human judgment using weight-of-evidence considerations to measures of association (such as relative risks, odds ratios, population attributable fractions, burden-of-disease estimates, and regression coefficients) to determine whether an inference of causality is supported has been widespread in epidemiology, even though some methodologists have argued that logically valid causal inferences cannot be derived from such associations in purely observational studies without interventions (Ward 2009). This makes natural experiments, where interventions such as pollution reductions occur differently for different subpopulations, potentially valuable aids to understanding causation.

The preceding calculations illustrate that a significant positive association between historical levels of PM2.5 and historical mortality rates does not necessarily provide a sound basis for inferring a positive association between changes in levels of PM2.5 and changes in mortality rates. This methodological point confirms the importance of using quasi-experiments or other appropriate formal methods of causal study design and analysis (Table 10.2) to draw causal conclusions. Free, publicly available data sets such as the EPA and CDC data sets used in this study, and free, publicly available software such as R and Python or the CAT software from Chapter 2, now make it relatively easy to test whether changes in PM2.5 and O₃ help to predict changes in disease mortality rates, on time scales from days to over a decade. We hope that this will encourage others to investigate further the relation between longitudinal changes in pollutant levels and changes in mortality rates, and to clarify the crucial distinction between positive statistical associations and evidence of causality in air pollution health effects epidemiology.

References

- Angrist JD, Pischke J-S (2009) *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press, Princeton
- Beelen R, Stafoggia M, Raaschou-Nielsen O et al (2014) Long-term exposure to air pollution and cardiovascular mortality: an analysis of 22 European cohorts. *Epidemiology* 25(3):368–378
- Buka I, Koranteng S, Osornio-Vargas AR (2006) The effects of air pollution on the health of children. *Paediatr Child Health* 11(8):513–516
- Campbell DT, Stanley JC (1966) Experimental and quasi-experimental designs for research. Rand McNally, Chicago
- Centers for Disease Control and Prevention (CDC) (2014) Wonder “compressed mortality, 1999–2010” database. <http://wonder.cdc.gov/cmf-icd10.html>
- Cesaroni G, Badaloni C, Gariazzo C, Stafoggia M, Sozzi R, Davoli M, Forastiere F (2013) Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environ Health Perspect* 121(3):324–331
- Clancy L, Goodman P, Sinclair H, Dockery DW (2002) Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 360(9341):1210–1214
- Cox LA Jr (2013) Improving causal inference in risk analysis. *Risk Analysis* 33(10):1762–1771
- Dai L, Zanobetti A, Koutrakis P, Schwartz JD (2014) Associations of fine particulate matter species with mortality in the United States: a multiplicity time-series analysis. *Environ Health Perspect* 122 (8):837–842
- Dominici F, Greenstone M, Sunstein CR (2014) Particulate matter matters. *Science* 344 (18):257–258
- Eichler M, Didelez V (2010) On Granger causality and the effect of interventions in time series. *Lifetime Data Anal* 16(1):3–32
- EPA (2006) Expanded expert judgment assessment of the concentration-response relationship between PM_{2.5} exposure and mortality. www.epa.gov/tn/ecas/regdata/Uncertainty/pm_ee_report.pdf
- EPA (U.S. Environmental Protection Agency) (2011) The benefits and costs of the Clean Air Act from 1990 to 2020. Final report—Rev. A. Office of Air and Radiation, Washington
- Fann N, Lamson AD, Anenberg SC, Wesson K, Risley D, Hubbell BJ (2012) Estimating the national public health burden associated with exposure to ambient PM_{2.5} and ozone. *Risk Anal* 32(1):81–95
- Freedman DA (2004) Graphical models for causation, and the identification problem. *Eval Rev* 28 (4):267–293
- Friedman N, Goldszmidt M (1998) Learning Bayesian networks with local structure. In: Jordan MI (ed) *Learning in graphical models*. MIT Press, Cambridge, pp 421–459
- Friedman MS, Powell KE, Hutwagner L, Graham LM, Teague WG (2001) Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma. *JAMA* 285(7):897–905
- Gilmour S, Degenhardt L, Hall W, Day C (2006) Using intervention time series analyses to assess the effects of imperfectly identifiable natural events: a general method and example. *BMC Med Res Methodol* 6:16
- Greenland S, Brumback B (2002) An overview of relations among causal modelling methods. *Int J Epidemiol* 31(5):1030–1037
- Greven S, Dominici F, Zeger S (2011) AN approach to the estimation of chronic air pollution health effects using spatio-temporal information. *J Am Stat Assoc* 106(494):396–406
- Hack CE, Haber LT, Maier A, Shulte P, Fowler B, Lotz WG, Savage RE Jr (2010) A Bayesian network model for biomarker-based dose response. *Risk Anal* 30(7):1037–1051
- Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, Finkelstein J (2006) The use and interpretation of quasi-experimental studies in medical informatics. *J Am Med Inform Assoc* 13(1):16–23
- Harvard Law Today (2014) <http://today.law.harvard.edu/improving-the-pollution-mortality-link/>

- Harvard School of Public Health (2002) Press release: “ban on coal burning in dublin cleans the air and reduces death rates”. www.hsppharvard.edu/news/press-releases/archives/2002-releases/press10172002.html
- Health Effects Institute (HEI) (2010) Impact of improved air quality during the 1996 Summer Olympic Games in Atlanta on multiple cardiovascular and respiratory outcomes. HEI research report #148. April, 2010. Peel JL, Klein M, Dana Flanders W, Mulholland JA, Tolbert PE. Health Effects Institute, Boston. <http://pubs.healtheffects.org/getfile.php?u=564>
- Health Effects Institute (HEI) (2013) Did the Irish coal bans improve air quality and health? HEI Update, Summer, 2013. <http://pubs.healtheffects.org/getfile.php?u=929>. Accessed 1 Feb 2014
- Helfenstein U (1991) The use of transfer function models, intervention analysis and related time series methods in epidemiology. *Int J Epidemiol* 20(3):808–815
- Hernán MA, Robins JM (2018) Causal inference. Chapman and Hall/CRC, Boca Raton. Forthcoming. See <https://www.hsppharvard.edu/miguel-hernan/causal-inference-book/>
- Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58 (5):295–300
- Imberger G, Vejlby AD, Hansen SB, Møller AM, Wetterslev J (2011) Statistical multiplicity in systematic reviews of anesthesia interventions: a quantification and comparison between Cochrane and non-Cochrane reviews. *PLoS One* 6(12):e28422
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8):e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Kelly F, Armstrong B, Atkinson R, Anderson HR, Barratt B, Beevers S, Cook D, Green D, Derwent D, Mudway I, Wilkinson P, HEI Health Review Committee (2011) The London low emission zone baseline study. *Res Rep Health Eff Inst* 163:3–79
- Krstić G (2011) Apparent temperature and air pollution vs. elderly population mortality in Metro Vancouver. *PLoS One* 6(9):e25101
- Künzli N, Tager IB (1997) The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies. *Environ Health Perspect* 105(10):1078–1083
- Lamm SH, Hall TA, Engel E, White LD, Ructer FH (1994) PM 10 particulates: are they the major determinant in pediatric respiratory admissions in Utah County, Utah (1985–1989)? *Ann Occup Hyg* 38:969–972
- Lehrer J (2012) Trials and errors: why science is failing us. *Wired*. <http://www.wired.co.uk/magazine/archive/2012/02/features/trials-and-errors?page=all>
- Lepeule J, Laden F, Dockery D, Schwartz J (2012) Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities Study from 1974 to 2009. *Environ Health Perspect* 120(7):965–970
- Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 21(3):383–388
- MacLure M (1991) Taxonomic axes of epidemiologic study designs: a refutationist perspective. *J Clin Epidemiol* 44(10):1045–1053
- Moore KL, Neugebauer R, van der Laan MJ, Tager IB (2012) Causal inference in epidemiological studies with strong confounding. *Stat Med* 31(13):1380–1404. <https://doi.org/10.1002/sim.4469>
- NHS (2012) Air pollution ‘kills 13,000 a year’ says study. www.nhs.uk/news/2012/04april/Pages/air-pollution-exhaust-death-estimates.aspx
- Ottenbacher KJ (1998) Quantitative evaluation of multiplicity in epidemiology and public health research. *Am J Epidemiol* 147:615–619
- Pelucchi C, Negri E, Gallus S, Boffetta P, Tramacere I, La Vecchia C (2009) Long-term particulate matter exposure and mortality: a review of European epidemiological studies. *BMC Public Health* 9:453
- Pope CA 3rd (1989) Respiratory disease associated with community air pollution and a steel mill, Utah Valley. *Am J Public Health* 79(5):623–628
- Powell H, Lee D, Bowman A (2012) Estimating constrained concentration-response functions between air pollution and health. *Environmetrics* 23(3):228–237

- Rothman KJ, Greenland S (2005) Causation and causal inference in epidemiology. *Am J Public Health* 95 Suppl 1:S144–S150
- Sarewitz D (2012) Beware the creeping cracks of bias. *Nature* 485:149
- Savitz DA (2012) Commentary: a niche for ecologic studies in environmental epidemiology. *Epidemiology* 23(1):53–54
- Stebbins JH Jr (1978) Panel studies of acute health effects of air pollution. II. A methodologic study of linear regression analysis of asthma panel data. *Environ Res* 17(1):10–32
- The Economist (2013) Trouble at the lab: scientists like to think of science as self-correcting. To an alarming degree, it is not. www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
- U.S. Environmental Protection Agency (EPA) (2014) www.epa.gov/airdata/ad_data_daily.html
- Ward AC (2009) The role of causal criteria in causal inferences: Bradford Hill's "aspects of association". *Epidemiol Perspect Innov* 6:2
- Wittmaack K (2007) The big ban on bituminous coal sales revisited: serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black-smoke exposure. *Inhal Toxicol* 19(4):343–350
- Yang Y, Li R, Li W, Wang M, Cao Y, Wu Z, Xu Q (2013) The association between ambient air pollution and daily mortality in Beijing after the 2008 Olympics: a time series study. *PLoS One* 8(10):e76759
- Yim SH, Barrett SR (2012) Public health impacts of combustion emissions in the United Kingdom. *Environ Sci Technol* 46(8):4291–4296
- Zeger SL, Dominici F, McDermott A, Samet JM (2008) Mortality in the Medicare population and chronic exposure to fine particulate air pollution in urban centers (2000–2005). *Environ Health Perspect* 116(12):1614–1619. <https://doi.org/10.1289/ehp.11449>

Chapter 11

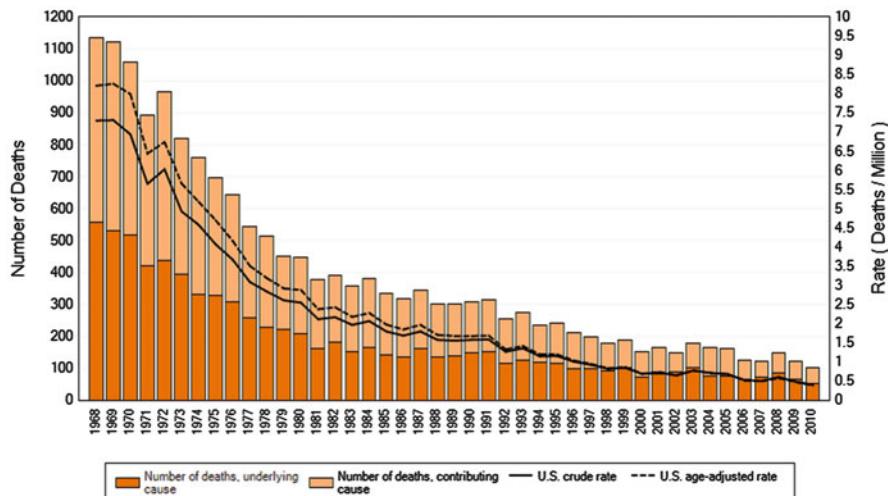
Evaluation Analytics for Occupational Health: How Well Do Laboratories Assess Workplace Concentrations of Respirable Crystalline Silica?



Introduction

Chapters 8 and 10 have introduced important themes of *evaluation analytics*: discovering through independent replication of previous work (Chap. 8) and by applying new methods such as modern predictive and causal analytics algorithms to previously collected observational data (Chap. 10) whether published claims are reproducible and whether predicted effects caused by changes in exposures have actually occurred. This chapter provides an example of evaluation analytics in a context where experimentation is possible. It illustrates how a designed experiment with samples having known properties can be used to evaluate how consistently and accurately the laboratory system used to assess compliance of workplaces with occupational safety standards for respirable crystalline silica (RCS) performs in correctly classifying exposure concentrations as above or below a desired level. In this context, causation appears to be clear: concentrations of RCS in air lead to concentrations on air filters sent to laboratories. However, as we shall see, there is enough unexplained noise or random variation in the process so that even control samples with no RCS are sometimes mistakenly identified as carrying significant positive loads of RCS (Cox et al. 2015). Thus, the causes of laboratory-reported values include substantial contributions from measurement errors.

Respirable crystalline silica (RCS), which consists of minute quartz particles (sand), causes increased risk of silicosis in people or animals exposed to sufficiently high concentrations for sufficiently long durations. Chapter 9 discussed some of the key causal mechanisms involved. To reduce or eliminate risk of silicosis in occupationally exposed workers, regulators, employers, employees, and labor organizations have worked together to reduce greatly exposure concentrations of RCS in the air of many workplaces since the 1960s. Figure 11.1 shows that, as hoped, silicosis mortalities in the U.S. have declined dramatically, by over 90% since the 0.10 mg/m³ PEL was established to protect worker health. Evidence from clinical, toxicological,



Source: U.S. Centers for Disease Control and Prevention, 2016

Fig. 11.1 Annual silicosis-associated mortalities have decreased by over 90% since the 0.10 mg/m³ PEL was established. Source: U.S. Centers for Disease Control and Prevention, 2016

epidemiological, and industrial hygiene studies, as well as the historical record in Fig. 11.1, suggest that it has been effective in doing so.

Despite this dramatic progress, silica exposures in some workplaces in recent decades have remained far above the current (and former) PEL levels. Such lack of compliance with the 0.10 mg/m³ PEL can harm human health. As noted by the Centers for Disease Control and Prevention (CDC) (64 MMWR 23, June 19, 2015):

“Results indicate that despite substantial progress in eliminating silicosis, silicosis deaths continue to occur. Of particular concern are silicosis deaths in young adults (aged 15–44 years). These young deaths likely reflect higher exposures than those causing chronic silicosis mortality in older persons, some of sufficient magnitude to cause severe disease and death after relatively short periods of exposure. A total of 12 such deaths occurred during 2011–2013, with nine that had silicosis listed as the underlying cause of death.”

From a risk management perspective, it is natural to wonder whether it is possible to further reduce silicosis mortalities and morbidities through improved compliance monitoring and enforcement and/or by reducing the current PEL. Would a lower PEL prevent more deaths, or, to the contrary, would rigorously enforcing the current 0.1 mg/m³ limit achieve all the human health benefits available from reducing exposures? To find out, it is important to consider how accurately workplace exposures are currently monitored and enforced and the possibilities for improving compliance with PELs. A regulation that tells employers “Do not exceed concentration C” can only be effective if it is possible to determine, with useful reliability, when a workplace is in compliance. As permitted concentrations become lower, determining compliance can become more difficult, requiring increasingly accurate

and precise laboratory measurements. Enjoining employers to comply with standards for which compliance cannot be determined reliably risks ineffective action, diverting limited resources to false positives (acting to reduce workplace exposures based on false findings that they are too high) or failing to act due to false negatives (i.e., laboratory findings mistakenly indicating that no action is needed). Such errors and potential for useless activity arise whenever occupational exposures are reduced until they are comparable to the “noise,” or random measurement error and variability, in laboratory results—or, conversely, whenever the random variability in laboratory results is large enough to obscure the effects that they seek to detect. The science-policy question of how best to set PEL concentrations to protect worker health then becomes complicated by the practical reality that compliance with target concentration levels cannot easily be determined.

The remainder of this chapter examines how well today’s workplace RCS concentrations can be determined from the laboratory results that are currently used in determining and enforcing compliance. A 2015 article described an experiment in which filters with known loads of RCS in different matrices (and also as pure crystalline silica samples) were sent to five different commercial laboratories to determine how accurately and reliably they measured these known loads (Cox et al. 2015). The overall results were striking: the laboratories did not consistently discriminate between concentrations that differed by a factor of 2, and even filters with no crystalline silica load were sometimes misidentified as containing substantial RCS concentrations. However, these findings came from an artificial experiment, leaving open the possibility that laboratories perform better in practice on real samples. The analysis in this chapter tests that possibility by quantifying the variability in results from commercial laboratories, including laboratories accredited by the American Industrial Hygiene Association (AIHA) Industrial Hygiene Proficiency Analytical Testing (IH-PAT) program (www.aihapat.org). These include laboratories used by numerous employers and other organizations to determine whether workplace exposures comply with PELs.

Data and Methods

Employers and other organizations often send workplace air sampling filters to accredited laboratories to determine current workplace air concentrations of RCS. The laboratories, in turn, may use several different analytical methods, such as X-ray diffraction, infrared spectroscopy, and calorimetry to estimate the quantity of RCS on a received filter; of these, X-ray diffraction is widely considered one of the most accurate and reliable analytic methods. All data discussed in this chapter were derived by X-ray diffraction. The laboratory returns to the submitting entity the estimated quantity of RCS (mg) on the filter from a given volume of air sampled. If the values (“lab results”) returned are sufficiently high, interventions to reduce workplace air concentrations of RCS may be triggered.

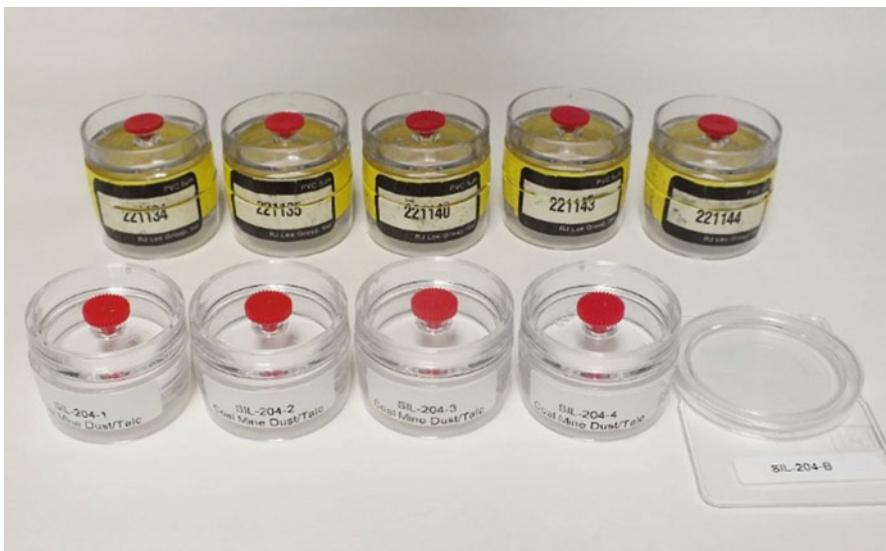


Fig. 11.2 Real-world (back row) and AIHA-prepared (front row) filters sent to laboratories for analysis

To maintain IH-PAT accreditation, laboratories must meet AIHA-specified criteria for proficiency in estimating quantities of RCS (and other substances) sent to them for analysis. This is done as follows. Four times per year (each being called a “round”), AIHA sends to each participating lab four spiked sample filters, prepared from continuously agitated homogeneous suspensions with four different known concentrations; thus the filters received by different laboratories contain approximately the same known loadings of RCS, called samples. Figure 11.2 shows a photograph of such samples (front row, clearly identifying the interfering substances) as well as typical real-world samples (back row) which display only unique sample numbers. As explained in AIHA methods documentation, the RCS analytes measured in these accreditation tests consist of “Free silica (quartz) on four 5.0 μm 37-mm PVC (polyvinyl chloride) filter samples containing differing silica concentrations and include a background matrix, on a rotating basis of coal mine dust, talc, calcite, or a combination” (AIHA 2016).

Each laboratory analyzes the IH-PAT-prepared samples and reports the estimated quantities of RCS back to IH-PAT. The reported results from different laboratories for corresponding samples in the same round should be the same if there are no errors or variability in the process. In practice, as discussed further in the Results section, there is considerable variability in the estimated quantities of RCS for these matched samples (although less than in earlier decades) consistent with previous literature (Harper et al. 2014; Shulman et al. 1992; Abell and Doemeny 1991). The key statistical criterion for determining a laboratory’s proficiency is that its results must fall within three standard deviations of the IH-PAT “assigned value” (or “reference mean”). A laboratory is rated non-proficient if it has failing scores

in 2 of the last 3 consecutive rounds (i.e., 2 of the last 12 consecutive test samples). To determine the assigned value for a given sample in a given round, IH-PAT first identifies a subgroup of the participating labs, termed reference labs, based on prior analytical performance. Using the reference labs' reported results, IH-PAT Winsorizes any outliers (replacing them with less extreme values, see <https://cran.r-project.org/web/packages/robustHD/robustHD.pdf>) and calculates arithmetic means and standard deviations of the reference lab results.

As described by OSHA (<https://www.osha.gov/dsg/etools/silica/faq/faq.html>, accessed 6/30/2016):

The PAT program is designed to help consumers select laboratories that are proficient. In the PAT program analyses of quartz, the “true” values against which a laboratory’s results are compared are based on results from reference laboratories that are a subset of the participating laboratories. Assuming that the PAT samples were made from accurately delivered consensus reference material and that the participants all used the same techniques, instrumentation and methodology, and that the samples are not otherwise flawed so as to introduce bias, the best accuracy that can be achieved by consensus analyses is limited by the standard error of the precision of that analysis [$SD/(n)^{1/2}$, where SD is the standard deviation in the results among the n reference labs]. . . . The current method of PAT quartz sample generation is by aerosol generation using “5 micron” Min-U-Sil 5 without cyclones. In addition to any errors in the generation process, this “total dust” approach introduces a sampling error that may not duplicate the sampling error associated with the use of a cyclone.

In the PAT program, these generation and sampling errors are recognized as significant and are evaluated in statistical tests conducted on sub-batches and batches of PAT samples by the contract laboratory that prepares them. . . . The results obtained by participants in the PAT program therefore include both the analytical error the participating laboratories introduce and an unknown but potentially large amount of error introduced in the generation and sampling of the aerosol. These latter errors may vary batch to batch.”

Multiple years of data on the estimated quantities of RCS returned by different laboratories in response to the spiked samples sent out via the AIHA-PAT program are available on-line in .pdf format at www.regulations.gov/#/documentDetail;D=OSHA-2010-0034-4188 from an AIHA-PAT program submission to OSHA; they are also available as Excel files from the present author. Table 11.1 shows the layout of the data used in all subsequent analyses. The “OrgId” code is a numerical code that uniquely identifies each laboratory (based on our re-coding of the original much longer codes). Data from 26 AIHA-PAT accredited labs (one per row) and for the most recent 2 rounds for which data are available (one per column) are shown, since the most recent data are assumed to be most representative and relevant for current testing conditions. Round 194 took place in July of 2013. (Including more years of data reinforces our findings, but risks losing relevance; Harper et al. 2014, discuss key changes over time in the IH-PAT program.) The numbers in Table 11.1 represent estimated quantities (mg) of RCS returned by each lab (row) for each sample and round (column); empty cells indicate missing values, e.g., because a lab dropped out of the AIHA-PAT program.

In analyzing these data, we emphasized exploratory and descriptive data analysis and non-parametric methods to avoid introducing potentially erroneous and biased modeling assumptions. The following sections present results of statistical plots of

Table 11.1 Data layout for AIHA-PAT data

	1	2	3	4	5	6	7	8	9
OrgId	SampleValue 193 01	SampleValue 193 02	SampleValue 193 04	SampleValue 193 04	SampleValue 194 01	SampleValue 194 02	SampleValue 194 03	SampleValue 194 03	SampleValue
1	1	0.08	0.06	0.21	0.09	0.08	0.09	0.08	0.14
2	4	0.08	0.06	0.22	0.12	0.12	0.15	0.08	0.20
3	5				0.11	0.11	0.15	0.08	0.18
4	7	0.09	0.07	0.21	0.11	0.12	0.12	0.07	0.17
5	8	0.10	0.07	0.21	0.11	0.10	0.13	0.06	0.18
6	9				0.11	0.12	0.12	0.07	0.18
7	12	0.10	0.08	0.24	0.13				
8	14	0.07	0.06	0.18	0.11	0.08	0.06	0.18	0.11
9	15	0.06	0.13	0.13	0.18	0.10	0.14	0.07	0.15
10	16	0.09	0.07	0.21	0.11	0.05	0.14	0.07	0.17
11	24	0.07	0.06	0.20	0.09	0.11	0.12	0.10	0.18
12	25	0.09	0.07	0.24	0.13	0.14	0.18	0.08	0.21
13	26	0.13	0.05	0.20	0.07	0.11	0.16	0.07	0.20
14	31	0.09	0.07	0.22	0.11	0.13	0.17	0.07	0.24
15	32	0.08	0.05	0.16	0.07	0.11	0.13	0.06	0.17
16	34	0.09	0.07	0.21	0.10	0.13	0.16	0.07	0.20
17	35	0.08	0.04	0.15	0.11	0.11	0.15	0.09	0.20
18	36	0.08	0.06	0.17	0.11	0.10	0.15	0.07	0.17
19	38	0.07	0.05	0.21	0.12	0.10	0.14	0.06	0.18
20	39	0.07	0.05	0.19	0.08	0.10	0.13	0.06	0.18
21	40	0.09	0.08	0.17	0.14	0.12	0.14	0.06	0.11
22	45	0.09	0.07	0.24	0.07	0.11	0.15	0.07	0.19
23	46					0.13	0.16	0.08	0.21
24	47	0.09	0.06	0.21	0.12	0.12	0.13	0.07	0.20
25	48	0.09	0.07	0.22	0.14	0.13	0.13	0.08	0.19
26	49					0.12	0.15	0.08	0.20

conditional empirical cumulative frequency distributions, nonparametric (smooth) regression, and Spearman's rank correlations to compare and visualize laboratory-specific results vs. approximate true values, i.e., the reference values.

Results

Figure 11.3 shows individual laboratory estimates of RCS amounts for each reference (estimated true) value. Each small circle represents one result returned by a laboratory. Reference values are on the x axis and the laboratory's corresponding estimated values are on the y axis. If all estimates were perfectly correct and there were no errors in the process, all data points would fall on the line shown (the line of perfect calibration), which equates estimated (y axis) and reference (x axis) values. In practice, individual laboratories return estimated RCS levels that vary widely around the reference values, as indicated by the vertical spread of the individual laboratory results around the line. For example, the spiked samples with a reference value (i.e., estimated true value) of just under 0.14 mg elicited individual laboratory estimates ranging from about 0.06 to over 0.18 mg. Since there are only 26 participating accredited labs in this study, this wide range implies that an employer or other entity who receives a laboratory report of, say, 0.06 or 0.09 mg (the two lowest

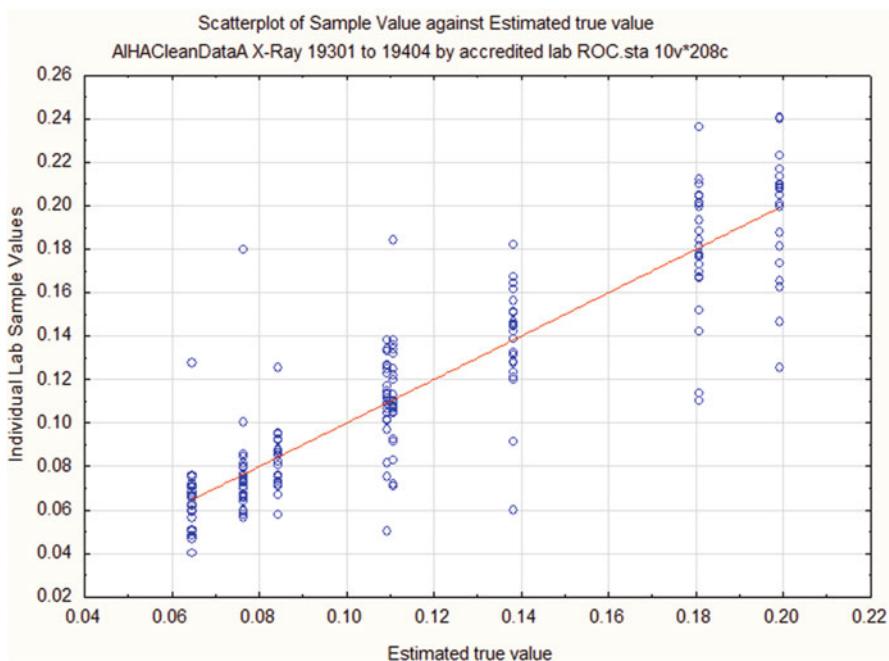


Fig. 11.3 Individual laboratory values (circles) are widely spread around reference mean values (line). All values are in mg

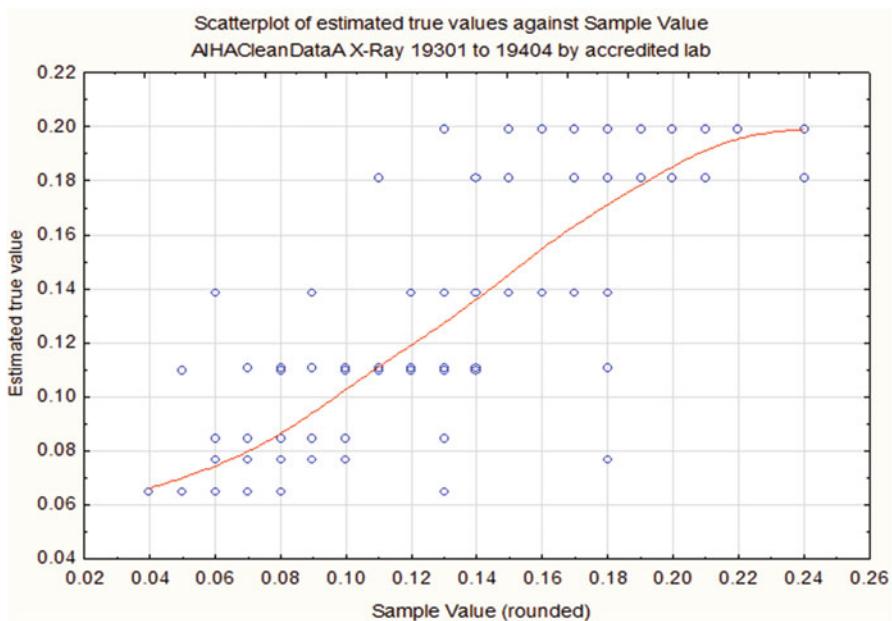


Fig. 11.4 Estimated true values (y axis) are widely spread around sample values (x axis). All values are in mg

individual laboratory estimated values for this reference value) cannot be reasonably sure (e.g., 95% confident) that the true value is less than 0.10 mg.

Conversely, the wide vertical scatter of estimates around the reference values that are below 0.10 mg implies that receiving a lab result of 0.12, or even 0.18 mg, does not imply that the true value exceeds 0.10 mg.

Figure 11.4 provides a different perspective on this variability in laboratory results by showing laboratory-estimated values (rounded to the nearest mg) on the x axis and corresponding reference values on the y axis. A nonlinear regression (nonparametric smoother) curve is fit to this scatterplot. Very high sample values returned by laboratories tend to over-estimate the reference values (e.g., a sample value of 0.24 mg returned by a laboratory corresponds to an average reference value of about 0.20 mg, as estimated by the regression curve); conversely, very low individual laboratory values tend to under-estimate corresponding reference values. There is very substantial variability in the reference values corresponding to a single estimated sample value, as shown by the vertical range of results (small circles) for specific sample values on the x axis.

The small circles in Figs. 11.3 and 11.4 show that, collectively, laboratory results are quite variable, despite the care exercised by the AIHA-PAT program in preparing spiked samples that should all yield closely similar values in the absence of laboratory error. It is natural to wonder whether this might be due to a few individual laboratories that are consistently higher or lower than the rest. To find out, we used

Spearman's rank correlations to test whether laboratories that gave higher (or lower) estimates than most others in one quarter also tended to do so over time. The test was carried out by computing the 28 ($= (8 \times 7)/2$) pairs of ordinal correlations between rankings of labs based on sizes of RCS estimates (for the same reference value) in each of the 8 rounds in Table 11.1. Six of the 28 pairs of Spearman's rank correlations differed significantly from zero at the conventional $p = 0.05$ significance level, and all six were positive (with numerical values between 0.39 and 0.73). This provides significant evidence that laboratories that give relatively high or low results compared to others in one quarter are more likely to do so again in another quarter. However, the effect is relatively small, and the wide range of variability in sample RCS estimates shown in Figs. 11.3 and 11.4 reflects variability that is more pervasive than one or a few laboratories.

Discussion

The practical implications of the variability in laboratory estimates of RCS quantities are potentially important to employers, employees, and regulators who rely on such results to determine compliance and need for interventions. If a laboratory returns an estimated value of 0.06 mg for a submitted air sample filter, for example, then Fig. 11.4 shows that the corresponding reference values (and hence the true values that they approximate) range from a low of about slightly above 0.06 mg to a high of about 0.14 mg (based on the empirically observed range of reference values that generated estimated values of 0.06 mg when sent to the 26 accredited laboratories). One of the four distinct reference values shown for an estimated value of 0.06 mg is 0.14 mg, and all four are higher than the estimated value of 0.06 mg, so there can be little confidence based on a returned value of 0.06 mg that the true value is less than a value corresponding to approximately 0.10 mg. Similarly, a laboratory estimate of 0.13 could correspond to a true value anywhere between about 0.06 mg and 0.20 mg, while an estimate of 0.18 mg could correspond to a true value anywhere between about 0.08 and 0.20 mg. Thus, no laboratory result between 0.06 and 0.18 mg can be relied on by the employer, employee, or regulator to confidently (e.g., with 95% confidence, or even 90% confidence) discriminate between true (or reference) values above and below 0.10 mg. To the contrary, returned values between 0.06 mg and 0.18 mg (spanning most of the design range for spiked sample values, which run from 0.05 to 0.20 mg) convey very little information about the probable true value, since any returned value between 0.06 mg and 0.18 mg is compatible with a wide range of true values.

That laboratories provide such relatively uninformative estimates for individual samples in no way contradicts or undermines the fact that, on average, higher sample values do indeed correspond to higher reference values, as shown by the nonparametric regression curve in Fig. 11.4. However, individual employers and employees cannot get the benefits of this useful aggregate relation, since they receive only the individual results of submitted sample filters, and these individual results are too

variable to provide trustworthy indications of whether the sampled workplaces are above or below a given limit. Acting to reduce RCS exposures (e.g., by increasing dust controls or administrative controls and use of respirators), or failing to take such measures, on the basis of laboratory-measured values does not provide a reliable approach to taking action when, and only when, appropriate.

Conclusions

Various authoritative agencies attribute great value to the information provided by laboratory analyses of RCS. For example, the US Occupational Health and Safety Administration states that “Analytical results on the quartz content of the air samples are necessary to evaluate whether the OSHA PEL is exceeded” (www.osha.gov/dsg/etools/silica/faq/faq.html). AIHA states that “The purpose of proficiency testing is to provide interested parties with objective evidence of a laboratory’s capability to produce data that is both accurate and repeatable for the activities listed in its scope of accreditation. A laboratory’s competence can be demonstrated through favorable proficiency testing data. This is important to clients, potential customers, accreditation bodies, and other external entities.” Likewise, the International Standards Organization (ISO), for which the IH-PAT program provides conformity assessment, states that “The need for ongoing confidence in laboratory performance is not only essential for laboratories and their customers but also for other interested parties, such as regulators, laboratory accreditation bodies and other organizations that specify requirements for laboratories” (ISO/IEC 2010). These statements about the importance of trustworthy laboratory performance, however true, stop short of addressing the fundamental challenge revealed by the data in this study: the variability in laboratory results is large enough compared with a 0.10 mg/m^3 limit so that results returned by laboratories do not reliably indicate whether workplace RCS concentrations are above or below that limit. The current (as of October, 2013) AIHA-PAT program considers laboratories proficient as long as their results fall within an interval of six standard deviations (three in each direction) of the mean for reference laboratories for at least 75% of the silica samples at least 2/3 of the time (i.e., in at least 2 of each 3 consecutive rounds): “IHPAT participant results are rated acceptable or unacceptable for each unique analyte sample number. . . . A passing score is 75% or more acceptable results for an analyte group. A participant is rated proficient for the applicable IHPAT analyte group if the participant has a passing score for the applicable IHPAT analyte group in two (2) of the last three (3) consecutive PT rounds.” (www.aihapat.org/Programs/IHPAT/Documents/IHPAT%20Scheme%20Plan%20R2.pdf, p.16) For the IH-PAT program data in Table 11.1, Figs. 11.3 and 11.4 show that this is not a sufficiently demanding criterion to assure usefully accurate and reliable results.

Previous research has noted that large differences among laboratories contribute to variability in PAT program estimates of the RCS content in samples (Shulman et al. 1992), although variability has declined since the introduction of the

PAT program in 1972, in part due to changes in RCS samples and in laboratory procedures (*ibid* and Harper et al. 2014). Maciejewska (2006) found that regular use of quality control methods for free silica determination was positively associated with proficiency of laboratories, suggesting that the variability of PAT estimates for RCS can potentially be reduced by such methods. Thus, although Figs. 11.3 and 11.4 show that current (2013) variability in laboratory estimates is too great to discriminate reliably among reference concentrations in the range of approximately 0.06–0.18 mg (and hence to assess compliance with PELs for workplaces with concentrations between about half and about double a PEL corresponding to 0.10 mg), it is plausible that this variability could be reduced by stricter quality control for laboratories making RCS determinations, as OSHA's 2016 final rule reducing the PEL from 010 to 0.05 mg/m³ requires, but demonstrating that it has been successfully accomplished appears to be a prerequisite for obtaining reliable results (Lee et al. 2016). More stringent statistical quality requirements may also be essential for obtaining more useful results. For example, instead of requiring only that laboratories come within three standard deviations of the reference value on 75% or more of samples in at least 2 of 3 consecutive rounds, AIHA might adopt the NIOSH criterion that “the method must provide results that are within $\pm 25\%$ of the expected (“true”) values at least 95 times out of 100” (Ashley 2015). Such accuracy would require greatly reducing the variability shown in Figs. 11.3 and 11.4, where far more than 5% of results lie further than 25% away from the expected values (given by the regression curves). This might be accomplished either by having individual laboratories make greater use of quality control measures for RCS determinations or by modifying compliance determination rules to make greater use of *averages* of RCS values assessed by multiple laboratories, to adjust for the fact that average values are relatively reliable, but individual sample values are currently too variable to meet accuracy criteria such as NIOSH's.

Our findings also have potentially important implications for monitoring and enforcement of the new (2016) OSHA limit of 0.05 mg/m³ and the new action level of 0.025 mg/m³. In general, as quantified in earlier studies (Cox et al. 2015), current laboratory procedures do not reliably quantify crystalline silica levels that differ by a factor of 2 (e.g., 0.10 vs. 0.05 mg/m³), and performance was not improved at lower exposure levels, so the information obtained from laboratories today is inadequate for monitoring and enforcing compliance even with the 0.10 mg/m³ PEL (Lee et al. 2016), and *a fortiori* for the new, lower PEL (Cox et al. 2015; Lee et al. 2016). Until laboratory practices and/or statistical protocols for determining compliance are radically improved to give demonstrably more accurate results, the current high prevalence of false-positive and false-negative conclusions about compliance implied by Fig. 11.4 of this study can be expected to inhibit effective allocation of resources to improve and protect worker health. Compliance with the recent 0.10 mg/m³ or the current 0.05 mg/m³ PEL cannot be determined reliably without substantial improvement in current laboratory and statistical practices. Until these improvements are made and credibly demonstrated, enforcement activities that are based primarily on random noise or error in laboratory results will provide neither incentives nor capability to continue reducing out-of-compliance levels of exposure that could threaten worker health.

References

- Abell MT, Doemeny LJ (1991) Monitoring the performance of occupational health laboratories. *Am Ind Hyg Assoc J* 52(8):336–339
- AIHA (2016) Scope of accreditation to ISO/IEC 17043:2010. AIHA Proficiency Analytical Testing Programs. www.a2la.org/scopepdf/3300-01.pdf
- Ashley K (2015) NIOSH manual of analytical methods 5th edition and harmonization of occupational exposure monitoring. *Gefahrst Reinhalt Luft* 2015(1–2):7–16
- Cox LA Jr, Van Orden DR, Lee RJ, Arlauckas SM, Kautz RA, Warzel AL, Bailey KF, Ranpuria AK (2015) How reliable are crystalline silica dust concentration measurements? *Regul Toxicol Pharmacol* 73(1):126–136
- Harper M, Sarkisian K, Andrew M (2014) Assessment of respirable crystalline silica analysis using Proficiency Analytical Testing results from 2003–2013. *J Occup Environ Hyg* 11(10):D157–D163
- ISO/IEC (2010) ISO/IEC 17043:2010(en) conformity assessment — general requirements for proficiency testing www.iso.org/obp/ui/#iso:std:iso-iec:17043:ed-1:v1:en
- Lee RJ, Van Orden DR, Cox LA, Arlauckas S, Kautz RJ (2016) Impact of muffle furnace preparation on the results of crystalline silica analysis. *Regul Toxicol Pharmacol* 80:164–172
- Maciejewska A (2006) Analysis of the competences of workplace inspecting laboratories for the determination of free crystalline silica (FCS), based on proficiency testing results. *Med Pr* 57 (2):115–122. Polish
- Shulman SA, Groff JH, Abell MT (1992) Performance of laboratories measuring silica in the Proficiency Analytical Testing program. *Am Ind Hyg Assoc J* 53(1):49–56

Part V

**Risk Management: Insights from
Prescriptive, Learning, and Collaborative
Analytics**

Chapter 12

Improving Individual, Group, and Organizational Decisions: Overcoming Learning Aversion in Evaluating and Managing Uncertain Risks



The descriptive, causal, predictive, and evaluation analytics illustrated in Chaps. 3–11 are largely about *risk assessment*. That is, they are about quantifying how large risks are now; predicting how much smaller they would become if costly interventions were undertaken (e.g., shifting pigs from closed to open production or further reducing air pollution levels); and evaluating how effective past interventions have been and how well current systems that help to monitor and control potential risks are performing. Such analytics help to inform decision-makers about current risks and the probable effectiveness and tradeoffs among objectives created by proposed risk management actions. This chapter and those that follow turn to prescriptive *risk management* issues: deciding what to do next and learning how to better achieve desired goals. This chapter reviews principles of benefit-cost analysis and practical psychological pitfalls that make it difficult for individuals, groups, and organizations to learn optimally from experience. It proposes possible ways to overcome these obstacles, drawing on insights from learning analytics and adaptive optimization from Chap. 2. Chapter 13 offers advice on how to help move organizations toward effective risk management practices by recognizing and rejecting common excuses that inhibit excellent collective risk management decision-making and by taking advantage of opportunities to learn and collaborate in sensing, interpreting, and responding to warning signs. Chapter 14 considers how regulatory and judicial institutions can work together to promote improved societal risk management and to advance the public interest by assuring that sound causal analytics, manipulative causation, and valid causal inferences, are made the basis for regulatory interventions. Chapter 15, which concludes this book, brings together and extends these prescriptive threads by considering philosophical, game-theoretic, and economic models for how to make risk management decisions with consequences that span multiple generations.

Introduction

Risk management decisions and learning should make good use of information from descriptive, causal, predictive, evaluation, and learning analytics, but it is often not easy to do so. Decision biases distort perceptions and cost-benefit evaluations of uncertain risks and the value of reducing them, leading to risk management policy decisions with predictably high retrospective regret. Well-documented psychological heuristics and biases used in decision-making encourage a phenomenon that we will call *learning aversion*: sub-optimal learning and premature decision-making in the face of high uncertainty about the costs, risks, and benefits of proposed changes. Narrow framing, overconfidence, confirmation bias, optimism bias, ambiguity aversion, and hyperbolic discounting of the immediate costs and delayed benefits of learning (Kahneman 2011) all contribute to deficient individual and group learning, avoidance of information-seeking, under-estimation of the value of further information, and hence needlessly inaccurate risk-cost-benefit estimates and sub-optimal risk management decisions. In practice, such biases can create predictable regret in selection of potential risk-reducing regulations.

The reinforcement learning and adaptive optimization method surveyed in Chap. 1 suggest principles that can help to improve decisions by more effectively balancing exploration (deliberate experimentation and uncertainty reduction) and exploitation (taking actions to maximize the sum of expected immediate reward, expected discounted future reward, and value of information). This chapter discusses how these and related ideas might be used to understand and overcome learning aversion and implement low-regret learning strategies, using regulation of air pollutants with uncertain health effects as an example.

Benefit-Cost Analysis

For most of the past century, economists have sought to apply methods of benefit-cost analysis (BCA) (Portney 2008) to help policy makers identify which proposed regulations, public projects, and policy changes best serve the public interest. BCA provides methods to evaluate quantitatively, in dollar terms, the total economic costs and benefits of proposed changes. In versions commonly used by regulators and analysts, BCA prescribes that decisions be made to *maximize the expected net present value* (NPV) of resulting time streams of net benefits (i.e., monetized benefits minus costs), with delayed and uncertain impacts being appropriately discounted to yield a net present value for each option being evaluated. Similarly, in law-and-economics analyses of negligence torts, the Learned Hand Rule prescribes a duty to take care to prevent or reduce risk if the cost of doing so is less than the expected benefit (Grossman et al. 2006). In regulatory BCA, benefits are typically measured as the greatest amounts that people who want the changes would be willing to pay (WTP) to obtain them. Costs are measured by the smallest

amounts that people who oppose the changes would be willing to accept (WTA) as full compensation for them (Portney 2008). Recommending alternatives with the greatest expected NPV helps to adjudicate the competing interests of those who favor and those who oppose a proposed change.

Example: A Simple BCA Justification for Banning Coal Burning

Recall from Chapter the 2002 study in the *Lancet* suggesting that a relatively simple public health intervention, banning burning of coal in Dublin County, Ireland, created substantial health benefits (Clancy et al. 2002). The study concluded that “Reductions in respiratory and cardiovascular death rates in Dublin suggest that control of particulate air pollution could substantially diminish daily death...Our findings suggest that control of particulate air pollution in Dublin led to an immediate reduction in cardiovascular and respiratory deaths.” In a press release, one of the authors explained that “The results could not be more clear, reducing particulate air pollution reduces the number of respiratory and cardiovascular related deaths immediately” (Harvard School of Public Health 2002). Citing these estimated benefits, policy makers extended the bans more widely, reasoning that “Research has indicated that the smoky coal ban introduced in Dublin in 1990 resulted in up to 350 fewer deaths...per year. It has clearly been effective in reducing air pollution with proven benefits for human health and our environment...” (Department of the Environment Community and Local Government 2012).

As a simple example to illustrate BCA ideas and principles, suppose that “350 fewer deaths per year” is well-defined, with each such postponed death being valued at \$1M for purposes of BCA. (Technically, what actually changes is presumably the *ages* at which deaths occur, rather than the *number* of deaths per year. In steady state, the deaths postponed from this year to next year would be exactly offset by the number of deaths postponed from last year to this year, so the number of deaths per year would actually remain unchanged (and on average equal to the number of births per year, since each birth eventually generates one death), even though everyone now lives a year longer. However, for purposes of illustration, we will assume that total benefits of \$350 M per year from increased longevity is a realistic estimate of the benefits in question.) Assume that extending the coal-burning ban to a wider area is estimated to double the effect of the original ban, creating another “350 fewer deaths per year.” Also for simplicity, suppose that the total costs of the proposed extended coal ban are estimated as \$100 M per year (e.g., from diminished coal producer and consumer surpluses, increased costs of gas and electricity as these are substituted for coal burning, unsatisfied demand for more heat at an affordable price in the winter, etc.) For purposes of illustration, only these costs and benefits will be considered. Since total estimated benefits from extending the ban greatly exceed total estimated costs, creating a total net benefit per year from extending the ban that

is estimated to be $\$350\text{ M} - \$100\text{ M} = \$250\text{ M}$ per year, the BCA recommendation would be to extend the ban. Even if the estimated cost were doubled or the estimated benefit were halved (but not both), the estimated net benefit would still be positive. Such sensitivity analysis is often used to check whether policy recommendations are robust to plausible uncertainties, as they appear to be here. (This example is continued below.)

Arguably, seeking to maximize net social benefit in this fashion promotes a society in which everyone expects to gain from public decisions on average and over time, even though not everyone will gain from every decision. Hence, BCA offers a possible approach to collective choice that appears to meet minimal standards for justice (it might be favored by everyone from an initial position behind Rawls's veil of ignorance) and economic efficiency (those who favor an adopted change gain more from it than those who oppose it lose). At first glance, BCA appears to have developed a decision-making recipe that circumvents the daunting impossibility theorems of collective choice theorists (e.g., Hylland and Zeckhauser 1979; Mueller 2003; Man and Takayama 2013; Nehring 2007; Othman and Sandholm 2009), which imply that no satisfactory way exists in general to use individual preferences to guide economically efficient social choices while protecting other desirable properties (such as voluntary participation and budget balance). For that is precisely what BCA seeks to do.

However, this chapter argues that, whatever its conceptual strengths and limitations might be for *homo economicus*, or purely rational economic man, BCA for real-world regulations or projects with risky outcomes often leads to *predictably regrettable* collective choices in practice (and does not really succeed in bypassing impossibility results in principle). More useful recommendations can be developed by seeking to *minimize expected rational regret*, rather than to maximize expected NPV, especially when probabilities for different costs and benefits are unknown or uncertain. This criterion, explained further later, is also better suited to the needs of real decision-makers with realistically imperfect information about the costs and benefits of proposed changes than is the principle of maximizing expected NPV.

Example (Cont.): A BCA Justification for Banning Coal Burning May Be Regrettable

In the Dublin study, the original researchers' conclusion that "The results could not be more clear, reducing particulate air pollution reduces the number of respiratory and cardiovascular related deaths immediately" (Harvard School of Public Health 2002)" was later questioned by methodologists, who noted that the study lacked key elements, such as a control group, needed to draw valid causal conclusions. Wittmaack 2007 pointed out that mortality rates were already declining long before the ban, and occurred in other parts of Europe and Ireland not affected by it, and concluded that "Serious epidemics and pronounced trends feign excess mortality

previously attributed to heavy black-smoke exposure.” Similarly, Pelucchi et al. 2009 noted that “However, during the same period, mortality declined in several other European countries. Thus, a causal link between the decline in mortality and the ban of coal sales cannot be established.” As of 2012, when the ban was extended to additional areas and towns, there was thus some reason to question whether the original health benefits estimates were credible, or whether they might be simply an artifact of poor statistical methodology. However, the question was primarily of interest to methodologists, and played no significant role in policy-making, which assumed that the original health benefits estimates were at least approximately correct (DECLG 2012).

Such discrete uncertainties (e.g., will proposed interventions actually cause their intended and projected consequences?) cannot be resolved by simple BCA sensitivity analyses that vary inputs over ranges around the best point estimates of their values. They require confronting the discrete possibility that the true benefits might be zero, or extremely different from the estimated levels (here, around \$350 M/year), due to flaws in the underlying assumptions of the BCA. How best to incorporate such discrete uncertainties into BCA has long been a challenge and topic of controversy among BCA scholars (Graham 1981). It is no easy task to assess and justify specific probabilities for them, and any such probability would be based on information and assumptions that others might disagree with. Thus, the question arises of how to do BCA when there are substantial uncertainties about the underlying premises, modeling assumptions, and policy-relevant conclusions of the cost and benefits models (here, for health risk reductions) being used.

If the original data are available for reanalysis, then methodological issues and challenges can be openly surfaced and discussed, and whether the original BCA conclusions and recommendations change when different methodological choices can be examined. For air pollution studies, original data are not always made available to other investigators. In the case of the Dublin study, however, the Health Effects Institute (HEI) funded the original investigators to re-do their analysis, taking into account methodological considerations such as the need to compare declines in mortality inside and outside the areas affected by the ban. The main result (HEI 2013) was that, “...In contrast to the earlier study, there appeared to be no reductions in total mortality or in mortality from other causes, including cardiovascular disease, that could be attributed to any of the bans. That is, after correcting for background trends, similar reductions were seen in ban and non-ban areas. The study by Dockery and colleagues shows that accounting for background trends in mortality can be crucial, since the earlier Dublin study appears likely to have overestimated the effects of the 1990 coal ban on mortality rates from diseases that were already declining for other reasons.” Thus, when uncertainty about benefits from a coal ban was finally reduced by further investigation in 2013, it turned out that the originally projected health benefits that had seemed to provide a strong BCA rationale for coal-burning bans were no longer supported. The decision to extend the bans might be considered regrettable, if, in hindsight, the true benefits of doing so turned out to be less than the true costs.

A striking feature of this example is that the analysis done in 2013, comparing reductions in mortality risks from before to after the ban across areas affected and not affected by the ban, could have been done just as easily in 2002 as in 2013. However, there was no felt need to do so. The investigators and the recipients of the analysis believed that the correct interpretation was at hand and was obviously correct (“could not be more clear”), justifying prompt action (the ban) intended to protect the public interest, and making further investigation both unnecessary and undesirable.

The following sections suggest that this pattern is no accident. Rather, there is a strong tendency, which we refer to as *learning aversion*, to stop BCA calculations and data collection prematurely (Russo and Schoemaker 1989). Confident recommendations for action may be based on BCA estimates in which the sign of estimated net benefits could easily be reversed by further data or analysis. Sensitivity or uncertainty analyses and additional information may be presented that bolster confidence in results (e.g., by showing that even if the best estimates of costs and benefits are changed by some factor, the recommendations do not change) while doing little to highlight fundamental remaining uncertainties about whether the key premises of the BCA calculations are correct (e.g., that banning coal burning measurably reduces all-cause and cardiovascular mortality risks). In short, rather than, or in addition to, “analysis-paralysis,” the reverse problem of making high-stakes decisions prematurely, when more information having high decision-analytic value-of-information (VOI) is readily available, is also a threat to effective use of BCA. This behavior is unsurprising in light of findings from behavioral economics on how people respond to uncertainties (especially “ambiguous” ones that cannot easily be quantified via known probabilities). Once recognized, it is easily avoided, e.g., by shifting the driving metaphor for BCA away from maximizing expected net benefits based on present information, and toward minimizing later regrets (Russo and Schoemaker 1989).

The remainder of this chapter is structured as follows. The next section discusses common aspirations and motivations for BCA and discusses its promise and limitations for improving collective choices in societies of *homo economicus*. Then we discuss key features of purely rational individual decision-making and some impossibility results from collective choice theory for purely rational agents. Turning to how real people make decisions, including many “predictably irrational” ones (Ariely 2009), we argue that a web of well-documented decision heuristics and biases calls into question the usual normative prescriptive use of elicited or inferred WTP and WTA amounts in many practical applications. Both WTP and WTA amounts are sensitive to details of framing, context, perceptions of fairness and rights, feelings about social obligations and entitlements, and other factors that depart from the simplified economic models (e.g., quasi-linear preferences with additively separable costs and benefits) envisioned in the usual foundations of BCA. Psychological phenomena such as ambiguity aversion (reluctance to bet on unknown or highly uncertain subjective probabilities) imply several forms of what we will call *learning aversion*, i.e., refusal to use available information to improve decision-making. Simple examples illustrate mechanisms of learning aversion for

organizations as well as individuals. In following the prescriptions of BCA, real people and organizations (whether individuals, companies, regulatory agencies, or legislators and policy-makers) typically spend too much to get too little, for a variety of reasons rooted in decision psychology and political theory. We not only systematically over-estimate the prospective value (net benefit) of projects with uncertain outcomes (as in the planning fallacy (Kahneman and Tversky 1979)), but we also typically fail to test and learn enough about the likely consequences of alternative courses of action before acting (Russo and Schoemaker 1989). Hence we make collective bets on social programs and regulations that are excessively risky, in the sense that their benefits do not necessarily, or with high probability, outweigh their costs. We also usually fail to study and learn enough after acting to optimally improve decision-making models and assumptions over time. In effect, our policy-making and regulatory institutions are often *learning-averse*, with a strong bias toward premature action and insufficient prospective investigation of alternatives or retrospective learning and evaluation of decisions and outcomes (Russo and Schoemaker 1989). They show a revealed preference for acting as if we already have sufficient information to identify the best course of action with confidence now, even if available information is actually inadequate to do so, and even if a careful decision analysis (based on value-of-information analysis for maximizing expected utility) would prescribe postponing a choice.

The last part of the chapter considers how to do better. For deciding which alternative action to take next (from among those being considered, e.g., to pass or not to pass a proposed new regulation), the BCA prescription “*Choose the alternative that maximizes expected NPV*” is often less good than the advice from other rules, such as: “*Choose the alternative that minimizes expected rational regret*,” or “*Do not choose yet, but continue to learn from small-scale trials before making a final choice for large-scale deployment*.” Results from machine learning and psychology suggest that seeking to minimize regret can be a highly adaptive strategy for uncertain environments in which relevant probabilities of decision outcomes are initially unknown—that is, environments where ambiguity-aversion is likely to be especially important in decision-making. The chapter concludes with comments on the prospects for using regret minimization as an alternative to expected NPV maximization as a foundation for more practical and valuable BCA.

Aspirations and Benefits of BCA

To improve the rationality and effectiveness of collective choices, such as whether to implement a costly regulation or to undertake a costly public works project, economic benefit-cost analysis (BCA) attempts to calculate and compare the total cost of each alternative being considered to the total benefit that it would produce. If an alternative’s costs clearly exceed its benefits, it can be rejected outright. Conversely, it can be considered further for possible adoption if its benefits exceed its costs, and if no other feasible alternative would create a clearly preferable distribution of costs

and benefits. Even if costs and benefits are uncertain, one can seek to implement only those alternative(s) that produce preferred probability distributions of net benefits (e.g., distributions that are not stochastically dominated by the distributions from other choices). Thus, BCA seeks to inject rationality, objectivity, and optimization into public discourses about what to do with limited resources.

BCA comparisons are admittedly complicated by the need to make trade-offs over time, under uncertainty, and across individuals and groups, especially when those who bear most of the costs of an intervention do not receive most of its benefits. Despite these difficulties, a welcome element of common sense and benign rationality seem to infuse basic BCA prescriptions, such as *Don't take actions whose costs are expected to exceed their benefits*; or *Take actions to produce the greatest achievable net benefits*. People may argue about how best to quantify costs and benefits, including how to evaluate opportunity costs, delayed or uncertain rewards, real options, and existence values. They may disagree about how best to characterize uncertainties—e.g., what information, models, and assumptions should be used in estimating the probabilities of different possible outcomes. But the key concept of submitting proposed courses of action to the relatively objective-seeming tests of quantitative BCA comparisons, rather than letting pure politics or other processes drive public decisions about expensive actions, has appealed powerfully to many scholars and some policy makers over the past half century.

It is easy to understand why. Without such guidance, collective decisions—even those taken under a free, democratic rule of law—may harm all involved, as factional interests and narrow focusing on incremental changes take precedence over more dispassionate and comprehensive calculations for identifying which subsets of changes are most likely to truly serve the public interest.

Example: Majority Rule Without BCA Can Yield Predictably Regrettable Collective Choices

Table 12.1 shows five proposed changes that a small society, consisting of individuals 1–3 (“players,” in game theory terminology) is considering adopting. The proposed changes, labeled A–E, are shown in the rows, of the table. These might

Table 12.1 A hypothetical example of changes in annual incomes (e.g., in thousands of dollars) for each of three people from each of five alternatives

Proposed change	Player 1's income change	Player 2's income change	Player 3's income change
A	-3	1	1
B	1	-3	1
C	1	1	-3
D	3	-1	-1
E	0	0	0

represent proposed regulatory acts, investment projects, initiatives, mandates, etc. The table presents resulting changes in annual incomes for player if each measure is adopted, measured in convenient units, such as thousands of dollars per year. (For simplicity, the impacts of the different measures are assumed to be independent of each other.) For example, project A, if implemented would cost player 1 three units of income (perhaps in the form of a tax on player 1's business or activities), and would produce benefits valued at one unit of income for each of players 2 and 3. Thus, its costs are narrowly concentrated but its benefits are widely distributed. Conversely, project D would impose a tax, or other loss of income, of one unit of income on each of players 2 and 3, but would produce three units of income for player 1. E is the *status quo*.

If the collective choice process used in this small society is direct majority rule, with each participant voting for or against each proposed change, A-E, then which proposed changes will be approved? Assuming that each voter seeks to maximize his own income (or minimize his own loss), measures A-C will be adopted, since a majority (two out of three) of the players prefer each of these to the *status quo*. Summing the changes in incomes for all three of the adopted measures A-C shows that each player would receive a net loss of 1 unit of income from these three adopted collective decisions. Thus, applying simple majority rule to each proposed change A-E creates a *predictably regrettable outcome*: it is clear that changes A-C will be adopted (the outcome is *predictable*) and it is clear that this will make all voters worse off than they would have been had they instead rejected the changes and maintained the status quo (the adopted changes are, in this sense, *jointly regrettable*).

The problem illustrated here is familiar: each voter is willing to have “society” (as embodied in the collective choice process) spend other people’s money to increase his own benefit. Yet, when each faction (a coalition, or subset of players, such as players 2 and 3, for change A) has the political power to adopt a measure that achieves gain for all its members at the expense of its non-members, the portfolio of alternatives that end up being adopted harms everyone, in the sense that everyone would have preferred the *status quo*. Political theorists have recognized this possibility for centuries; it loomed large in Federalist Paper Number 10, and in concerns about tyranny of the majority.

BCA seeks to remedy this ill by subjecting each alternative to a cost-benefit test. A familiar example is the potential compensation test: *Do the gainers gain more than the losers lose?* Would those who prefer adoption of a proposed alternative still prefer it if they had to fully compensate those who preferred the *status quo*? (This question makes sense under the usual assumptions of quasi-linear preferences (utility can be expressed as benefits minus costs) and if utility is assumed to be transferable and proportional to money. Although these assumptions, in turn, may be difficult to defend, they suffice to illustrate some key points about strengths and limitations of BCA even under such idealized conditions.) Alternatives A-C in Table 12.1 fail this test, but alternative D—which would not be selected by majority rule—passes. For example, if a tax of one income unit taken from each of individuals 2 and 3 allows individual 1 to gain a benefit (such as socially subsidized healthcare) evaluated as equivalent to three income units, it might be deemed an alternative worth considering

further, since individual 1 could (at least in principle) pay one unit of income to each of individuals 2 and 3 and still be better off (by one income unit) than before the change. BCA practitioners often apply such tests for potential Pareto improvements to determine whether a proposed change is worth making (Feldman 2004).

Of course, taking from some to benefit others, especially if potential compensation remains only a theoretical possibility, raises questions about rights and justice (e.g., is enforced wealth transfer a form of theft? Would individuals voluntarily choose to adopt procedures that maximize estimated net social benefits, if they made the choice from behind the veil of ignorance in Rawls's initial position?) Moreover, it is well known that potential compensation criteria can lead to inconsistencies when a proposed alternative to the *status quo* increases one good, e.g., clean air, but reduces another, e.g., per-capita income. (Those who prefer a change in the *status quo* might still do so if they had to fully compensate those who prefer it; and yet those who prefer the *status quo* might still do so if they had to fully compensate those who do not (Feldman 2004).) Thus, potential compensation tests are by no means free of conceptual and practical difficulties. Nonetheless, the idea that a proposed change should not be adopted unless its benefit (defined as the sum of willingness-to-pay (WTP) amounts from those who want it) exceeds its cost (defined as the sum of willingness-to-accept (WTA) amounts needed to fully compensate those who don't) provides a plausible and much-cited screen for eliminating undesirable proposals (Portney 2008).

Decision-Making by *Homo economicus*

BCA was developed by economists, and is most applicable to societies of purely rational individual decision-makers. *Homo economicus*, or ideally rational economic man, has several admirable characteristics not widely shared by real people (Gilboa and Schmeidler 1989; Smith and von Winterfeldt 2004); these are briefly recalled now. He does not engage in activities whose costs are clearly greater than their benefits, or make purchases or lifestyle choices that he is certain to regret later—unlike many real-world recipients of predictable credit card bills and doctor's admonishments. He does not over-value present as opposed to delayed rewards, or certain as opposed to uncertain ones, or losses compared to gains (neither of which distracts him from a dispassionate focus on final outcomes, independent of framing and reference point effects). He does not succumb to temptations that he knows are against his rational long-term self-interest, in the sense of making current choices that he knows he will later regret. He welcomes any relevant information that increases the *ex ante* expected utility of his decisions, whether or not it supports his preconceptions. He seeks and uses such information rationally and effectively whenever the cost of acquiring it is less than its benefits in increased expected utility. He learns from new information by conditioning crisp, coherent priors on it and then acts optimally—that is, to maximize subjective expected utility (SEU)—in light of the resulting posteriors and what is known about future opportunities and constraints.

Homo economicus is a dispassionate fellow, unswayed by useless regrets (no crying over spilt milk), endowment effects (grapes are not sweetened by ownership), status quo bias (he neither fears nor seeks change for its own sake), or sunk cost bias (being in for a penny does not affect his decision about whether to be in for a pound). No business or investment strikes him as being too big to fail if failure has become the rational choice). He experiences neither thrills nor anxiety from gambling once his bets have been optimally placed; he does not hold on to losing stocks to avoid the pain of selling them and acknowledging a loss; and he never seeks to win back with an unfavorable bet what he has already lost. His Prospect Theory weighting function for probabilities is a 45° line, so that he neither overestimates the probabilities of rare events (thus driving over-investment in protecting against them), nor underestimates the probabilities of more common and familiar ones (thus driving under-investments in prudent investments to protect against predictable risks, e.g., from floods or hurricanes). He does not especially favor his own prior opinions, intuitions and beliefs (no confirmation bias) or eschew uncertain probabilities or outcomes compared to known ones (no Allais or Ellsberg paradoxes, no ambiguity aversion). His choices are dynamically consistent: what he plans today for his future self to do, it actually does when the future arrives. These and other characteristics of *homo economicus* can be succinctly summarized by saying that he is a *subjective expected utility (SEU) decision-maker*, conforming to the usual (Savage-style) axioms for rational behavior (Gilboa and Schmeidler 1989; Smith and von Winterfeldt 2004).

However, perfect individual rationality does not necessarily promote effective collective choice. Numerous impossibility results in game theory and the theory of collective choice reveal the difficulty of constructing collective choice procedures (“mechanisms”) that will produce desirable results based on voluntary participation by rational people. Tradeoffs must be made among desirable characteristics such as budget balance (a mechanism should not run at a net loss), *ex post* Pareto-efficiency (a mechanism should not select an outcome that every participants likes worse than one that was rejected), voluntary participation, and nondictatorship (a mechanism should reflect the preferences of more than one of the participants) (e.g., Mueller 2003; Man and Takayama 2013; Othman and Sandholm 2009). Similar tradeoffs, although less well known, hold when collective decisions must be made by rational individuals with different beliefs about outcomes (Hylland and Zeckhauser 1979; Nehring 2007), as well as when they have different preferences for outcomes.

Example: Pareto-Inefficiency of BCA with Disagreements About Probabilities

Suppose that members of a society (or an elected subset of members representing the rest) must collectively decide whether to pay for an expensive regulation with uncertain health benefits (or other uncertain benefits). Uncertainties for individuals

will be represented by subjectively assessed probabilities, and the fact that these probabilities are not objectively determined is reflected in the fact that different people assess them differently. For concreteness, suppose that the collective choice to be made is whether to implement a costly proposed regulation to further reduce fine particulate air pollution in order to promote human health and longevity. Each individual believes that the benefits of the proposed regulation will exceed its costs if and only if (a) Air pollution at current levels causes significantly increased mortality risks; and (b) The proposed regulation would reduce those (possibly unknown) components of air pollution that, at sufficiently high exposure concentrations and durations, harm health. Each individual favors the regulation if and only if the joint probability of events (a) *and* (b) exceeds 20%. That is, the product of the probabilities of (a) and (b) must exceed 0.2 for the estimated benefits of the proposed regulation to exceed its costs (as these two events are judged to be independent).

As a mechanism to aggregate their individual beliefs, the individuals participating in the collective choice have agreed to use the arithmetic averages of their individual probabilities for relevant events, here (a) and (b). They will then multiply the aggregate probability for (a) and the aggregate probability for (b) and pass the regulation if and only if the resulting product exceeds 0.2. (Of course, many other approaches to aggregating or reconciling expert probabilities can be considered, but the point illustrated here with simple arithmetic averaging holds generally.)

Individual beliefs can be described by two clusters with quite different world views and subjective probability assessments. Half of the community (“pessimists”) fear both man-made pollution and our inability to control its consequences: they believe that air pollution probably does increase mortality risk, but that not enough is known for a regulation to reliably target and control the unknown components that harm human health. Specifically, they assign probability 0.8 to event (a) (exposure causes risk) and probability 0.2 to event (b) (regulation reduces relevant components of exposures). The other half of the community (“optimists”) is skeptical that that exposure increases risk, but believe that, if it does, then it is probably the components targeted by the regulation that do so (i.e., fine particulate matter rather than sulfates or something else). They assess a probability of only 0.2 for event (a) and a probability of 0.8 for event (b). Note that both sets of beliefs are consistent with the postulates that all individuals are perfectly rational, since the axioms of rationality do not determine how prior probabilities should be set (in this case, reflecting two different world views about the likely hazards of man-made pollution and our ability to control them).

Using arithmetic averaging to combine the subjective probability estimates of participating individuals (assumed to be half optimists and half pessimists), the average probability for event (a) is $(0.8 + 0.2)/2 = 0.5$, and the average probability for event (b) is likewise $(0.2 + 0.8)/2 = 0.5$. These group probability assessments imply that the collective joint probability of events (a) and (b) is $0.5 \times 0.5 = 0.25$. Since this is above the agreed-to decision threshold of 0.2, the regulation would be passed. On the other hand, every individual computes that the joint probability of events (a) and (b) is only $0.8 \times 0.2 = 0.16$. Since this is below the decision threshold of 0.2 required for projected benefits to exceed costs, no individual wants the

regulation passed. Thus, aggregating individual beliefs about events leads to a decision that no one agrees with—a regrettable outcome.

The important point illustrated by this example is not that one should not average probabilities, or that other mechanisms might work better. To the contrary, an impossibility theorem due to Nehring (2007) demonstrates that *no* method of aggregating individual beliefs and using them to make group decisions can avoid selecting dominated decisions (other than such trivial procedures as selecting a single individual as a “dictator” and ignoring everyone else’s beliefs). For *any* aggregation and decision rule that treats individuals symmetrically, one can construct examples in which the group’s decision is not favored by any of its members. (For example, using a geometric mean instead of an arithmetic means would resolve the specific problem in this example, but such a procedure would also select dominated choices in slightly modified versions of the example.) Thus, the general lesson, illustrated here for the specific aggregation mechanism of averaging individual probabilities to get collective ones, is that when probabilities of events are not known and agreed to, and opinions about them are sufficiently diverse, then calculations (collective decision mechanisms) that combine the probability judgments of multiple experts or participants to determine what acts should be taken in the public interest risk producing regrettable collective choices with which no one agrees.

Example: Impossibility of Pareto-Efficient Choices with Sequential Selection

A possible remedy for the Pareto-inefficient outcomes in the preceding example would be not to combine individual beliefs about component events at all, but instead to elicit from individuals their final, holistic preferences for, or evaluations of, collective actions. For example, each individual might estimate his own net benefit from each alternative action (pass or reject the proposed regulation, with a proposed tax or other measure to pay for it if it is passed), and then society might take the action with the largest sum of estimated individual net benefits. This would work well in the preceding example, where everyone favors the same collective choice (albeit for different reasons, based on mutually inconsistent beliefs). But it leaves the resulting decision process squarely in the domain of other well-known impossibility theorems that apply when individuals directly express preferences for alternatives.

As an example, suppose a society of three people (or a Congress of three representatives of a larger society) makes collective choices by voting among various proposed regulatory alternatives as the relevant bills are brought forward for consideration. Suppose that the legislative history is such that, in the following list of possible alternatives, the choice between A and B comes to a vote first (e.g., because advocates for PM2.5 reduction organize themselves first or best), and that later the winner of that vote is run off against alternative C (perhaps because O3 opponents propose their bill later, and it is assumed that the current cost-constrained

political environment will allow at most one such pollution reduction bill to be passed in the current session). Finally (maybe in the next session, with an expanded regulatory budget, or perhaps as a rider to an existing bill), alternative D is introduced, and run off against whichever of alternatives A-C has emerged as the collective choice so far. Here are the four alternatives considered:

- A: Do not require further reductions in any pollutant
- B: Require further reductions in fine particulate matter (PM2.5) emissions only
- C: Require further reductions in ozone (O₃) only
- D: Require further reductions in both PM2.5 and O₃.

Individual preferences are as follows (with “>” interpreted as “is preferred to”):

1. A > D > C > B
2. B > A > D > C
3. C > B > A > D

For example, individual 1 might believe that further reducing air pollution creates small (or no) health benefits compared to its costs, but believes that, if needless costs are to be imposed, they should be imposed on both PM2.5 and O₃ producers (with a slight preference for penalizing the latter, if a choice must be made). Individual 2 believes that PM2.5 is the main problem, and that dragging in ozone is a waste of cost and effort; individual 3 believes that ozone is the main problem.

Applying these individual preferences to determine majority votes, it is clear that B will be selected over A (since B is preferred to A by both of individuals 2 and 3). Then, B will lose to C (since 1 and 3 prefer C to B). Finally, D will be selected over C (since 1 and 2 prefer D to C). So, the predictable outcome of this sequence of simple majority votes is that alternative D will be the society’s final collective choice, i.e., require further reductions in both pollutants. But this choice is clearly Pareto-inefficient (and, in that sense, regrettable): everyone prefers option A (no further reduction in pollutants), which was eliminated in the first vote, to option D (further reductions in all pollutants), which ended up being adopted.

A central theme of collective choice theory for societies of rational individuals is that such perverse outcomes occur, in the presence of sufficiently diverse preferences, for all possible collective choice mechanisms (including those in which BCA comparisons are used to compare pairs of alternatives), provided that non-dictatorship or other desired properties hold (e.g., Mueller 2003; Man and Takayama 2013).

How Real People Evaluate and Choose Among Alternatives

Real people are quite different from *homo economicus* (Gilboa and Schmeidler 1989; Smith and von Winterfeldt 2004). Psychologists, behavioral economists, marketing scientists, and neuroscientists studying choices have demonstrated convincingly that most people (including experts in statistics and decision science)

depart systematically from all of the features of purely rational decision-making discussed above (e.g., Kahneman 2011). To a very useful first approximation, most of us can be described as making rapid, intuitive, emotion-informed judgments and evaluations of courses of action (“System 1” judgments, in the current parlance of decision psychology), followed (time and attention permitting) by slower, more reasoned adjustments (“System 2” thinking) (*ibid*).

The Affect Heuristic Effects Risky Choice and BCA Evaluations via a Network of Decision Biases

Much of System 1 thinking, in turn, can be understood in terms of the *affect heuristic*, according to which gut reaction—a quick, automatically generated feeling about whether a situation, choice, or outcome is good or bad—drives decisions. For most decisions and moral judgments, including those involving how to respond in risky situations, the alternative choices, situations, or outcomes are quickly (perhaps instinctively) categorized as “bad” (to be avoided) or “good” (to be sought). Beliefs, perceptions, and System 2 rationalizations and deliberations then tend to align behind these prompt evaluations. This approximate account, while over-simplified, successfully explains many of the departures of real preferences and choice behaviors from those prescribed by expected utility theory, and is consistent with evidence from neuroeconomics studies of how the brain processes risks, rewards, delays, and uncertainties (including unknown or “ambiguous” ones) in arriving at decisions. For example, immediate and certain rewards are “good” (positive valence). They are evaluated by different neural circuits than rewards that are even modestly delayed, or uncertain, perhaps explaining the observed “certainty effect” of relative over-weighting of rewards received with certainty. Conversely, immediate, certain losses are typically viewed as “bad” and are disproportionately avoided: many people will not buy with cash (immediate loss) what they will buy with credit cards (delayed loss). More generally, real people often exhibit time preferences that exhibit approximately hyperbolic discounting, and hence dynamic inconsistency: someone who would always prefer \$1 now to \$2 6 months from now may nonetheless also prefer \$2 in 36 months to \$1 in 30 months. The conflict between the high perceived value of immediate temptations and their lower perceived value (or even negative net benefit) when viewed from a distance in time explains many a broken resolution and resulting predictable regret.

Figure 12.1 provides a schematic sketch of some suggested relations among important decision biases. Although there are numerous details and a vast literature about relations among biases, the core relations in Fig. 12.1 can be summarized succinctly as: *WTP* ← *Affect heuristic* → *Learning aversion* → *Overspending* → *Rational regret*. These components are explained next. The arrows in Fig. 12.1 indicate a range of implication relations of various strengths and degrees of speculation, ranging from relatively weak (the bias at the tail of an arrow plausibly

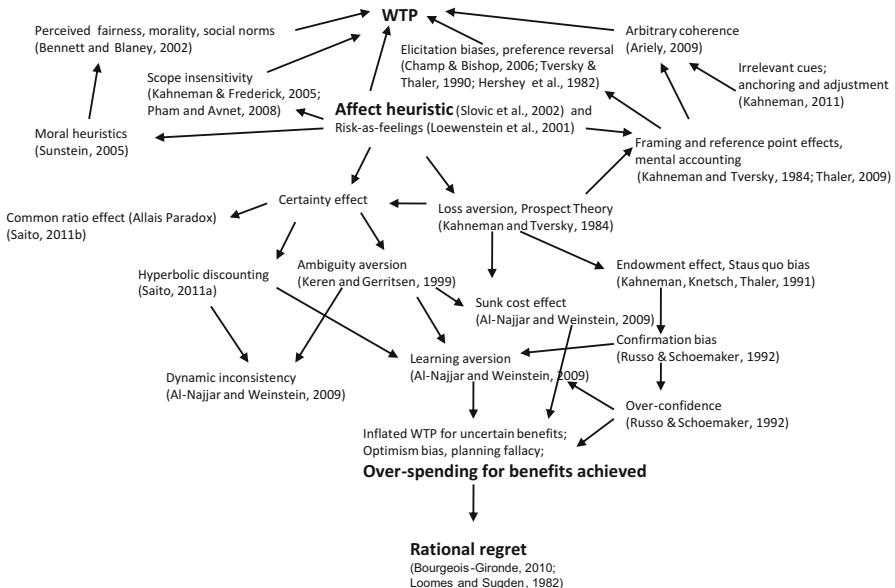


Fig. 12.1 Suggested relations among decision biases. (An arrow from A to B indicates that bias A implies, contributes to, or facilitates bias B)

contributes to, facilitates, or helps to explain the one at its head) to strong (the bias at the tail of an arrow mathematically implies the one at its head under quite general conditions). For example, it may seem plausible that the certainty effect helps to explain hyperbolic discounting if delayed consequences are interpreted by the brain as being uncertain (since something unknown might happen in the future to prevent receiving them—one might be hit by a bus later today) (Prelec and Loewenstein 1991; Saito 2011a). Establishing or refuting such a speculation empirically might take considerable effort for an experimental economist, behavioral economist, or neural economist (Dean and Ortoleva 2012; Epper and Fehr-Duda 2014). But mathematical conditions under which the certainty effect implies hyperbolic discounting (and also the common ratio effect found in the Allais Paradox) can be established fairly easily, e.g., Saito 2011a, b). The arrows in Fig. 12.1 suggest several such implications having varying degrees of support in the literature; the cited references provide details.

Some of the most striking implications in Fig. 12.1 concern the consequences of *ambiguity aversion*, i.e., reluctance to take action based on beliefs about events with unknown objective probabilities (and willingness to pay to reduce uncertainty about probabilities before acting). An ambiguity-averse decision maker would prefer to use a coin with a known probability of heads, instead of a coin with an unknown probability of heads, whether betting on heads or on tails; this is inconsistent with SEU (since revealing a preference for the coin with known probability of heads when betting on heads implies, in SEU, that one considers the other coin to have a smaller

probability of heads, and hence a larger probability of tails). Proposed normative models of decision-making with ambiguity aversion lead to preferences for acts that can be represented as *maximizing the minimum possible subjective expected utility* when the probabilities of consequences for acts, although unknown, belong to a set of multiple priors (the Gilboa-Schmeidler multiple priors representation); or to more general representations in which an additional penalty is added to each prior (Maccheronia et al. 2006). However, recent critiques of such proposed “rational ambiguity-aversion” models have pointed out the following implications (Al-Najjar and Weinstein 2009):

- *Ambiguity aversion implies that decisions do not ignore sunk costs*, as normative theories of rational decision-making would prescribe;
- *Ambiguity aversion implies dynamic inconsistency*, i.e., that people will make plans based on assumptions about how they will behave if certain contingencies occur in the future, and then not actually behave as assumed.
- *Ambiguity aversion implies learning aversion*, i.e., unwillingness to receive for free information that might help to make a better (SEU-increasing) decision.

Decision Biases Invalidate Straight-Forward Use of WTP Values

One clear implication of the network of decision biases in Fig. 12.1 is that they make WTP amounts (both elicited and revealed) untrustworthy as a normative basis for quantifying the benefits of many risk-reducing measures, such as health, safety, and environmental regulations (Casey and Delque 1995). Important, systematic departures of elicited WTP from normative principles include the following:

- *Affect heuristic*. People (and other primates) are willing to pay more for a small set of high-quality items than for a larger set that contains the same items, with some lower-quality one added as well (Kralik et al. 2012). More generally, in contrast to the prescriptions of SEU theory, expanding a choice set may change choices even if none of the added alternatives is selected, and may change satisfaction with what is chosen (Poundstone 2010).
- *Proportion dominance*. Willingness-to-pay is powerfully, and non-normatively, affected by use of proportions. For example, groups of subjects typically are willing to pay more for a safety measure described as saving “85% of 150 lives” in the event of an accident than for a measure described as saving “150 lives” (Slovic et al. 2002, 2004) (Similarly, one might expect that many people would express higher WTP for saving “80% of 100 lives” than for saving “10% of 1000 lives,” even though all would agree that saving 100 lives is preferable to saving 80.) The high percentages act as cues triggering positive-affect evaluations, but the raw numbers, e.g., “150 lives,” lack such contextual cues, and hence do not

- elicit the same positive response. This aspect of choice as driven by contextual cues is further developed in Ariely's theory of arbitrary coherence (Ariely 2009).
- *Sensitivity to wording and framing.* Describing the cost of an alternative as a "loss" rather than as a "cost" can significantly *increase* WTP (Casey and Delque 1995). The opportunity to make a small, certain payment that leads to a large return value with small probability, and else to no return, is assessed as more valuable when it is called "insurance" than when it is called a "gamble" (Hershey et al. 1982). Describing the risks of medical procedures in terms of mortality probabilities instead of equivalent survival probabilities can change preferences among them (Armstrong et al. 2002), since the gain-frame and loss-frame trigger loss-averse preferences differently, in accord with Prospect Theory.
 - *Sensitivity to irrelevant cues.* A wide variety of contextual cues that are logically irrelevant can nonetheless greatly affect WTP (Poundstone 2010). For example, being asked to write down the last two digits of one's Social Security Number significantly affects how much is willing to pay for consumer products (with higher SSNs leading to higher WTP amounts) (Ariely 2009). The "anchoring and adjustment" heuristic (Kahneman 2011) allows the mind to anchor on irrelevant cues (as well as relevant ones) that then shape real WTP amounts and purchasing behaviors (Poundstone 2010).
 - *Insensitivity to probability.* If an elicitation method or presentation of alternatives gives different salience to attributes with different effects on affect (e.g., emphasizing amount vs. probability of a potential gain or loss), then choices among the alternatives may change (the phenomenon of *elicitation bias*, e.g., Champ and Bishop 2006). Similarly, although rational (System 2) risk assessments consider the probabilities of different consequences, System 1 evaluations may be quite insensitive to the magnitudes of probabilities (e.g., 1 in a million vs. 1 in 10,000), and, conversely, overly sensitive to the change from certainty to near-certainty: "When consequences carry sharp and strong affective meaning, as is the case with a lottery jackpot or a cancer... variation in probability often carries too little weight. ...[R]esponses to uncertain situations appear to have an all or none characteristic that is sensitive to the possibility rather than the probability of strong positive or negative consequences, causing very small probabilities to carry great weight." (Slovic et al. 2002)
 - *Scope insensitivity.* Because the affect heuristic distinguishes fairly coarsely between positive and negative reactions to situations or choices, but lacks fine-grained discrimination of precise degrees of positive or negative response, WTP amounts that are largely driven by affect can be extraordinarily insensitive to the quantitative magnitudes of benefits involved. As noted by Kahneman and Frederick (2005), "In fact, several studies have documented nearly complete neglect of scope in CV [contingent valuation stated WTP] surveys. The best-known demonstration of scope neglect is an experiment by Desvouges et al. (1993), who used the scenario of migratory birds that drown in oil ponds. The number of birds said to die each year was varied across groups. The WTP responses were completely insensitive to this variable, as the mean WTP's for saving 2000, 20,000, or 200,000 birds were \$80, \$78, and \$88, respectively. ... [Similarly],

Kahneman and Knetsch (see Kahneman 1986) found that Toronto residents were willing to pay almost as much to clean up polluted lakes in a small region of Ontario as to clean up all the polluted lakes in Ontario, and McFadden and Leonard (1993) reported that residents in four western states were willing to pay only 28% more to protect 57 wilderness area than to protect a single area.”

- *Perceived fairness, social norms, and moral intensity.* How much individuals are willing to pay for benefits typically depends on what they think is fair, on what they believe others are willing to pay, and on whether they perceive that the WTP amounts for others reflect moral convictions or mere personal tastes and consumption preferences (e.g., Bennett and Blaney 2002). The maximum amount that a person is willing to pay for a cold beer on a hot day may depend on whether the beer comes from a posh hotel or a run-down grocery store, even though the product is identical in either case (Thaler 1999).

Many other anomalies (e.g., preference reversal, endowment effect, *status quo* bias, etc.) drive further gaps between elicited WTP and WTA amounts, and between both and normatively coherent preferences (see Fig. 12.1). Taken together, they rule out any straight-forward use of WTP values (elicited or inferred from choices) for valuing uncertain benefits. Indeed, once social norms are allowed as important influencers of real-world WTP values (unlike the WTPs in textbook BCA models of quasi-linear individual preferences), the question arises of whether coherent (mutually consistent) WTP values necessarily exist at all. Simple examples show that they may not.

Example: Non-existence of WTP in a Social Context

As a trivial example of the non-existence of mutually consistent individual WTP amounts when social influences are important, consider a society of two people with the following preferences for funding a proposed project:

1. Individual 1 is willing to pay up to the maximum WTP amount that anyone else pays (in this case, just individual 2), so that no one can accuse him of failing to pay his fair share. If no one else pays anything, then individual 1 is willing to pay \$100.
2. Individual 2 is willing to pay what he considers his fair share, namely, the total social benefit of the project (which he defines as the sum of WTPs from everyone else—in this case, just individual 1—divided by the number of people in society, in this case, 2).

With these preferences, there is no well-defined set of individual WTP amounts. Letting A denote the WTP for individual 1 and B the WTP for individual 2, there is no pair of WTP amounts, (A, B), satisfying the individual preference conditions that $A = B$ for $B > 0$, else $A = 100$; and $B = A/2$.

Multiple Decision Biases Contribute to Learning Aversion

The network of decision biases in Fig. 12.1 shows a prominent role for *learning aversion*, meaning reluctance to seek or use information that might change a decision for the better. The term “learning aversion” (Louis 2009) is not widely used in decision science. However, we believe it is central to understanding how to avoid premature action and to improve the practice and outcomes of BCA. For example, Table 12.2 summarizes ten well-documented “decision traps,” or barriers to effective decision-making by individuals and organizations, discussed in a popular book (Russo and Schoemaker 1982). Most of these traps involve failing to take sufficient care to collect, think about, appropriately use, and deliberately learn from relevant information that could improve decisions. Not keeping track of decision results (number 9), failing to make good use of feedback from the real world (number 8), failing to collect relevant information because of overconfidence in one’s own judgment (number 4), and trusting too much in the most readily available ideas and information (number 5) are prominent examples of failure to learn effectively from experience. Although most of the examples in the *Decision Traps* book (Russo and Schoemaker 1989) are drawn from the world of business, the same failings are

Table 12.2 Ten decision traps (from Russo and Schoemaker 1989)

1. **Plunging In**—Beginning to gather information and reach conclusions without first taking a few minutes to think about the crux of the issue you’re facing or to think through how you believe decision like this one should be made
2. **Frame Blindness**—Setting out to solve the wrong problem because you have created a mental framework for your decision, with little thought, that causes you to overlook the best options or lose sight of important objectives
3. **Lack of Frame Control**—Failing to consciously define the problem in more ways than one or being unduly influenced by the frames of others
4. **Overconfidence in Your Judgment**—Failing to collect key factual information because you are too sure of your assumptions and opinions
5. **Shortsighted Shortcuts**—Relying inappropriately on “rules of thumb” such as implicitly trusting the most readily available information or anchoring too much on convenient facts
6. **Shooting From The Hip**—Believing you can keep straight in your head all the information you’ve discovered, and therefore “winging it” rather than following a systematic procedure when making the final choice
7. **Group Failure**—Assuming that with many smart people involved, good choices will follow automatically, and therefore failing to manage the group decision-making process
8. **Fooling Yourself About Feedback**—Failing to interpret the evidence from past outcomes for what it really says, either because you are protecting your ego or because you are tricked by hindsight
9. **Not Keeping Track**—Assuming that experience will make its lessons available automatically, and therefore failing to keep systematic records to track the results of your decisions and failing to analyze these results in ways that reveal their key lessons
10. **Failure to Audit Your Decision Process**—Failing to create an organized approach to understanding your own decision-making, so you remain constantly exposed to all the above mistakes

pervasive in applied risk analysis, policy analysis, and BCA. For example, in the Dublin coal-burning ban example previously considered, the original researchers failed to collect relevant information (what happened to mortality rates outside the ban area over the same period?), while expressing great confidence in their own judgments that the correct interpretation of the data was obvious (“could not be more clear”) (Harvard School of Public Health 2002).

Figure 12.1 suggests that such learning aversion is not only a product of overconfidence (which, in turn, might reflect a predilection to consider only information and interpretations that support the views with which one is already endowed, to avoid the loss of those comfortable views and the negative affect associated with such a loss). Hyperbolic discounting and ambiguity aversion are also shown as contributors to learning aversion. Hyperbolic discounting implies that the immediate costs of learning (e.g., costs of having to collect new information that might disconfirm current beliefs, and costs of having to update current beliefs and decision rules that depend on them) may overwhelm (at present) the potential future benefits of being able to make better decisions based on the new information—even if, in retrospect, the potential (but delayed) benefits would be judged much larger than the costs of learning. Ambiguity aversion, as axiomatized by Gilboa and Schmeidler (1989) and others (Maccheronia et al. 2006) implies that a decision-maker will sometimes refuse free information that could improve decisions (Al-Najjar and Weinstein 2009). For example, in principle, an ambiguity-averse decision-maker might refuse sufficiently informative, free genetic information that is highly relevant for decisions on lifestyle, healthcare planning, and insurance purchasing (Hoy et al. 2014). Empirically, fuller disclosure of scientific uncertainties to women facing cancer treatment choices does not necessarily improve the quality of their decisions (by any measure evaluated), but does significantly reduce their subsequent (post-decision) satisfaction with the decisions that are eventually made (Politi et al. 2011).

BCA facilitates learning-averse decision-making. Its golden rule is to choose the action (from among those being evaluated) that maximizes the *expected* discounted net benefit. There is no requirement that expected values must be calculated from adequate information, or that more information collection must continue until some optimality condition is satisfied before a final BCA comparison of alternatives is made. In this respect, BCA differs from other normative frameworks, including decision analysis with explicit value-of-information calculations, and optimal statistical decision models (such as the Sequential Probability Ratio Test) with explicit optimal stopping rules and decision boundaries for determining when to stop collecting information and take action. Since learning-averse individuals (Hoy et al. 2014) and organizations (Russo and Schoemaker 1989) typically do not collect enough information (as judged in hindsight) before acting, prescriptive disciplines should explicitly encourage optimizing information collection and learning as a prelude to evaluating, comparing, and choosing among final decision alternatives (Russo and Schoemaker 1989). Helping users to overcome learning aversion is therefore a potentially valuable direction for improving the current practice of BCA.

In a collective choice context, learning aversion may be strategically rational if discovering more information about the probable consequences of alternative

choices could disrupt a coalition's agreement on what to do next (Louis 2009). But collective learning aversion may also arise because of free-rider problems or other gaps between private and public interests.

Example: Information Externalities and Learning Aversion in Clinical Trials

In clinical trials, a well known dilemma arises when each individual seeks his or her own self-interest, i.e., the treatment that is expected to be best for his or her own specific case, given presently available information. If everyone uses the same treatment, then the opportunity to learn about potentially better (but possibly worse) treatments may never be taken. Given a choice between a conventional treatment that gives a 51% survival probability with certainty and a new, experimental treatment that is equally likely to give an 80% survival probability or a 20% survival probability, and that will give the same survival probability (whichever it is) to all future patients, each patient might elect the conventional treatment (since $51\% > 0.5*0.2 + 0.5*0.8 = 50\%$). But then it is never discovered whether the new treatment is in fact better. The patient population continues to endure an individual survival probability of 51% for every case, when an 80% survival probability might well be available (with probability 50%). The same remains true even if there are many possible treatment alternatives, so that the probability that at least some of them are better than the current incumbent approaches 100%. Ethical discussions of the principle of clinical equipoise (should a physician prescribe an experimental treatment when there is uncertainty about whether it performs better than a conventional alternative, especially when opinions are divided?) recognize that refusal to experiment with new treatments (possibly due to ambiguity-aversion) in each individual case imposes a costly burden from failure to learn on the patient population as a whole, and on each member of it when he or she must choose among options whose benefits have not yet been well studied (Gelfand 2013). The principle that maximizing expected benefit in each individual case can needlessly reduce the expected benefit for the entire population is of direct relevance to BCA, as discussed further in the next example.

Example: Desirable Interventions with Uncertain Benefits Become Undesirable When They Are Scaled Up

Many people who would be willing to pay \$1 for a 50–50 chance to gain \$3 or nothing (expected net value of \$1.50 expected benefit – \$1 cost = \$0.50) might baulk at paying \$100,000 for a 50–50 chance to gain \$300,000 or nothing. Indeed, for risk-averse decision-makers, scaling up a favorable prospect with uncertain

benefits by multiplying both costs and benefits by a large enough factor can make the prospect unacceptable. (As an example, for a decision-maker with exponential utility function evaluating a prospect with normally distributed benefits having mean M and variance V , the certainty equivalent of n copies of the prospect, where all of n of them share a common uncertainty and the same outcome, has the form $CE = nM - kn^2V$, where k reflects subjective relative risk aversion. Since the first term grows linearly and the second term grows quadratically with the scaling factor n , the certainty equivalent is negative for sufficiently large n .) Now consider a local ordinance, such as a ban on coal-burning, that has uncertain health benefits and known implementation costs, such that its certainty equivalent is assessed as positive for a single county. If the same ban is now scaled up to n counties, so that the same known costs and uncertain benefits are replicated n times, then the certainty equivalent will be negative for sufficiently large n . A bet worth taking on a small scale is not worth taking when the stakes are scaled up too many times. Yet, top-down regulations that apply the same action (with uncertain benefits) to dozens, hundreds, or thousands of counties or individuals simultaneously, based on assessment that $CE > 0$ for each one, implies that essentially the same bet is being made many times, so that the total social CE will be negative if the number of counties or individuals is sufficiently large. This effect of *correlated uncertainties* in reducing the net benefits of regulations with uncertain benefits that are widely applied is omitted from BCA calculations that only consider expected values.

Learning Aversion and Other Decision Biases Inflate WTP for Uncertain Benefits

The decision biases network in Fig. 12.1 has a potentially surprising implication: Real people typically over-estimate highly uncertain benefits and under-estimate highly uncertain costs, and hence are willing to pay too much, for projects (or other proposed changes) with unknown or highly uncertain benefits and/or costs. Intuitively, one might expect exactly the reverse: that ambiguity aversion would reduce the perceived values or net benefits of such projects. But in fact, ambiguity aversion (and other drivers of learning aversion) mainly cut off information collection and analyses needed for careful evaluation, comparison, and selection of alternatives, leading to premature and needlessly risky decisions (see Table 12.2). Then, overconfidence and optimism bias take over (Fig. 12.1). From the perspective of obtaining desirable outcomes, members of most decision-making groups spend too much time and effort convincing each other that their decisions are sound, and increasing their own confidence that they have chosen well. They spend too little effort seeking and using potentially disconfirming information that could lead to a decision with more desirable outcomes (Russo and Schoemaker 1989). Moreover, in assessing the likely future outcomes of investments in risky projects, individuals and groups typically do *not* focus on the worst plausible scenario (e.g., the worst-case

probability distribution for completion times of future activities), as theoretical models of ambiguity aversion suggest (Gilboa and Schmeidler 1989). To the contrary, they tend to assign low subjective probabilities to pessimistic scenarios, and to base plans and expectations on most-favorable, or nearly most-favorable, scenarios (e.g., Newby-Clark et al. 2000).

This tendency toward overly-optimistic assessment of both uncertain benefits (too high) and uncertain costs or delays (too low) has been well documented in discussions of optimism bias (and corollaries such as the planning fallacy). For example, it has repeatedly been found that investigators consistently over-estimate the benefits (treatment effects) to be expected from new drugs undergoing randomized clinical trials (e.g., Djulbegovic et al. 2011; Gan et al. 2012); conversely, most people consistently underestimate the time and effort needed to complete complex tasks or projects, such as new drug development (Newby-Clark et al. 2000). These psychological biases are abetted by statistical methods and practices that routinely produce an excess of false positives, incorrectly concluding that interventions have desired or expected effects that, in fact, they do not have, and that cannot later be reproduced (Nuzzo 2014; Sarewitz 2012; Lehrer 2012; Ioannidis 2005). Simple Bayesian calculations suggest that more than 30% of studies with reported P values of ≤ 0.05 may in fact be reporting false positives (Goodman 2001). Indeed, tolerance for, and even encouragement of, a high risk of false-positive findings (in order to reduce risk of false negatives and to continue to investigate initially interesting hypotheses) has long been part of the culture of much of epidemiology and public health investigations supposed to be in the public interest (e.g. Rothman 1990).

The bottom of Fig. 12.1 suggests that learning aversion and several related decision biases contribute to a willingness to take costly actions with highly uncertain benefits and/or costs. Other prominent decision biases that favor such willingness to bet on a positive outcome under substantial uncertainty include the following:

- (a) *Overconfidence* in subjective judgments when objective facts or probabilities are not available (Russo and Schoemaker 1992);
- (b) *Sunk-cost effect* (propensity to throw good money after bad, or escalating commitment to an uncertain project as past investment increases, in preference to stopping and acknowledging failure and the need to move on) (Navarro and Fantino 2005); and
- (c) *Optimism bias* (e.g., underestimating the probable effort, cost, success probability, or uncertainty to complete a complex undertaking; and overestimating the probable benefits of doing so).

These biases favor premature decisions to pay to achieve uncertain benefits, even in situations where free or inexpensive additional investigation would show that the benefits are in fact almost certainly much less than the costs.

Example: Overconfident Estimation of Health Benefits from Clean Air Regulations

Overconfidence and confirmation biases can be encoded in the modeling assumptions and analytic procedures used to develop estimates of cost and benefits for BCA comparisons. For example, the U.S. EPA (2011a, b) estimated that reducing fine particulate matter (PM2.5) air pollution in the United States has created close to 2 trillion dollars per year of annual health benefits, mainly from reduced elderly mortality rates. This is vastly greater than the approximately 65 billion dollars per year that EPA estimates for compliance costs, leading them to conclude that “*The extent to which estimated benefits exceed estimated costs and an in-depth analysis of uncertainties indicate that it is extremely unlikely the costs of 1990 Clean Air Act Amendment programs would exceed their benefits under any reasonable combination of alternative assumptions or methods identified during this study*” (emphasis in original). However, the benefits calculation used a quantitative approach to uncertainty analysis based on a Weibull distribution (assessed using expert guesses) for the reduction in mortality rates per unit reduction in PM2.5. The Weibull distribution is a continuous probability distribution that is only defined over non-negative values. Thus, the quantitative uncertainty analysis implicitly assumes a 100% certainty that reducing PM2.5 does in fact cause reductions in mortality rates (the Weibull distribution puts 100% of the probability mass on positive values), in direct proportion to reductions in PM2.5 pollutant levels, even though EPA’s qualitative uncertainty analysis states (correctly) such a causal relation has not been established. An alternative uncertainty analysis that assigns a positive probability to each of several discrete uncertainties suggests that “EPA’s evaluation of health benefits is unrealistically high, by a factor that could well exceed 1000, and that it is therefore very likely that the costs of the 1990 CAAA [Clean Air Act Amendment] exceed its benefits, plausibly by more than 50-fold. The reasoning involves re-examining specific uncertainties (including model uncertainty, toxicological uncertainty, confounder uncertainty, and uncertainty about what actually affects the timing of death in people) that were acknowledged qualitatively, but whose discrete contributions to uncertainty in health benefits were not quantified, in EPA’s cost-benefit analysis” (Cox 2012). If this analysis is even approximately correct, then EPA’s highly confident conclusion results from an uncertainty analysis that disregarded key sources of uncertainty. It implicitly encodes (via the choice of a Weibull uncertainty distribution) overconfidence and confirmation biases that may have substantially inflated estimated benefits from Clean Air Act regulations by assuming, similar to the Dublin coal ban analysis, that reducing PM2.5 concentrations causes reductions in mortality rates, while downplaying (by setting its subjectively assessed probability to zero) the possibility that this fundamental assumption might be wrong.

In the political realm, the costs of regulations (or of projects or other proposed expensive changes) can also be made more palatable to decision-makers by a variety of devices, long known to marketers and politicians and increasingly familiar to

behavioral economists, that exploit the decision biases in Fig. 12.1 (Poundstone 2010). Among these are the following: postponing costs by even a little (to exploit hyperbolic discounting, since paying now provokes an adverse reaction that paying even slightly later does not); emphasizing annual costs instead of larger total costs; building in an annual rate increase (so that increases become viewed as part of the *status quo*, and hence acceptable without further scrutiny); paying from unspecified, obscure, or general funds (e.g., general revenues) rather than from specific accounts (so that trade-offs, opportunity costs and outgoing payments are less salient); adding comparisons to alternatives that no one would want to make the recommended one seem more acceptable; creating a single decision point for committing to a stream of expenses, rather than instituting multiple review and decision points (e.g., a single yes/no decision, with a limited time window of opportunity, on whether to enact a costly regulation that will last for years, rather than a contingent decision for temporary funding with frequent reviews to ask whether it has now served its purpose and should be discontinued); considering each funding decision in isolation (so that proposal can be evaluated based on its affect when viewed outside the context of competing uses to which the funds could be put); framing the cost as protecting an endowment, entitlement, or option (i.e., as paying to avoid losing a benefit, rather than as paying to gain it); and comparing expenditures to those of others (e.g., to how much EU or Japan is spending on something said to be similar). These and related techniques are widely used in marketing and advertising, as well as by business leaders and politicians seeking to “sell” programs to the public (Gardner 2009). They are highly effective in persuading consumers to spend money that, in retrospect, they might feel would have been better spent on something else (Ariely 2009; Poundstone 2010).

Doing Better: Using Predictable Rational Regret to Improve BCA

Figure 12.1 and the preceding discussion suggest that a variety of decision biases can lead to both individual and collective decision processes that place too little value on collecting relevant information, rely too heavily on uninformed or under-informed judgments (which tend to be over-optimistic and over-confident), and hence systematically over-value prospects with uncertain costs and benefits, creating excessive willingness to gamble on them. One result is *predictable disappointment*: consistent over-investment in uncertain and costly prospects that, predictably, will be seen in retrospect to have (on average) cost more and delivered less than expected. A second possible adverse outcome is *predictable regret*: investing limited resources in prospects with uncertain net benefits when, predictably, it will be clear in hindsight that the resources could have been better spent on something else. Standard BCA facilitates these tendencies by encouraging use of current expected values to make choices among alternatives, instead of emphasizing more complex, but potentially less costly

(on average), optimal sequential strategies that require waiting, monitoring, and inaction until conditions and information justify costly interventions (Stokey 2009 for economic investment decisions; Matheny et al. 2011 for hospital operations). This section considers how to do better, and what “better” means.

A long-standing tradition in decision analysis and normative theories of rational decision-making complements the principle of maximizing expected utility with various versions of minimizing expected rational regret (e.g., Loomes and Sugden 1982; Bell 1985). Formulations of rational regret typically represent it as a measure of the difference between the reward (e.g., net benefit, in a BCA context) that one’s decision *actually* achieved and the greatest reward that *could* have been achieved had one made a different (feasible) decision instead (Hart 2005; Hazan and Kale 2007). Adjusting decision rules to reduce rational regret plays a crucial role in current machine-learning algorithms, as well as in neurobiological studies of human and animal learning, adaptation, and decision-making, within the general framework of computational reinforcement learning (e.g., Li and Daw 2011; Schönberg et al. 2007). (By contrast, related concepts such as elation or disappointment (Delquié and Cillo 2006) reflect differences between expected or predicted rewards and those actually received. They do not necessarily attribute the difference to one’s own decisions, or provide an opportunity to learn how to make more effective decisions.)

Intuitively, instead of prescribing that current decisions should attempt to maximize prospective expected reward (or expected discounted net benefits), rational regret-based theories prescribe that they should be made so that, even in hindsight, one has no reason to change the decision process to increase average rewards. In effect, instead of the advice “Choose the alternative with the greatest expected value or utility,” normative theories of regret give the advice “Think about how, in retrospect, you would want to make decisions in these situations, so that no change in the decision procedure would improve the resulting distribution of outcomes. Then, make decisions that way.” In this context, a no-regret rule (Chang 2007) is one that, even in retrospect, one would not wish to modify before using again, since no feasible modification would lead to a preferred distribution of future consequences. Equivalently, if various options are available for modifying decision rules to try to improve the frequency distribution of rewards that they generate, then a no-regret rule is one that cannot be improved upon: it is a fixed point of the decision rule-improvement process (Hazan and Kale 2007). (These concepts apply to what we are calling rational regrets, i.e., to regrets about not making decisions that would have improved reward distributions.)

Example: Rational vs. Irrational Regret

Suppose that a decision maker’s reward (or “payoff,” in the game-theoretic terminology often used) is determined by her choice of an act, A or B, together with a random state of nature (e.g., the outcome of one toss of a fair die, with faces 1–6

being equally likely, revealed only after the choice has been made. Possible payoffs range between 1 and 6.1 dollars, as described by the following table.

	State:	1	2	3	4	5	6
Decision							
Act A:		1	2	3	4	5	6.1
Act B:		2	3	4	5	6	1

Expected utility theory unequivocally prescribes choosing act A (since its probability distribution of rewards stochastically dominates that of act B, as $6.1 > 6$), even though act B yield a higher payoff than A 5/6 of the time. Minimizing rational regret also prescribes choosing act A, since any decision rule that prescribes choosing B in this situation (always or sometimes) will yield a payoff frequency distribution that is inferior to (stochastically dominated by) the payoff distribution from always choosing act A. In this simple case, choosing act A and then observing that choosing B would have yielded a higher reward provides no reason for a rational decision-maker to deviate from the optimal strategy of always choosing act A. Thus, minimizing rational regret recommends A, not B.

Other concepts of regret and regret-avoidance are linked to personality psychology. These include making decisions with low potential for regret to protect damaging already low self-esteem (Josephs et al. 1992), as well as preferring to avoid learning outcomes in order to avoid possible regrets. From a biological perspective, it has been proposed that the *emotion* of regret, when used as an error signal to adaptively modify decision rules in individual decision-making, is a “rational emotion” that helps us to learn and adapt decision-making effectively to uncertain and changing environments (e.g., Bourgeois-Gironde 2010). Although these psychological and biological aspects of regret are important for some kinds of decision-making under risk, it is primarily proposed concepts of rational regret, as just discussed, that we believe are most useful for improving the practice of BCA. The rest of this section explains how.

Does the shift in perspective from maximizing prospective expected net benefits to minimizing expected retrospective regret make any practical difference in what actions are recommended? Not for *homo economicus*. For an ideally rational SEU decision-maker, the principle of maximizing SEU, while optimally taking into account future plans (contingent on future events) and the value of information, is already a no-regret rule. But for real decision-makers (whether individuals or groups) who are not able to formulate trustworthy, crisp, agreed-to probabilities for the consequences of each choice, the shift to minimizing regret has several powerful practical advantages over trying to maximize expected net benefits. Among these are the following:

- *Encourage prospective hindsight analysis.* A very practical aid for reducing overconfidence and optimism bias is for decision-makers to imagine that a contemplated project or investment ends badly, and then to figure out what could have caused this and how it might have been prevented. Such “prospective hindsight”

or “premortem” exercises have been used successfully in business to help curb under-estimation of costs and over-estimation of benefits when both are highly uncertain (Russo and Schoemaker 1989). In the domain of regulatory benefit-cost analysis, they prompt questions such as: Suppose that, 20 years from now, we rigorously assess the health benefits and economic costs actually achieved by extending Clean Air Act amendments, and find that the costs were on the order of a trillion dollars (EPA 2011a, b), but that the projected benefits of reduced mortality rates caused by cleaner air never materialized. How might this have happened? Could it have been discovered sooner? What might we do now or soon to prevent such an outcome? As discussed in Chap. 2, Bayesian network technology can help to support and automate such prospective hindsight analyses by entering undesirable outcomes as assumed findings and then seeking most-probable explanations for them. When such questions are asked on a small scale, as in the Dublin coal-ban example, they lead to simple answers, such as to use a control group (people outside the affected area) to determine whether the bans actually produced their predicted effects (HEI 2013). On a national level a similar openness to the possibility of errors in projections, and vigilance in frequently testing uncertain assumptions against data as the effects of expensive regulations become manifest, might likewise be used to anticipate and prevent the BCA failure scenarios imagined in premortem exercises. In the U.S., for example, learning from the experiences of cities, counties, or states (such as California) who are early adopters of policies or initiatives that are later proposed for national implementation provides opportunities to check assumptions against data relatively early, and to modify or optimally slow-roll (Stokey 2009) the implementation of national-level policies as needed to reduce expected regret.

- *Increase feedback and learning.* Items 8–10 in Table 12.2 describe failures to learn from real-world feedback based on the gaps between what was expected and what actually occurred, or between what was achieved and what could have been achieved by better decisions (if this is known). Formal models of how to adaptively modify decision processes or decision rules to reduce regret—for example, by selecting actions next time a situation is encountered in a Markov decision process, or in a game against nature (with an unpredictable, possibly adversarial, environment) using probabilities that reflect cumulative regret for not having used each action in such situations in the past—require explicitly collecting and analyzing such data (Robards and Sunehag 2011; Hazan and Kale 2007). Less formally, continually assessing the outcomes of decisions and how one might have done better, as required by the regret-minimization framework, means that opportunities to learn from experience will more often be exploited instead of missed.
- *Increase experimentation and adaptation.* An obvious possible limitation of regret-minimization is that one may not know what would have happened if different decisions had been made, or what probabilities of different outcomes would have been induced by different choices (Jaksch et al. 2010). This is the case when relevant probabilities are unknown or ambiguous. It can arise in practice when no states or counties have been early (or late) adopters of a

proposed national policy, and so there is no comparison group to reveal what would have happened had it not been adopted. In this case, formal models of regret reduction typically require exploring different decision rules to find out what works best. Such learning strategies (called “on-policy” learning algorithms, since they learn only from experience with the policy actually used, rather than from information about what would have happened if something different had been tried) have been extensively developed and applied successfully to regret reduction in machine learning and game theory (Chang 2007; Yu et al. 2009; Robards and Sunehag 2011). They adaptively weed out the policies that are followed by the least desirable consequences, and increase the selection probabilities for policies that are followed by preferred consequences. Many formal models of regret-minimization and no-regret learning strategies (e.g., Chang 2007; Jaksch et al. 2010 for Markov decision processes) have investigated how best to balance exploration of new decision rules and exploitation of the best ones discovered so far. Under a broad range of conditions, such adaptive selection (via increased probabilities of re-use) of the decision rules that work best empirically soon leads to discovery and adoption of optimal or near-optimal (“no-regret”) decision rules (i.e., maximizing average rewards) (Chang 2007; Robards and Sunehag 2011; Hazan and Kale 2007). Of course, translating these mathematical insights from the simplified world of formal decision models (e.g., Markov decision processes with initially unknown transition and reward probabilities and costs of experimentation) to the real world requires caution. But the basic principle that the policies that will truly maximize average net benefits per period (or discounted net benefits, in other formulations) may initially be unknown, and that they should then be discovered via well-designed and closely analyzed trials, has powerful implications for the practice of BCA and policy making. It emphasizes the desirability of conducting, and carefully learning from, pilot programs and trial evaluations (or natural experiments, where available) before rolling out large-scale implementations of regulations or other changes having highly uncertain costs or benefits. In effect, the risk of failure or substantially sub-optimal performance from programs whose assumptions and expectations about costs and benefits turn out to be incorrect can be reduced by small-scale trial-and-error learning, making it unnecessary to gamble that recommendations based on BCA using current information will turn out to coincide with those that will be preferred in hindsight, after key uncertainties are resolved.

- *Asymptotic optimization of decision rules with initially unknown probabilities for consequences.* In formal mathematical models of no-regret reinforcement learning with initially unknown environments and reward probabilities, swift convergence of the prescriptions from empirical regret-minimization algorithms to approximately optimal policies holds even if the underlying process tying decisions to outcome probabilities is unknown or slowly changing (Yu et al. 2009). This makes regret-minimization especially relevant and useful in real-world applications with unknown or uncertain probabilities for the consequences of alternative actions. It also provides a constructive approach for avoiding the fundamental limitations of collective choice mechanisms that require combining

the subjective probabilities (or expected values) of different participants in order to make a collective choice (Hylland and Zeckhauser 1979; Nehring 2007). Instead of trying to reconcile or combine discrepant probability estimates, no-regret learning encourages collecting additional information that will clarify which among competing alternative policies work best. Again, the most important lesson from the formal models is that adaptively modifying policies (i.e., decision rules) to reduce empirical estimates of regret based on multiple small trials can dramatically improve the final choice of policies and the final results produced (e.g., average rewards per period, or discounted net benefits actually achieved). From this perspective, recommending *any* policy based on analysis and comparison of its expected costs and benefits to those of feasible alternatives will often be inferior to recommending a process of trials and learning to discover what works best. No-regret learning (Chang 2007) formalizes this intuition.

In summary, adjusting decision processes to reduce empirical estimates of regret, based on actual outcomes following alternative decisions, can lead to much better average rewards or discounted net benefits than other approaches. Real-world examples abound of small-scale trial and error leading to successful adaptation in highly uncertain business, military, and policy environments (e.g., Harford 2011).

Conclusions

This chapter has argued that a foundational principle of traditional BCA, choosing among proposed alternatives to maximize the expected net present value of net benefits, is not well suited to guide public policy choices when costs or benefits are highly uncertain. In principle, even if preferences are aligned (so that familiar collective choice paradoxes and impossibility results caused by very different individual preferences do not arise)—for example, even if all participants share a common goal of reducing mortality risks—there is no way (barring such extremes as dictatorship) to aggregate sufficiently diverse probabilistic beliefs to avoid selecting outcomes that no one favors (Hylland and Zeckhauser 1979; Nehring 2007). BCA does not overcome such fundamental limitations in any formulation that requires combining probability estimates from multiple participants to arrive at a collective choice among competing alternatives—including using such probabilities to estimate which alternative has the greatest expected net benefit.

In practice, a variety of well-known decision biases conspire to make subjectively assessed expected value calculations and WTP estimates untrustworthy, with highly uncertain benefits often tending to be over-estimated, and highly uncertain costs tending to be under-estimated. Biases that contribute to unreliable expected net benefit and WTP estimates range from the affect heuristic, which we view as fundamental, to optimism, over-confidence, and confirmation biases, ambiguity aversion, and finally to what we have called learning aversion (Fig. 12.1). As a result of this network of biases, it is predictable that projects and proposals with

highly uncertain costs and benefits will tend to be over-valued, leading to potentially regrettable decisions, meaning decisions that, in retrospect, and upon rational review, one would want to have made differently. Similar results have been demonstrated for groups and for individuals (Russo and Schoemaker 2009). The net result is a proclivity to gamble on excessively risky proposals when the benefits and costs are highly uncertain.

To help overcome these difficulties, we have proposed shifting to a different foundation for BCA calculations and procedures: minimizing rational regret. Regret minimization principles been developed in both decision analysis (e.g., Loomes and Sugden 1982; Bell 1985) and extensively in more recent machine learning, game theory, and neurobiological models of reinforcement learning (Hart 2005; Chang 2007; Hazan and Kale 2007; Li and Daw 2011; Schönberg et al. 2007). Although the idealized mathematical models and analyses of these fields are not necessarily directly applicable to real-world BCA settings, they do suggest several practical principles that have proved valuable in improving real-world individual and collective decisions when potential costs and benefits are uncertain enough so that the best course of action (given clarity on goals) is not clear. In particular, we propose that BCA under such conditions of high uncertainty can be improved by greater use of prospective hindsight (or “*premortem*”) analyses to reduce decision biases; explicit data collection and careful retrospective evaluation and comparison of what was actually achieved to what was expected, and to what could have been achieved by different choices (when this can be determined); and deliberate learning and adaptation of decision rules based on the results of multiple small-scale trials in settings for which this is practical. Not all of these principles are applicable in all BCA situations, of course. Whether to build a bridge in a certain location cannot be decided by multiple small-scale trials, for example. But for many important health, safety, and environmental regulations with substantial costs and substantial uncertainty about benefits, learning from experiences on smaller scales (e.g., from the changes in mortality rates following different histories of pollution reductions in different counties) can powerfully inform and improve BCA analyses that are intended to guide larger-scale (e.g., national) policy-making. The main proposed shift in emphasis is from *guessing* what will work best (in the sense of maximizing the expected NPV of net benefits, as assessed by experts or other participants in the decision-making process), and then perhaps betting national policies on the answer, to *discovering* empirically what works best, when it is practical to do so and when the answer is initially highly uncertain.

References

- Al-Najjar NI, Weinstein J (2009) The ambiguity aversion literature: a critical assessment. *Econ Philos* 25(Special Issue 03):249–284
Ariely D (2009) Predictably irrational: the hidden forces that shape our decisions. Revised and expanded edition. HarperCollins, New York

- Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel PA (2002) Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. *Med Decis Mak* 22(1):76–83
- Bell DE (1985) Putting a premium on regret. *Manag Sci* 31(1):117–120. <https://doi.org/10.1287/mnsc.31.1.117>
- Bennett R, Blaney RJP (2002) Social consensus, moral intensity and willingness to pay to address a farm animal welfare issue. *J Econ Psychol* 23(4):501–520
- Bourgeois-Gironde S (2010) Regret and the rationality of choices. *Philos Trans R Soc Lond Ser B Biol Sci* 365(1538):249–257. <https://doi.org/10.1098/rstb.2009.0163>
- Casey JT, Delquie P (1995) Stated vs. implicit willingness to pay under risk. *Organ Behav Hum Decis Process* 61(2):123–137
- Champ PA, Bishop RC (2006) Is willingness to pay for a public good sensitive to the elicitation format? *Land Econ* 82(2):162–173
- Chang YC (2007) No regrets about no-regret. *Artif Intell* 171:434–439
- Clancy L, Goodman P, Sinclair H, Dockery DW (2002) Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 360(9341):1210–1214
- Cox LA Jr (2012) Reassessing the human health benefits from cleaner air. *Risk Anal* 32(5):816–829
- Dean M, Ortoleva P (2012) Estimating the relationship between economic preferences: a testing ground for unified theories of behavior. Working Paper, Department of Economics. Brown University. Providence, RI. http://www.econ.brown.edu/fac/Mark_Dean/papers.shtml. Last Retrieved 1 February 2014
- DECLG Department of the Environment Community and Local Government, March 9, 2012. New Smoky Coal Ban Regulations will bring Cleaner Air, Fewer Deaths and can help efficiency. <http://www.environ.ie/en/Environment/Atmosphere/AirQuality/SmokyCoalBan/News>MainBody,31034.en.htm>. Last Retrieved 1 February 2014
- Delquié P, Cillo A (2006) Disappointment without prior expectation: a unifying perspective on decision under risk. *J Risk Uncertain* 33(3):197–215. <https://doi.org/10.1007/s11166-006-0499-4>
- Djulbegovic B, Kumar A, Magazin A, Schroen AT, Soares H, Hozo I, Clarke M, Sargent D, Schell MJ (2011) Optimism bias leads to inconclusive results—an empirical study. *J Clin Epidemiol* 64 (6):583–593. <https://doi.org/10.1016/j.jclinepi.2010.09.007>
- EPA (2011a) The benefits and costs of the Clean Air Act from 1990 to 2020: summary report. U.S. EPA, Office of Air and Radiation, Washington, DC. <http://www.epa.gov/air/sect812/aug10/summaryreport.pdf>
- EPA (2011b) The benefits and costs of the Clean Air Act from 1990 to 2020. Full report. U.S. EPA, Office of Air and Radiation, Washington, DC. <http://www.epa.gov/oar/sect812/feb11/fullreport.pdf>
- Epper T, Fehr-Duda H (2014) The missing link: unifying risk taking and time discounting. www.thomasepper.com/papers/RaT2.pdf. Last Retrieved 9 February 2018
- Feldman AM (2004) Kaldor-Hicks compensation. In Newman P (ed), *The new Palgrave Dictionary of economics and the law*, vol 2, E-O. Macmillan, London, pp 417–412
- Gan HK, You B, Pond GR, Chen EX (2012) Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. *J Natl Cancer Inst* 104(8):590–598. <https://doi.org/10.1093/jnci/djs141>
- Gardner D (2009) *The science of fear: how the culture of fear manipulates your brain*. Penguin Group, New York
- Gelfand S (2013) Clinical equipoise: actual or hypothetical disagreement? *J Med Philos* 38 (6):590–604. <https://doi.org/10.1093/jmp/jht023>. Epub 2013 Jul 22
- Gilboa I, Schmeidler D (1989) Maxmin expected utility with a non-unique prior. *J Math Econ* 18:141–153
- Goodman SN (2001) Of P-values and Bayes: a modest proposal. *Epidemiology* 12(3):295–297. No abstract available

- Graham DA (1981) Cost-benefit analysis under uncertainty. *Am Econ Rev* 71(4):715–725. http://www.nber.org/papers/w0194.pdf?new_window=1. Last Retrieved 1 February 2014
- Grossman PZ, Cearley RW, Cole DH (2006) Uncertainty, insurance, and the Learned Hand formula. *Law Probab Risk* 5(1):1–18
- Harford T (2011) Adapt: why success always starts with failure. Farra, Straus and Giroux, New York
- Hart A (2005) Adaptive heuristics. *Econometrica* 73(5):1401–1430. <http://www.math.huji.ac.il/~hart/papers/heurist.pdf>
- Harvard School of Public Health (2002) Press Release: “Ban On Coal Burning in Dublin Cleans the Air and Reduces Death Rates”. www.hsphs.harvard.edu/news/press-releases/archives/2002-releases/press10172002.html
- Hazan E, Kale S (2007) Computational equivalence of fixed points and no regret algorithms, and convergence to equilibria. *Adv Neural Inf Process Syst* 20:625–632
- Health Effects Institute (HEI) (2013) Did the Irish coal bans improve air quality and health? HEI Update, Summer, 2013. <http://pubs.healtheffects.org/getfile.php?u=929>. Last Retrieved 1 Feb 2014
- Hershey JC, Kunreuther HC, Schoemaker PJH (1982) Sources of bias in assessment procedures for utility functions. *Manag Sci* 28(8):936–954
- Hoy M, Peter R, Richter A (2014) Take-up for genetic tests and ambiguity. *J Risk Uncertain* 48 (2):111–133
- Hylland A, Zeckhauser RJ (1979) The impossibility of Bayesian Group decision making with separate aggregation of beliefs and values. *Econometrica* 47(6):1321–1336
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8):e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *J Mach Learn Res* 11:1563–1600
- Josephs RA, Larrick RP, Steele CM, Nisbett RE (1992) Protecting the self from the negative consequences of risky decisions. *J Pers Soc Psychol* 62(1):26–37
- Kahneman D (2011) Thinking fast and slow. Farrar, Straus, and Giroux, New York
- Kahneman D, Frederick S (2005) A model of heuristic judgment. In: Holyoak KJ, Morrison RG (eds) *The Cambridge handbook of thinking and reasoning*. Cambridge University Press, New York, pp 267–293
- Kahneman D, Tversky A (1979) Intuitive prediction: biases and corrective procedures. *TIMS Stud Manage Sci* 12:313–327
- Kahneman D, Tversky A (1984) Choices, values and frames. *Am Psychol* 39:341–350
- Kahneman D, Knetsch JL, Thaler RH (1991) Anomalies: the endowment effect, loss aversion, and status quo bias. *J Econ Perspect* 5(1):193–206. Winter 1991
- Keren G, Gerritsen LEM (1999) On the robustness and possible accounts of ambiguity aversion. *Acta Psychol* 103(1–2):0149–0172
- Kralik JD, Xu ER, Knight EJ, Khan SA, Levine WJ (2012) When less is more: evolutionary origins of the affect heuristic. *PLoS One* 7(10):e46240. <https://doi.org/10.1371/journal.pone.0046240>
- Lehrer J. Trials and errors: why science is failing us. *Wired*. January 28, 2012. <http://www.wired.co.uk/magazine/archive/2012/02/features/trials-and-errors?page=all>
- Li J, Daw ND (2011) Signals in human striatum are appropriate for policy update rather than value prediction. *J Neurosci* 31(14):5504–5511. <https://doi.org/10.1523/JNEUROSCI.6316-10.2011>
- Loomes G, Sugden R (1982) Regret theory: an alternative theory of rational choice under uncertainty. *Econ J* 92(368):805–824
- Louis P (2009) Learning aversion and voting rules in collective decision making, mimeo, Universitat Autònoma de Barcelona
- Maccheroni F, Marinacci M, Rustichini A (2006) Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica* 74(6):1447–1498
- Man PTY, Takayama S (2013) A unifying impossibility theorem. *Economic Theory* 54(2):249–271

- Matheny ME, Normand SL, Gross TP, Marinac-Dabic D, Loyo-Berrios N, Vidi VD, Donnelly S, Resnic FS (2011) Evaluation of an automated safety surveillance system using risk adjusted sequential probability ratio testing. *BMC Med Inform Decis Mak* 11:75. <https://doi.org/10.1186/1472-6947-11-75>
- Mueller DC (2003) *Public Choice III*. Cambridge University Press, New York
- Navarro AD, Fantino E (2005) The sunk cost effect in pigeons and humans. *J Exp Anal Behav* 83 (1):1–13
- Nehring K (2007) The impossibility of a Paretian Rational: a Bayesian perspective. *Econ Lett* 96 (1):45–50
- Newby-Clark IR, Ross M, Buehler R, Koehler DJ, Griffin D (2000) People focus on optimistic scenarios and disregard pessimistic scenarios while predicting task completion times. *J Exp Psychol Appl* 6(3):171–182
- Nuzzo R (2014) Scientific method: statistical errors. P values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume. *Nature* 506:150–152. <https://doi.org/10.1038/506150a>
- Othman A, Sandholm T (2009) How pervasive is the Myerson-Satterthwaite impossibility? In: *Proceedings of the 21st international joint conference on artificial intelligence, IJCAI'09*. Morgan Kaufmann Publishers Inc. San Francisco, CA, pp 233–238
- Pelucchi C, Negri E, Gallus S, Boffetta P, Tramacere I, La Vecchia C (2009) Long-term particulate matter exposure and mortality: a review of European epidemiological studies. *BMC Public Health*. 9:453
- Pham MT, Avnet T (2008) Contingent reliance on the affect heuristic as a function of regulatory focus. *Organ Behav Hum Decis Process* 108:267–278. <http://www.columbia.edu/~tdp4/OBHDHP2009.pdf>. Last Retrieved 14 Feb 2014
- Politi MC, Clark MA, Ombao H, Dizon D, Elwyn G (2011) Communicating uncertainty can lead to less decision satisfaction: a necessary cost of involving patients in shared decision making? *Health Expect* 14(1):84–91. <https://doi.org/10.1111/j.1369-7625.2010.00626.x>. Epub 2010 Sep 23
- Portney PR (2008) Benefit-cost analysis. In: Henderson DR (ed) *The concise encyclopedia of economics*. Library of Economics and Liberty. Last Retrieved 1 February 2014. <http://www.econlib.org/library/Enc/BenefitCostAnalysis.html>
- Poundstone W (2010) *Priceless: the myth of fair value (and how to take advantage of it)*. Scribe Publications, Melbourne
- Prelec D, Loewenstein GF (1991) Decision making over time and under uncertainty: a common approach. *Manag Sci* 37(7):770–786
- Robards M, Sunehag P (2011) Near-optimal on-policy control. Paper presented at the 9th European workshop on reinforcement learning, 9–11 Sept. Athens, Greece. https://ewrl.files.wordpress.com/2011/08/ewrl2011_submission_5.pdf. (Last downloaded 5-10-18)
- Rothman KJ (1990) No adjustments are needed for multiple comparisons. *Epidemiology* 1:43–46
- Russo JE, Schoemaker PJH (1989) Decision traps: ten barriers to brilliant decision-making and how to overcome them. Doubleday, New York
- Russo JE, Schoemaker PJH (1992) Managing overconfidence. *Winter. Sloan Manag Rev* 33 (2):7–17
- Saito K (2011a) A relationship between risk and time preferences. <http://www.hss.caltech.edu/~saito/papers/Saito2011Allais.pdf>
- Saito K (2011b) Strotz meets Allais: diminishing impatience and the certainty effect: comment. *Am Econ Rev* 101(5):2271–2275
- Sarewitz D (2012) Beware the creeping cracks of bias. *Nature* 485:149
- Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J Neurosci* 27(47):12860–12867
- Slovic P, Finucane M, Peters E, MacGregor DG (2002) Rational actors or rational fools: implications of the affect heuristic for behavioral economics. *J Socioecon* 31(4):329–342

- Slovic P, Finucane M, Peters E, MacGregor D (2004) Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Anal* 24(2):311–322
- Smith JE, von Winterfeldt D (2004) Decision analysis in “Management Science”. *Manag Sci* 50 (5):561–574
- Stokey NL (2009) The economics of inaction: stochastic control models with fixed costs. Princeton University Press, Princeton
- Sunstein C (2005) Moral heuristics. *Behav Brain Sci* 28:531–573
- Thaler RH (1999) Mental accounting matters. *J Behav Decis Mak* 12:183–206
- Tversky A, Thaler RH (1990) Anomalies: preference reversals. *J Econ Perspect* 4(2):201–211
- Wittmaack K (2007) The big ban on bituminous coal sales revisited: serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black-smoke exposure. *Inhal Toxicol* 19(4):343–350
- Yu JY, Mannor S, Shimkin N (2009) Markov decision processes with arbitrary reward processes. *Math Oper Res* 34(3):737–757

Chapter 13

Improving Risk Management: From Lame Excuses to Principled Practice



Introduction

Three classic pillars of risk analysis are *risk assessment* (how big is the risk and how sure can we be?), *risk management* (what shall we do about it?), and *risk communication* (what shall we say about it, to whom, when and how?). Chapter 1 proposed two complements to these: *risk attribution* (who or what addressable conditions actually caused an accident or loss?) and *learning from experience* about risk reduction (what works, how well, and how sure can we be?) Failures in complex systems usually evoke blame, often with insufficient attention to root causes of failure, including some aspects of the situation, design decisions, or social norms and culture. Focusing on blame, however, can inhibit effective learning, instead eliciting excuses to deflect attention and perceived culpability. Productive understanding of what went wrong, and how to do better, thus requires moving past recrimination and excuses.

This chapter, which is a slight update of Paté-Cornell and Cox (2014), identifies common blame-shifting “lame excuses” for poor risk management. These generally contribute little to effective improvements and may leave real risks and preventable causes unaddressed. We propose principles from risk and decision sciences and organizational design to improve results. These start with organizational leadership. More specifically, they include: deliberate testing and learning—especially from near-misses and accident precursors; careful causal analysis of accidents; risk quantification; candid expression of uncertainties about costs and benefits of risk-reduction options; optimization of trade-offs between gathering additional information and immediate action; promotion of safety culture; and mindful allocation of people, responsibilities, and resources to reduce risks. We propose that these principles provide sound foundations for improving successful risk management.

Why Do Catastrophes Happen? Bad Luck Is Rarely the Whole Answer

Even excellent risk management decisions usually do not usually reduce risk to zero. Flawless risk management decisions typically do not eliminate all risk. However, failures of large systems and operations, such as the *Challenger* space shuttle disaster in 1986, the core breaches at the Fukushima Daiichi reactors in 2011, or the fire at the BP Mobile Drilling Unit Deepwater Horizon also in 2011, are often rooted in flawed decision-making at high levels of the organization, with disregard or poor use of available information and of the actual effects of the incentive structure (see Chap. 12; Paté-Cornell 1990; Murphy and Paté-Cornell 1996). Human, organizational, and management factors that predispose to human errors and operation failures can and should be addressed explicitly in a systems analysis to support risk management decisions and reduce accident risks. These decisions in the design, manufacturing, construction and operation of complex systems, affect their longevity and failure risk as well as their daily performance and productivity. Too often, however, catastrophic failures lead to a tight focus, after the fact, on assigning blame, and to expensive litigation over who knew (or should have known) what and when, and who should have done things differently. People design decision rules or operating practices, attract blame and tend to be replaced (Harford 2011). Although hindsight bias may add asperity to the prosecution, the defense frequently finds that questionable (or ridiculous) arguments have been advanced for why things were done as they were, often in an effort to deflect blame from those who actually made the basic decisions.

Post hoc assignment of blame is prominent in our culture and in our justice system. It provides material for daily news and discussion: political spin on who is to blame for flooding and power outages in the wake of Hurricane Sandy; decisions about whether scientists should be jailed in Italy for opining, following small tremors, that earthquake risks in L'Aquila were modest if not negligible right before a lethal one struck (International Seismic Safety Organization 2012); and debate and recriminations over why Ambassador Stevens' calls for help did not prompt additional actions to save his life in Benghazi. More generally, the question is why precursors and near-misses were systematically ignored or misread, as they were for instance, during drilling of the Macondo well (NAE 2004).

Psychologists, however, have convincingly documented that both prospective risk assessment and retrospective blame assignment are often flawed by various heuristics and biases (e.g., Kahneman 2011), as discussed in Chap. 12. In assessing risky prospects and projects, several classic fallacies have been identified in the literature. People systematically tend to over-estimate benefits and under-estimate costs (the planning fallacy).¹ Optimistic bias leads some to accept risky prospects that they might reject if their expectations and perceptions were more accurate

¹This was the case for example, of the US space shuttle, when the risk of mission failure was originally estimated in the order of 1/5000.

(illusion of control and overconfidence). In retrospect, they may believe that whatever happened was inevitable (hindsight bias). They may prefer to continue with systems and assumptions in which large investments have already been made, rather than acknowledging, in light of new experience, that they are flawed and should be updated (sunk-cost bias). They may also blame bad luck and other people for undesired outcomes while attributing success to their own skill and efforts (self-serving bias), and altogether distort their understanding of what happened and why. Yet, identifying mistakes and their fundamental causes after a failure or a near-miss is key to learning effectively about what went wrong and how to do better in the future (Paté-Cornell 2009a, b). Therefore, to learn from costly experience how to improve risk management, it is essential to do realistic post-mortems, and not to let the opportunities for learning dissipate in a cloud of evasion and misdirection. Accordingly, this paper focuses on a set of well-worn “lame excuses” often advanced to justify the decisions and behaviors that preceded catastrophic failures of complex systems. It then proposes some principles to improve risk assessment and management by cutting through these excuses to identify needed changes in design, operations, and culture.

“It’s Not Our Fault”: Some Common Excuses for Bad Risk Management

Many arguments advanced to deflect the blame for conspicuous failures are based on claims of unpredictability: “it was a black swan”, or extremely low probability and not reasonably foreseeable; “it was a perfect storm,” or extremely rare conjunction of conditions; and so forth (Paté-Cornell 2012a, b). More generally, they are attempts to put the blame elsewhere, claiming that other people, nature, or even supernatural influences (bad luck or an Act of God) were responsible. Excuses in this category are seldom justified. The failure might well be rooted, for example, in flawed procedures as might be expected, in case of poor monitoring, from the theory of principal-agent problems (Garber and Paté-Cornell 2012).

- “It was not our job: we were not paid for it”, or “it was not our responsibility to report the deterioration of the structure or design errors that threaten its integrity.”

This kind of reasoning can be attributed to poor safety culture and the detachment of the individuals from the proper functioning of an organization or a system. Misplaced faith in the *status quo* encourages us to accept how things are and how work gets done until something breaks badly enough to force us to recognize that we should have made changes sooner (Hammond et al. 1998).

- “It was an act of God”, implying that natural forces were involved that are uncontrollable and therefore the blame cannot be put on any human being.

Of course, this is a fallacy if human choices create the exposure to the risk and determine a system's capacity to survive naturally occurring stresses. Some of the initial claims regarding the accident at the Fukushima Daiichi nuclear power plant belong in that category. Earthquakes of magnitude greater than 8 (and the large tidal waves that came with them) had occurred several times in recorded history (Epstein 2011), but the siting of the reactor and the initial choice of a 5.7 m tsunami design criterion were human decisions.

- “It was a black swan” meaning that it was both unprecedeted and unimaginable before the fact.

This one has become a favorite excuse after Taleb’s 2007 book comparing unimaginable (or, at least, unanticipated) events to the discovery of black swans in the seventeenth century by Dutch sailors who, until then, had seen only white ones (Taleb 2007). Intended by the author to explain financial crises that only a few had seen coming, it has become a much-used descriptor for events such as the attack on the US on 9/11/2001. Yet, an attack on the world trade center had occurred in 1993 and FBI agents had detected suspicious flying training in the preceding months.² With preparation, vigilance, and effort, much more is reasonably foreseeable than might be expected (Russo and Schoemaker 1990).

Deliberate exercises in applying “prospective hindsight” (i.e., assuming that a failure will occur in the future, and envisioning scenarios of how it could happen) and probabilistic analysis using systems analysis, event trees, fault trees and simulation can be used to overcome common psychological biases that profoundly limit our foresight (Russo and Schoemaker 1990). These include anchoring and availability biases (Kahneman 2011; Kahneman and Tversky 1974), confirmation bias, group-think (Janis 1984), *status quo* bias, or endowment effects (Russo and Schoemaker 1990; Hammond et al. 1998). In the case of 9/11, for example, a similar terrorist attack had been mounted against an Air France flight in 1994 but thwarted by French security forces before it reached Paris. Yet, the experience had faded in the *status quo* of 2001.

- “It was a perfect storm”, i.e., it required such a rare conjunction of unlikely events that it seemed that it could be safely ignored.

This phrase became popular after the publication of a book and the release of a movie describing a convergence of several storms in the Northern Atlantic in 1991, in which a ship sank and the crew perished. Such conjunctions of unusual conditions, although individually rare, are collectively quite common (Paté-Cornell 2012a, b); but their probabilities are often underestimated because dependencies are unrecognized or misunderstood. In technical systems, these conjunctions sometimes come from common causes of failure such as external loads (e.g., extreme winds) or human errors that affect the whole system (e.g., flawed maintenance of all

²The 9/11 commission report (2004) points to a “failure of imagination” to anticipate these attacks, given past experience and new signals.

engines of an aircraft). They happen regularly in the economic sector and in supply chains, for instance, if difficult market situations and lean inventories are compounded by a natural catastrophe (independent events in this case). They are even more likely to occur in the financial industry or other tightly coupled organizations where the failure of one institution may have devastating effects on related ones. This may occur for instance, because the failure factors are economically and statistically dependent (risk “contagion”), or because psychological reactions to one event are likely to cause further failures (bank runs).

- “It never did that before” (e.g., “the system had not blown up yet in previous near-misses, so, we thought that we were fine”).

Ignoring near-misses and considering them successes is a classic reaction. Indeed, responding effectively to prevent near-misses from developing into full-blown catastrophes may reflect competence in managing hazardous situations and be a justifiable source of pride and confidence.³ Yet, interpreting near-misses as justification for complacency can make one miss potentially valuable lessons. This was the case, for example, of tire blowouts on the SST Concorde, which had happened 57 times before a piece of rubber from one of them punctured the fuel tank and caused the death of everyone on board in July 2000 (BEA 2002).⁴

- “Those in charge did not know, and could not reasonably have known, what was about to happen”.

Excusable ignorance is a common plea in organizations that fail to pass messages, especially bad news, from the front lines to decision makers higher in the hierarchy. Indeed, “plausible deniability” is sometimes sought as a way to deflect responsibilities from top decision makers. Clearly, many signals are gathered every day and organizations need some filters to function effectively (Lakats and Paté-Cornell 2004). Incentives sometimes are such that an agent can rationally take shortcuts to meet a resource constraint rather than bringing the problem to the attention of the principal (Garber and Paté-Cornell 2012). Ineffective elicitation and use of the information known to the members of an organization is both common and costly when the objective should be to align the goals of the employees and those of the organization. Therefore, changing incentives and procedures of deliberations and decisions can do much to elicit, exploit, and reward information that might otherwise remain hidden (Russo and Schoemaker 1990).

- “We did not know that things had changed so much”

³This was the case of the US Airways flight 1549 out of JFK, which, in January 2009, safely landed on the Hudson when it had been crippled minutes earlier by a bird strike and could not reach a close airport.

⁴This was also the case of the Deepwater Horizon platform, which was destroyed by explosions and fire in 2011 after close calls that should have alerted the operators and altered their course of action (NAE 2012). Yet, neither the operators nor the regulators saw reason to intervene earlier since no accident had happened.

Status quo bias lulls into assuming that things will remain what they are: the environment will not change and our system will remain what it is. This is seldom true. A general failure to monitor the situation and its evolution (including for instance markets, competitors and employees' performance) and to disregard or misinterpret signals that a new one is looming is a natural tendency (Russo and Schoemaker 1990). This change in business environment is at the core of enterprise risk management and critical in these times of globalization and quick emergence of new technologies.

- “It was permitted and convenient, so we did it.”

This was the case for instance, of the design of the Piper Alpha platform, where, against common sense, the control rooms had been located above the production modules (Paté-Cornell 1993) for the convenience of operators who could easily go from one to another. Yet, an explosion in the production modules could (and eventually did) destroy possibilities of controlling the situation.

- “We took every reasonable precaution and followed standard operating procedure.”

This would be a convincing excuse if the precautions and standard operating procedures were effectively applied to the situations for which they were intended. Yet, as potential catastrophes begin to unfold, the system, the environment, or the situation may change in ways that make the standard procedures and precautions inadequate. Blind rule-following may be disastrous if the assumptions behind the rules no longer hold. Deterioration, for example, affects automobile safety as well as that of airplanes, especially when maintenance on schedule fails to address some obvious parts such as the fuselage of an aircraft⁵ and when there is not sufficient latitude for maintenance on demand. A quick shift from standard to crisis operation mode and creative improvisation to respond to the new and unforeseen situation may then be essential to avert a disaster (Harford 2011).

- “Everybody does it”

The everybody-does-it-defense, commonly used by teenagers, implicitly assumes that it is no one's obligation to examine current *status quo* practices and their implications, especially in a changing environment (Hammond et al. 1998) and that imitating others justifies one's actions. Heedless imitation and herd-following behavior can, of course, multiply the consequences of failure if tight technological couplings make imitation easy and reflection more difficult. This is the case, for instance, of computer reactions to financial market situations, where automatic trading platforms and rules amplify the effects of initially minor price fluctuations. Similarly, destructive memes of harmful behaviors (whether teenage

⁵This was the case for example of the Aloha Airlines in 1988 in Hawaii where part of the airplane fuselage broke apart in an explosive decompression and peeled off exposing the inside of the plane (NTSB 1989).

hyperventilation, recreational drug use or copycat crimes) can spread rapidly through social media. Imitative learning is a powerful force of social coherence, but thoughtless imitation, as well as following orders without questioning their ethics and consequences can destabilize systems, spread destructive habits, and amplify risks.

- “All signals indicated that we were doing fine.”

Good test results may be falsely reassuring if the tests were performed in the wrong environment, or the sample size was too small. Over-confidence in biomedical circles has become so prevalent that commentators are starting to worry that science is failing us (Lehrer 2012). For lack of operating experience, engineering and physical models as well as expert opinions may be needed to complete the picture. Besides, “we used only the best/most credible/reliable results” may reflect a self-serving selection bias.⁶

- “But everyone agreed that it seemed a good idea at the time!”

A common reaction to failure is that everyone agreed beforehand to the course of action. “Our best experts approved, our reviewers and stakeholders (Congress, clients, funders, public, etc.) loved our analysis, our models were detailed and coherent, the results were perfectly clear it all seemed to make sense”. In reality, such consensus may reflect “groupthink” and mutual influence among the players (Janis 1984). That clients receive analytical results that fit their interest may be the result of the incentives or the information that they have given to the analysts. These results are thus directly affected by motivated reasoning, confirmation and interpretation bias, and premature closure.

In practice, full validation of a risk analysis model in the traditional statistical sense may be impossible for new systems or systems that have changed because the statistics at the system level are not yet available. Yet, these models can be justified based on what is known of the functions and dependencies of the various components, and of the external factors (loads, environment) that affect their robustness.

- “It was an operator error”

Blaming the personnel in charge at the lower level is sometimes convenient to an organization. For instance, more than 60% of recent airline accidents have been blamed on pilot errors (FAA 2013). Yet, in some cases, a system design may be the accident root cause and must be corrected to allow for some pilot mistakes before they cause a crash. In other cases, the pilots may not have been sufficiently trained, for instance, to understand the signals that they get from electronic monitors or to operate in absence of these signals. Similarly, in the medical field, some accidents are directly caused by residents who do not have sufficient experience; but the true responsibility may lie with the supervisors if they are 15 min away from the

⁶The same is true of data selection, for instance, as mentioned earlier, the choice of the Fukushima reactor designers to ignore all tsunami data older than 1000 years.

operating room when they should be accessible in 2 min. More generally, managers sometimes blame the operators for their own failures when the leadership did not provide the training, the supervision, the information or the incentives required to face critical situations.

Foundations for Better Risk Management

Valuable lessons about risk reduction can be derived from these accidents and from the excuses that are invoked after the fact. This section proposes some constructive foundations for improved risk management practices. They are selected from the management science and risk management literatures and reinforced by the cases described earlier. As witnessed by the extensive literatures on improving organizational decision-making (Russo and Schoemaker 1990) and building high-reliability organizations (Weick and Sutcliffe 2007), we are far from the first to suggest principles and practices for overcoming the foregoing limitations. But we believe that the following recommendations, which emphasize causal understanding, quantitative methods, and deliberate design for learning, can add value to earlier improvement efforts.

Understand the Causes of the Hazard, Then Its Potential Effects

“Acts of God” such as earthquakes, tsunamis or hurricanes often have a history, and their mechanisms and recurrence times can be at least partly understood. There is no more valuable tool for reducing risk than correctly analyzing and understanding causes (Paté-Cornell 1993; Cox 2013). This requires identifying the factors affecting the performance of people and systems, and their technical characteristics, as well as the environment in which they operate. It is essential, in particular, to understand who can control what, and how incentive and information structures affect agents’ decisions, response times, and error rates (e.g., Murphy and Paté-Cornell 1996). In the oil industry, for instance, rewarding production alone will likely discourage interrupting operations when immediate maintenance is needed. A general nonchalant attitude towards safety can be corrected by training, changing incentives, and making top managers aware of the true costs of risk and of opportunities to manage them more effectively. Because risks are often invisible until they have struck, it is an easy and common practice to dismiss them or to leave them to actuaries and insurers to price and manage their financial side. The human costs, however, cannot be truly redressed after the fact.

Risk analysis can make the cumulative impacts of risks on a company and its employees, customers, and partners more vivid. It can quantify, with emphasis on

uncertainties, avoidable possible losses as well as potential changes in insurance premiums and cost of capital, and can highlight cost-effective opportunities for enterprise risk management. Other risk factors are simply facts that can only be accounted for. The realities of shrinking global markets may not be changeable, but some level of diversification, innovation, and decoupling meant to protect a system from cascading failures may go a long way towards reducing risks.

More complex cases are those in which the risk is caused in large part by the activities or the threats of intelligent adversaries such as drug gangs or insurgents. The key to analyzing the risk that they present is to understand who they are, their intent, their capabilities and the types of response that one is willing to implement given the possibilities of deterrence but also escalating conflicts (Paté-Cornell 2012a, b). The issue here is thus to address not only the symptoms (e.g., immediate threats of attacks) but also the basic problems, although sometimes, as in saving a patient, treating the symptoms may have to come first.⁷

Characterize Risk Magnitudes and Uncertainties

Once a hazard—a possible source of risk—is identified, the next step is to try to figure out what one is dealing with and how large the risk might be (Kaplan and Garrick 1981). Are probabilities and magnitudes of potential losses large enough, compared to the costs of reducing them, to warrant urgent attention, or are they small enough that waiting to address them implies little possible loss? This is where science reporters often fail as risk communicators, by publishing articles exclaiming that exposures have been “linked” to adverse effects, but without noting the absolute sizes of the risks involved or, often, even failing to check whether the claimed “links” are causal (Gardner 2009). If the risk is uncertain, can this uncertainty be clarified for a cost that is less than the value of information obtained by additional investigation, and the benefits of improved decisions that it would make possible?⁸ If so, acting quickly out of concern about uncertain risks may be less prudent than first collecting better information.

This quantification of the risk and of the associated uncertainties may be a difficult task depending on the nature and the relevance of the available evidence. Quantitative risk assessment (QRA) or Probabilistic Risk Analysis (PRA), developed originally for engineered systems, involve all available information that can help to answer practical risk management and uncertainty reduction questions. These methods are based both on systems analysis and on probability, including essential

⁷That was the dilemma in managing the financial crisis of 2008, when governments were facing several options: some favored the injection of stimulus capital first then the regulation of banking reserves and others the reverse.

⁸Of course, additional information may also increase the uncertainties about a risk and justify increased safety measures.

functions, feedback loops and dependencies caused by external events and common causes of failure (Paté-Cornell 2009a, b). For a structural system, these external events can be earthquakes or flooding that affect simultaneously several subsystems and components. In these cases, the risk analysis is based on an assessment of the probability distributions of loads and capacities, and computation of the chances that the former exceeds the latter. When needed, the results should include, and if needed display separately, the effects of aleatory as well as epistemic uncertainties⁹ to accurately characterize the limitations of the analytical results.

Realistic PRAs, including those for failures of complex technological systems, must also include human and organizational factors. This analysis can be achieved, starting from a functional and probabilistic analysis of system failures, then considering the potential decisions and actions of the actors directly involved (errors as well as competent moves), and linking these to the environment created by the management (Murphy and Paté-Cornell 1996). This requires examining in details the procedures, the structure and the culture of the organization, including the information passed along to operators, the resource constraints and the incentive system.

These risk analyses, imperfect as they are, can be invaluable tools in identifying risks that were not considered or were underestimated before,¹⁰ and in setting priorities among safety measures.

Identify Possible Risk Reduction Measures and Candidly Assess Their Costs, Benefits, and Uncertainties

For risks that are worth acting on now, the next step is to identify the risk mitigation alternatives and challenges in an implementing them, and to assess how much difference they would make. This is an essential step in rational (“System 2”) thinking, which is often entirely missing from emotional and intuitive (e.g., outrage-driven or amygdala-mediated) “System 1” responses to risk (Sanfey and Chang 2008). Our emotions, often based on recent events that have been widely advertised, may tell us that a situation is unacceptable and urge us to adopt a change to address the problem (“Ban that product!”). Indeed, the “precautionary principle”

⁹Aleatory uncertainties are caused by randomness. They remain even when the probability of an event is known with certainty. Epistemic uncertainty refers to imperfect basic knowledge: it is the uncertainty about the probability of an event, for example when several hypotheses are possible, when experts disagree, etc. Separating the two types of uncertainty in the display of the results as a family of risk curves is especially useful either when the analysis applies to several systems and/or over several time periods, or when the decision maker is “ambiguity averse” and epistemic uncertainties affect his/her preference function (Paté-Cornell and Davis 1994; Paté-Cornell and Fischbeck 1995; Paté-Cornell 1996).

¹⁰This was the case of the importance of the auxiliary feed-water systems in the safety of nuclear power plants as emphasized in the first PRA’s performed for these systems.

to implement such bans systematically when there remain uncertainties has been adopted by some governments (European Commission 2000). Yet, reasonable (if imperfect) calculations of how much difference alternative interventions would actually make are needed to guide risk management actions to achieve desired results.

The challenge, again, is to ensure that these assessments are as objective as humanly possible. Algorithmic techniques such as those in Chap. 2 may help. Separating facts and values may sometimes require that an analyst waste no time working for someone who will disregard fact-based results, or who insists on constraining or influencing them based on values and preconceptions, for instance by forcing some inputs. For example, if the U.S. EPA required its experts to express their uncertainty about “lives saved per microgram per cubic meter” of reduction in fine particulate matter by using Weibull distributions, which are constrained to show a 100% probability of positive life savings (no matter what the data say), then analysts might insist on being given the flexibility to use other distributions that could also assign positive probability to zero (or negative) values if that is what the data indicate (Cox 2012a). An illustration of the “risk of no risk analysis” is, again, the choice of a surprisingly low tsunami design criterion at the Fukushima Daiichi nuclear reactor, despite a recorded history of such events over more than a 1000 years as mentioned earlier. Insisting that risk management be guided by risk analyses is particularly critical for new nuclear reactors, whose design criteria must meet the characteristics of each site and the local hazards of external loads at a time where 68 new nuclear power plants are under construction across the world.

Assess the Urgency of Immediate Action and the Value of Information

Is collecting (or waiting for) additional information before acting more costly than it is worth? The value of gathering new information depends on the possibility that it will permit better (e.g., higher expected-utility) decision making. It therefore depends on the uncertainties faced by the decision maker, as well as his or her risk attitude. When deciding the urgency of action and evaluating whether to wait for additional information, a risk manager should consider:

1. Is the system stable? If not, how quickly is it deteriorating?
2. Are the benefits of gathering or waiting for additional information, which might improve the decision, expected to outweigh the costs?
3. What does one know (and can expect) of new technologies that may allow elimination of the risk altogether, for instance by replacing a hazardous technology at an acceptable cost?

An example of the first consideration—deterioration—is the speed at which one might expect, for instance, deterioration of the climate with and without proposed

interventions, with an assessment of its likely impacts (both beneficial and harmful) on human population in different parts of the globe. Examples of the second consideration—value of information and optimal stopping—include the choice of whether to perform additional medical tests before an operation, whether to engage in more invasive medical tests on a routine basis, or whether to delay a repair in a car or a chemical factory to ensure that the potential risk reduction benefits justify the costs of an immediate fix. An example of the third type—risk reduction by the substitution of a new technology—might be the decision to live with the consequence of coal burning to generate electricity, after closing nuclear plants and before solar energy becomes truly economical, understanding the pace and the costs of such new development and the actual potential for future risk reduction.

Anticipate, Monitor, and Prepare for Rare and Not-So-Rare Events

Not-so-rare events can generally be analyzed relatively easily because there is a base of experience, either with the system itself (e.g., classic earth dams) or with its components and subsystems (e.g., an aircraft that has been in service for decades, so that there is substantial operating experience with its subsystems and components). Rare events that result from the conjunction of known components (“perfect storms”) with or without dependencies may be a bit more difficult to analyze if either the probabilities or the dependencies among them are difficult to establish.

Rare or unknown events for which there is little or no information as to whether or not they can actually occur are especially difficult to manage sensibly. Starting with the most difficult case, genuine “black swans” that one knows nothing about and cannot reasonably anticipate, the best strategy may be to monitor for signals of unusual occurrences (e.g., of new diseases) and to put in place a “resilient” structure of organizational connections, financial reserves and access to human intelligence and knowledge that allows for quick, creative local responses (Paté-Cornell 2012a, b; Cox 2012b). For instance, new types of flu occur on average every 2 years. A system managed by the World Health Organization (WHO) permits monitoring and sharing of information across countries and identification of virus types. Although imperfect, that system allows relatively quick response to new strains of flu viruses such as H1N1. But the slow response to the spread of AIDS illustrates the difficulty of identifying and responding to a new type of pathogen.¹¹ Managing the more straightforward case of “perfect storms” is easier in that it involves “anticipating the unexpected” but imaginable (Augustine 1996), and observing conjunctions of dangerous events such as the convergence of storms, loads on a system, or economic problems.

¹¹The HIV retrovirus had been present in the human population for decades before it was clearly identified and researched (Gallo 2006).

Deliberately Test and Learn

As detailed in Chap. 12, an avoidable pitfall in organizational risk management is to fail to deliberately acknowledge and test key assumptions, to learn from experience and to capture data and lessons for future reference as opportunities arise (Russo and Schoemaker 1990). The world is full of “natural experiments”—unplanned but potentially highly informative shocks to systems or changes in conditions over time, as in Chap. 10—which can be used to test and refine critical assumptions underlying risk assessment and risk management... if we remember to do so. For example, if air pollution in Beijing during a winter inversion soars to dozens of times higher concentrations than are permitted in the U.S., but mortality rates do not increase correspondingly, the common risk assessment assumption that life expectancies decrease in direct proportion to pollution concentrations (Pope et al. 2009; Correia et al. 2013) should be revisited in light of the new data. Nor, outside the domain of human health, is it always necessary to wait for natural experiments. Intelligence and security professionals know that deliberately testing their systems (e.g., by “red teaming,” which grew more popular after 9/11) and trying to bypass or disable safeguards is a key to active identification and elimination or mitigation of exploitable vulnerabilities.

Learn from Near-Misses and Identify Accident Precursors

Many accidents have been preceded by close-calls, for instance when only one event did not occur in a known accident sequence. That these have not turned into a disaster has sometimes been viewed as evidence that the system needs no correction. Pro-active risk management of course, is the best way to avoid disasters. Yet, industries and regulators seem to believe at times that they should not intervene because the system has worked and no disaster has occurred—even if only by chance. The experts who claimed after a small tremor that all was safe in L’Aquila (Italy) where a large earthquake then occurred shortly after were relying on a recent occurrence of a false alert and failed to communicate to the public the fact that small shocks can also be precursors of large ones. In the case of the 2011 accident at the Macondo well, the regulators as well as the three companies involved did not intervene when they knew that some worrisome near-misses had occurred (presuming that they were doing well enough), and decided to ignore precursors and test results (NAE 2012) presumably for a variety of immediate benefits.

Establish and Maintain a Culture of Safety

It is possible to deliberately create and maintain a safety culture that reduces accident risks and losses. This requires acting beyond the classic ritual statements of “safety first”. A safety culture starts at the head of an organization, with a true commitment

to recognize and properly manage unavoidable tradeoffs, and by training those who are closest to operations to make appropriate decisions when needed. Therefore the deliberate design and development of highly reliable organizations (HROs) typically emphasize adopting a vigilant, risk-aware mind set and instilling the following five principles throughout the organization: preoccupation with failure at all levels and by all hands; reluctance to jump to conclusions or simplify interpretations of data and anomalies; sensitivity to operations at all levels; commitment to resilience; and deference to relevant expertise, rather than to authority (Weick and Sutcliffe 2007).

A key part of a safety culture thus involves the incentives provided by the management. The structure and the procedures of organizations such as an oil company reflect an attitude at the top of the corporation that permeates all levels of the organization. Concretely, the incentives, constraints, and directions explicitly communicated to employees shape their decisions, especially when they have little time to react or little information to evaluate their decisions. This was one of many problems at the Fukushima Daiichi nuclear power plant, where operators had to wait for hours before deciding on their own to flood a crippled reactor. It was also true on the Deepwater Horizon platform, where ignoring negative pressure tests results contributed to the already high risks of an accident (NAE 2012). Economic incentives that encourage motivated reasoning may thus distort risk-taking and risk-management decisions. As pointed out earlier, organizations that reward exclusively a production level and *de facto* penalize those who slow down production, put at risk not only their employees but also possibly, the general public both from a safety and a financial point of view.

Put the Right People in the Right Place with the Right Knowledge, Incentives and Resources

Training and learning are two of the most important requirements for effective risk management. Risk analysis can clarify the effectiveness and performance of risk management decisions and their importance in affecting outcomes. The results, in turn, allow assessing where additional resources and training, as well as changes in incentives and responsibilities, are most likely to pay off in reduced risks and improved performance. Having examined what drives the operators of a complex system (e.g., the conductors of high-speed trains), one can also review management procedures, structure and culture for fitness to meet the needs of both regular operations and responses to crisis. In normal operations, disciplined rule-following can protect us against the temptations, heuristics and biases that undermine so much human decision-making. These range from succumbing to short-run impulses that we may come to regret such as hyperbolic discounting (Chap. 12; Lehrer 2009; Gardner 2009), to letting fears, doubts, and desire control decisive actions which, upon reflection, no one favors upon reflection. (Recall that hyperbolic discounting describes “present-biased” preferences in which the same delay in reward is valued

at different rates in the present than in the future, e.g., if \$10 now is preferred to \$20 in 1 year, yet \$20 in 6 years is preferred to \$10 in 5 years.) On the other hand, when reality turns to crisis or emergency situations, narrow rule-following can lead to blinkered vision and to abdication of the responsibility, creativity, and active cooperation needed for adaptive responses to the unexpected (Harford 2011). A key challenge in many organizations is to know when to shift from normal procedures to emergency response, which implies that crisis signals have been observed and transmitted in time for quick effective response.

In that context, where operators can face unexpected delays and problems, it is essential to provide people with reasonable amounts of resources and deadlines, and to be willing to make adjustments. Otherwise, agents might satisfy the managers by cutting corners in ways that they may not even imagine until and unless they see consequent failures (Garber and Paté-Cornell 2012). Therefore, when managers set these constraints they have to ask themselves what are their “shadow price”, i.e., by how much would one reduce the failure risk if one relaxed that constraint by one unit (one more day?); or on the contrary, whether one can tighten these constraints at a net benefit.

Clearly Define Leadership and Responsibilities

Key to the effectiveness of managers is their leadership in providing role models, and setting the tone for the organization’s performance. Leadership in a risk management context implies not only having (or deferring to) relevant knowledge and authority but also establishing clear lines of accountability and building trust from the people involved that their leaders can and will make proper and prudent decisions in difficult situations.

Who is responsible for avoiding accidents and mishaps? There are often several lines of responsibility and accountability, which should be properly defined and coordinated. The feeble defense of “responsible but not guilty” was used, for instance, by a high government official head of a health ministry in Europe in 1991, after contaminated blood infected a large number of people. The question of course, is: what constitutes guilt on the part of a leader who fails to define proper procedures and ensure their application? Another failure of leadership can occur when a conflict of authority emerges from a two-head structure. For instance, a surgeon and an anesthesiologist who disagree when neither of them has the ultimate decision making power can cause (and have caused) the death of a patient (Paté-Cornell et al. 1997). It may be possible to pinpoint precisely an error at the bottom of the organizational hierarchy that has led to an accident sequence. (In the case of the Piper Alpha accident in 1988, a young worker made the mistake of leaving the work of fixing a pump unfinished at the end of a day and failed to tag the pump as remaining to be fixed.) But, as in the case of rogue traders, the overall question of supervision, incentives and safety culture emanate directly from the leadership of the company and the regulators.

Leadership is thus a key ingredient of a solid system of risk management decision making in which the decision maker hears the message on time, understands it (and the uncertainties involved if any) and is able and ready to act when needed. The decision makers must be willing to know the truth, to make difficult choices and trade-offs of what is tolerable and what is not, and to decide when it is time to shift from regular operations to crisis management with the ability to make quick, well informed decisions. When the risk is born by a group of people, this requires a collective decision process, able to balance the interests and the safety of different groups, and the overall costs and benefits.

Share Knowledge and Experience Across Organizations

Not all risk management responsibilities can or should be defined within a specific organization. Distributed control of risks, shared among multiple organizations or individuals, also creates a need for legal and institutional frameworks to clearly define and enforce rights and duties. Clarifying whose responsibility it is to avoid risks that arise from joint decisions¹² can reduce the average costs of risk management. To provide a rational basis for coordinating liability and incentives to reduce the costs of risk externalities and jointly caused risks in a society of interdependent agents, one might adopt several possible principles. In the economic analysis of law (Posner 1998), the Learned Hand formula, discussed further in the next chapter, states that parties should take additional care if and only if the expected marginal benefit of doing so exceeds the expected marginal cost (Feldman and Kim 2005). Similarly, the “cheapest cost avoider” principle states that the party who can most cheaply avoid a jointly created risk should do so.

Accidents sometimes reveal the existence of information in some parts of industry that could have saved others. Some organizations successfully permit sharing that critical information. The Institute of Nuclear Power Operations (INPO) provides a practical example of such an organization (Reilly 2013). Created in the wake of the Three Mile Island accident, INPO provides a forum where industry managers can discuss existing problems behind closed doors with the support of the regulator (in the US, the NRC). It has the role of an internal watchdog, regularly rating each power plant. These ratings, in turn, influence the insurance rate of the plants thus promoting strong incentives for excellence in safety. What makes such an organization successful is the combination of peer pressures, of a forum for internal discussion of potential problems, blunt assessment of plant performance and the “teeth” provided by financial incentives. What sometimes makes it difficult to generalize the model is the competition among the organizations involved and the global nature of some industries such as the oil market.

¹²An example is a consumer’s decision to stir his soup with a hair dryer, together with the manufacturer’s decision not to affix a label warning against this use.

Conclusions

Successful risk management is usually a cooperative enterprise. Successful cooperation, in turn, requires moving past blame-casting and excuse-giving to understand the causes and conditions that contribute to catastrophes and improve the system, or conversely, that promote safety in the face of unanticipated challenges. Prominent among the addressable drivers of safety are vigilance and readiness to perceive and respond to anomalies, determination and ability to learn from experience, eagerness to continually probe and update assumptions in light of new information, and capacity to adapt creatively and cooperatively when conditions change. Clear lines of duty and responsibility for risk avoidance, together with discipline and training in following well-conceived procedures and rules for routine safe operation of complex systems, are key contributors to safety cultures that work. At the same time, having the wisdom, incentives, know-how, and experience in team problem-solving required to step outside such boundaries and improvise when needed, is essential for successful risk management in the face of novel threats. These are generally teachable and learnable skills.

We propose that improved practices of risk analysis, quantification and management should be built on technical and cultural foundations, which encompass expertise both in reducing routine risks and in responding to novel ones. Such risk management practices should rely less on blame-casting and excuse-making than in the past. They will need to acknowledge that human error is not necessarily the main driver of failures in an increasingly complex and interconnected world, and that systems should be designed to withstand such errors. Unprecedented hazards, fat-tailed distributions, and risk contagion leading to cascading failures are increasingly recognized as drivers of some of the most conspicuous modern risks, from power outages to epidemics to financial failures. Improved risk management practices should thus increasingly rely on intelligent engagement with our uncertain and changing world. They should build on the key principles we have touched upon: leadership and accountability; robust design (decoupling subsystem whenever possible); vigilant and open-minded monitoring; continual active testing of assumptions and systems; deliberate learning; optimal trade-offs of the costs and benefits of gathering further information before acting; well-trained and disciplined habits of coordination; and ability to cooperate quickly and effectively in response to new threats. These principles have been valuable foundations for effective risk management when they were applied in the past. They should become common practice in the future.

Acknowledgments We thank Warner North for useful suggestions that led to a shorter, clearer exposition.

References

- Augustine NR (1996) Augustine laws. American Institute of Aeronautics and Astronautics, Washington, DC
- BEA, French Accident Investigation Bureau (2002) Final report on the 2000 Concorde accident, Paris France
- Correia AW, Arden Pope C III, Dockery DW, Wang Y, Ezzati M, Dominici F (2013) Effect of air pollution control on life expectancy in the United States: an analysis of 545 U.S. counties for the period from 2000 to 2007. *Epidemiology* 24(1):23–31
- Cox LA Jr (2012a) Reassessing the human health benefits from cleaner air. *Risk Anal* 32 (5):816–829
- Cox LA Jr (2012b) Community resilience and decision theory challenges for catastrophic events. *Risk Anal* 32(11):1919–1934
- Cox LA Jr (2013) Improving risk analysis. Springer, New York
- Epstein W (2011) “A probabilistic risk assessment practitioner looks at the Great East Japan earthquake and tsunami”, a Ninokata Laboratory White Paper. Tokyo Institute of Technology, Tokyo
- European Commission (2000) Communication from the Commission on the precautionary principle. Brussels, Belgium
- Federal Aviation Administration (2013) Aviation rule making advisory committee report. FAA, Washington, DC
- Feldman A, Kim J (2005) The Hand Rule and United States v. Carroll Towing Co. Reconsidered. *Am Law Econ Rev* 7(2):523–543
- Gallo RC (2006) A reflection on HIV/AIDS research after 25 years. *Retrovirology* 3:7
- Garber RG, Paté-Cornell ME (2012) Shortcuts in complex engineering systems: a principal-agent approach to risk management. *Risk Anal* 32(5):836–854. <https://doi.org/10.1111/j.1539-6924.2011.01736>
- Gardner D (2009) The science of fear: how the culture of fear manipulates your brain. Penguin Group, New York
- Hammond JS, Keeney RL, Raiffa H (1998) The hidden traps in decision-making. Harvard Business Review
- Harford T (2011) Adapt: why success always starts with failure. Farrar, Straus and Giroux, New York
- International Seismic Safety Organization (2012) Position statement on earthquake hazard assessment and design load for seismic safety. Arsita, Italy
- Janis IL (1984) Groupthink: psychological studies of policy decisions and fiascoes. Houghton Mifflin, Boston
- Kahneman D (2011) Thinking, fast and slow. Farrar, Straus, and Giroux, New York
- Kahneman and Tversky (1974) Judgment under uncertainties: heuristics and biases. *Science* 85 (4157):1124–1131
- Kaplan S, Garrick JB (1981) On the quantitative definition of risk. *Risk Anal* 1(1):11–27
- Lakats LM, Paté-Cornell ME (2004) Organizational warnings and system safety: a probabilistic analysis. *IEEE Trans Eng Manag* 51(2):183–196
- Lehrer J (2009) How we decide. Houghton Mifflin Harcourt, New York
- Lehrer J (2012) Trials and errors: why science is failing us. Wired January 28, 2012. <http://www.wired.co.uk/magazine/archive/2012/02/features/trials-and-errors?page=all>
- Murphy DM, Paté-Cornell ME (1996) The SAM framework: a systems analysis approach to modeling the effects of management on human behavior in risk analysis. *Risk Anal* 16 (4):501–515
- National Academy of Engineering (2004) Accident precursor analysis and management: reducing technological risk through diligence. The National Academies Press, Washington, DC

- National Academy of Engineering (2012) Macondo well-deepwater horizon blowout: lessons for improving offshore drilling safety, report to the department of interior. The National Academies Press, Washington, DC
- National Commission on Terrorist Attacks Upon the United States (2004) The 9/11 Commission Report, Washington, DC
- National Transportation Safety Board (1989) Aircraft Accident Report- Aloha Airlines, flight 243, Boeing 737-200, - N73711, near Maui, Hawaii, 28 April 1988
- Paté-Cornell ME (1990) Organizational aspects of Engineering System Safety: the case of offshore platforms. *Science* 250:1210–1217
- Paté-Cornell ME (1993) Learning from the Piper Alpha accident: a post-mortem analysis of technical and organizational factors. *Risk Anal* 13(2):215–232
- Paté-Cornell ME (1996) Uncertainties in risk analysis: six levels of treatment. *Reliab Eng Syst Saf* 54:95–111
- Paté-Cornell ME (2009a) Probabilistic Risk Assessment. In: James J. Cochran Editor-in-Chief, The Wiley Encyclopedia of Operations Research and Management Science, Hoboken, Wiley
- Paté-Cornell ME (2009b) Accident Precursors. The Wiley encyclopedia of operations research and management science. In James J. Cochran Editor-in-Chief, Hoboken, Wiley
- Paté-Cornell ME, with the collaboration of Seth Guikema, Paul Kucik, David Caswell and Russ Garber (2012a) Games, risks and analytics: several case involving national security and management situations. *Decis Anal* 9(2):186–203
- Paté-Cornell ME (2012b) On black swans and perfect storms: risk analysis and management when statistics are not enough. *Risk Anal* 32(11):1823–1833
- Paté-Cornell E, Cox LA Jr (2014) Improving risk management: from lame excuses to principled practice. *Risk Anal* 34(7):1228–1239. <https://doi.org/10.1111/risa.12241>
- Paté-Cornell ME, Davis DB (1994) A challenge to the compound lottery axiom: a two-stage normative structure and comparison to other theories. *Theory Decis* 37(3):267–309
- Paté-Cornell ME, Fischbeck PS (1995) Probabilistic interpretation of command and control signals: Bayesian updating of the probability of nuclear attack. *Reliab Eng Syst Saf* 47(1):27–36
- Paté-Cornell ME, Lakats LM, Murphy DM, Gaba DM (1997) Anesthesia patient risk: a quantitative approach to organizational factors and risk management options. *Risk Anal* 17(4):511–523
- Pope CA 3rd, Ezzati M, Dockery DW (2009) Fine-particulate air pollution and life expectancy in the United States. *N Engl J Med* 360(4):376–386
- Posner R (1998) Economic analysis of law, 5th edn. Aspen Publishers, New York, NY
- Reilly W (2013) Valuing safety even when the markets do not notice. In: Goldwyn DL, Kalicki J (eds) Energy and security, 2nd edn. Johns Hopkins University Press, Baltimore
- Russo JE, Schoemaker PJH (1990) Decision traps: ten barriers to brilliant decision-making and how to overcome them. Doubleday, New York
- Sanfey AG, Chang LJ (2008) Multiple systems in decision-making. In: Tucker WT, Ferson S, Finkel A, Long TF, Slavin D, Wright P (eds) Strategies for risk communication: evolution, evidence, experience, vol 1128. Annals of the New York Academy of Science, New York, pp 53–62
- Taleb NN (2007) The Black Swan: the impact of the highly improbable. Random House, New York
- Weick KE, Sutcliffe KM (2007) Managing the unexpected: resilient performance in an age of uncertainty, 2nd edn. Wiley, San Francisco

Chapter 14

Improving Institutions of Risk Management: Uncertain Causality and Judicial Review of Regulations



This chapter continues to consider questions of applied benefit-cost analysis and effective risk management, building on themes introduced in the previous two chapters. It expands the scope of the discussion to include a law-and-economics perspective on how different institutions—regulatory and judicial—involved in societal risk management can best work together to promote the public interest. In the interests of making the exposition relatively self-contained, we briefly recapitulate distinctions among types of causality and principles of causal inference that are discussed in more detail in Chap. 2, as well as principles of benefit-cost analysis and risk psychology, including heuristics and biases, from Chap. 10. In this chapter, however, the focus is less on individual, group, or organizational decision-making than on how rigorous judicial review of causal reasoning might improve regulatory risk assessment and policy.

Law-and-economics suggests principles for deciding how best to allocate rights, duties, and legal liability for actions that cause harm or that fail to prevent it. These principles can be applied to suggest how to coordinate the overlapping activities of regulators and the courts in constraining and penalizing individual behaviors and business activities in an effort to increase net social benefits. This chapter reviews law-and-economics principles useful for benefit-cost analysis (BCA) and judicial review of regulations and public policies; highlights the crucial roles of causation and uncertainty in applying these principles to choose among competing alternatives; and discusses how net social benefits can be increased by applying these same principles to judicial review of regulations that are based on uncertain assumptions about causation of harm.

The real-world examples of food safety regulation and air pollution regulation introduced in Chaps. 5 and 10, respectively, illustrate that deference by the courts (including administrative law judges) to regulators is not likely to serve the public interest when harmful effects caused by regulated activities are highly uncertain and are assessed primarily by regulatory agencies. In principle, responsibility for increasing net social benefits by insisting on rigorous analysis of causality and remaining

uncertainties by both plaintiffs and defendants should rest with the courts. In practice, failure of the courts to shoulder this burden encourages excessive regulation that suppresses socially beneficial economic activities without preventing the harms or producing the benefits that regulators and activists project in advocating for regulation. Stronger judicial review of regulations by courts that require sound and explicit reasoning about causality from litigants is needed to reduce arbitrary and capricious regulations, meaning those in which there is no rational connection between the facts found and the choice made, and to promote the public interest by increasing the net benefits from regulated activities.

Introduction: Principles of Law-and-Economics and Benefit-Cost Analysis of Regulations

Building on principles of benefit-cost analysis (BCA) introduced in Chap. 12, a common normative principle for defining “good” laws and regulations and institutions to administer and enforce them is that they should be designed and operated to *maximize net social benefits*, defined as the sum over all individuals of the difference between expected total benefits and expected total costs (including opportunity costs) received. If time is important, expected net present values of costs and benefits are used. We will call this the *benefit-cost analysis (BCA) principle*. It has been widely used in engineering economics to identify the most beneficial project scales and portfolios of projects to undertake, given budget constraints; in public policy to justify proposed regulatory initiatives and policy interventions; and in healthcare to recommend technologies and practices that are most worth adopting. As illustrated in this section, it can also be viewed as a framework motivating and unifying various law-and-economics principles for using the law to maximize widely distributed social benefits from economic transactions. Later in this chapter, we will propose that it can also fruitfully be applied to the problem of determining how judicial review can and should be used to increase the net social benefits—henceforth called simply the net benefits—of regulations.

More rigorous theoretical formulations often replace expected net benefits with expected social utility (also called social welfare), usually represented as a weighted sum of individual utilities following a theorem of Harsanyi (Harsanyi 1955; Hammond 1992). Various proposals have been advanced for scaling, weighting, eliciting, and estimating individual utilities, despite an impressive body of theory on incentive-compatibility constraints and impossibility theorems for voluntary truthful revelation of private information about preferences, and despite progress in behavioral economics showing that elicited preferences do not necessarily reflect rational long-term self-interest or provide a secure normative foundation for decision-making. Whether the normative criterion is taken to be expected net benefit or expected social utility, its maximization is usually understood to be subject to constraints that protect individual rights and freedoms by preventing systematic

exploitation of any individual or class of individuals to benefit others. In the simplest cases, risk-aversion and other nonlinearities are ignored throughout, and decisions are made simply by comparing the expected net benefits of the alternatives being considered, summed over all the parties involved.

Example: The Learned Hand Formula for Liability Due to Negligence

The Learned Hand formula in law-and-economics holds that a defendant should be considered *negligent*, and hence liable for harm that his failure to take greater care in his activities has caused to a plaintiff, if and only it would have cost the defendant less to take such care than the expected cost of harm to the plaintiff done by not taking it, namely, the probability of harm to the plaintiff times the magnitude of the harm if it occurs. In symbols, the defendant should be found negligent if and only if $C < pB$, where C is the cost to the defendant of taking an expensive action to prevent the harm from occurring, p is the probability of harm occurring if that action is not taken, and B is the cost, or severity of the harm to the plaintiff expressed in dollars, if it does occur—and hence the benefit from preventing it if it would otherwise have occurred. The Learned Hand rule defines negligence as failure to take care not to harm another when the expected net social benefit of doing so is positive. It encourages each participant in a society of laws to consider the expected cost of one's actions to others as of equal weight in deciding what to do as costs or benefits to one's self. This dispassionate, equal weighting of the interests of all is precisely what is required for individual choices to maximize net social benefit, i.e. the sum of the gains and losses of all participants.

The Learned Hand rule encourages economic agents to act as utilitarians in decision-making, counting the consequences to others as of equal weight with consequences to self. By doing so, it also provides a utilitarian rationale for the calculus of negligence in tort law as promoting the public interest, construed as maximizing net social benefit. In other words, the rationale for individual choices promoted by the rule of negligence law and the rationale for adopting that law itself are the same: to maximize total (and hence average) net benefits from interactions among individuals in society. If each individual is as likely to be a potential defendant as a potential plaintiff in the many individual transactions and situations to which the law of negligence applies, then each expects to gain on average from the rule of this law.

The BCA principle of deciding what to do by comparing expected net social benefits from alternative choices (see Chap. 12) can be used by individuals to choose among different actions and levels of care, given a legal and regulatory system. It can also be used as a basis for designing institutions, for example, by choosing among alternative liability rules and regulations based on the behaviors that they are expected to elicit in response to the incentives they create, with the goal of

maximizing the estimated net social benefits arising from those behaviors. The BCA principle can be generalized well beyond the domain of negligence liability. BCA comparisons provide a widely applicable rationale for determining the legal duties of economic actors to take due care in their activities, refrain from excessively hazardous behaviors, provide pertinent information to others about risks, and bear liability for harm arising at least in part from their activities.

Example: The Cheapest Cost-Avoider Principle When Liability Is Uncertain

Suppose that either of two parties, a producer and a user or consumer of a potentially hazardous product such as a lawnmower, an electrical hair dryer, or a medical drug, can choose between taking more care or less care to prevent harm to the consumer from occurring. Taking more care is more costly than taking less care to the agent making that choice. If harm occurs that either party could have prevented by taking more care in manufacture or use, respectively, then how should liability for the harm be allocated between them? A standard answer when the respective costs for each party to have prevented the harm are common knowledge is the *cheapest cost avoider principle*: the party who should bear liability for the harm is the one who could have prevented it most cheaply. Like the Learned Hand principle, this again requires each agent, i.e., each of the two parties, to consider a dollar of cost to another to have the same importance as a dollar of cost to one's self in calculating net social benefits. This implies that the agent who can most cheaply avoid a harm or loss should do so. When the relevant costs are not common knowledge, however, this simple principle does not necessarily hold. For example, if each agent assesses a positive probability that the other will be found liable in the event of an accident that harms the consumer in using or consuming the product, then both may under-invest in safety, increasing the probability of harm above what it would be if costs and liability were common knowledge, and thereby reducing net social benefit. In this case, a regulation that requires the producer to use the higher level of care, combined with credible enforcement of the regulation via random inspections and large penalties for detected violations, can in principle increase the net benefits to both producers and consumers by making it common knowledge that it is in the producer's best interest to take a high level of care. Then consumers who might otherwise have been unwilling to buy the product because of fear that it is unsafe might be willing to buy it, and even to pay more for its higher level of safety, thus increasing both producer surplus and consumer surplus compared to the situation without regulation. A strict liability standard that makes it common knowledge that it is in the producer's best interest to take a high level of care could create the same benefits. In both cases, however, shifting all responsibility for safety to the producer could make an otherwise beneficial product too costly to produce, even if both producer and consumer would benefit if the consumer could be trusted to take due

care in using it. In general, designing liability standards and regulations to maximize total social benefits requires considering who knows what about care levels and product risks, the costs of acquiring such information, possibilities for credibly signaling it, and moral hazard and incentives. In practice, a mix of market, liability, insurance, warranty, and regulatory instruments is used to deal with these realistic complexities.

More generally, not only producers and consumers of potentially hazardous consumer products or services, but also employers and employees in hazardous occupations, owners and renters or lesers of properties, sellers and buyers of used cars or mortgages or insurance products or collateralized debt obligations, and owners and neighbors of hazardous or noxious facilities, can all be viewed as making choices that jointly affect net social benefits. The obligations and penalties imposed by legal and regulatory institutions can then be viewed from the law-and-economics perspective as seeking to promote choices to maximize widely distributed net social benefits. In turn, these institutions themselves, and the ways in which they interact with each other and with the public, can also be designed and evaluated by this criterion.

A powerful intuition underlying these applications of BCA principles to liability and regulation of economic activities is that a nation of laws serves its citizens best by implementing just those laws and policies whose total benefits outweigh their total costs, including opportunity as well as implementation and enforcement costs. To protect the rights of individuals, costs and benefits should also be distributed so that everyone expects to gain over a lifetime from application of the adopted laws and policies, even though specific individuals lose in specific cases, as when a court determines that one litigant and not the other wins a tort law case. This distribution requirement prevents adoption of laws and policies that systematically exploit one individual or class of individuals to benefit the rest. It encourages laws and regulations that, arguably, might be collectively chosen from behind a Rawlsian veil of ignorance (Hammond 1992). Identifying collective choices with positive net benefits and acceptable distributions of costs and benefits is a central goal of BCA.

Despite their intuitive appeal, making these principles precise is famously difficult. A vast literature in collective choice theory considers how to aggregate individual preferences, beliefs, and utilities to yield social utility functions to maximize, or at least decision rules for making Pareto-efficient collective decisions that would increase the net benefits to all individuals. This literature has yielded a rich variety of impossibility results and tradeoffs among proposed criteria for evaluating the performance of aggregation procedures. Criteria that have been considered include voluntary participation, freedom to have and to express any individual preferences or utility functions in a large set, incentive-compatibility in revealing the private information needed to make the process work (e.g., willingness-to-pay information), Pareto-efficiency of outcomes, complexity of the procedure (e.g., yielding a result in less than the lifetime of the participants), and balanced budget, if the procedure is costly to administer. For rational participants, only some subsets of such desiderata are mutually consistent. Some must be relinquished to obtain the rest.

As a practical matter, however, real people often do not reason or behave like the idealized rational agents to which such results apply (Thaler 2015). They are often more altruistic and cooperative than purely economic or game-theoretic reasoning would predict (Kahneman 2011). Thus, the question of how well different legal and regulatory institutions elicit choices and behaviors from real people that increase net social benefits is worth empirical as well as theoretical investigation. Likewise, even if economic agents have little or no choice about how to comply with regulations, other than perhaps the option of suing for relief if the regulations seem arbitrary and capricious or otherwise unjust, how well regulatory and legal institutions succeed in developing and enforcing policies that increase net social benefits—and how they might work together to do so better—are questions well worth empirical as well as theoretical investigation.

Uncertain Causation Encourages Ineffective and Potentially Harmful Regulations

The central problem to be investigated in the remainder of this chapter is *how to identify socially beneficial choices when causation is uncertain* and the sizes of the benefits caused by alternative choices are therefore unknown. In many important applications, the benefits caused by costly regulations, policy interventions, or investments are uncertain. Whether they exceed the costs may then also be uncertain. In such settings, BCA principles must be modified to deal with risk aversion, rather than considering only expected net benefits. Moreover, uncertainty about causation can encourage regulations that are socially harmful and that would not pass a BCA test that accounts for correlated uncertainties about its effects on many individuals. This creates a need for review and accountability that regulatory agencies are not well equipped to provide for themselves. The following paragraphs develop these points.

Uncertain Causation Encourages Socially Reckless Regulation

To understand how uncertainty about causation can lead to adoption of regulations whose risk-adjusted costs exceed their benefits, suppose that a regulatory ban or restriction on some activity or class of activities, such as emissions of a regulated air pollutant, imposes an expected cost of c on each of N economic agents and yields an uncertain benefit of b for each of them, with the expected value of b denoted by $E(b)$ and its variance denoted by $Var(b)$. If the uncertain benefit b has a normal distribution, then a standard calculation in decision analysis shows that its certainty-equivalent value to a decision maker with an exponential utility function is

$$CE(b) = E(b) - k^*Var(b),$$

where k is proportional to the decision maker's risk aversion. To a risk-averse decision maker (i.e., having $k > 0$), the uncertain benefit is worth less than a certain benefit of size $E(b)$ by the amount of the "risk premium," $k^*Var(b)$. If the uncertainty about b is due to uncertainty about the size of the effect that a policy or regulation would cause, e.g., the size of the reduction in annual mortality risk that would be achieved by a reduction or ban on a source of exposure, and if the size of this uncertain effect is the same for all N agents, then the net social benefit summed over all N agents is

$$N^*E(b) - N^2kVar(b) - Nc$$

(since the variance of N times b is N^2 times the variance of b). This can be written as

$$N^*[E(b) - c - NkVar(b)],$$

showing that the *per capita* net benefit after adjusting for risk aversion is $[E(b) - c - NkVar(b)]$. This will necessarily be negative if N is sufficiently large, since $kVar(b) > 0$. However, a regulatory agency that implements regulations with expected benefits exceeding their expected costs, i.e., with $E(b) - c > 0$, pays no attention to the risk premium term $NkVar(b)$ or to the size of N . It ignores the fact that individual benefits received are positively correlated, since a regulation that does not cause its intended and expected benefit can result in a net loss for each of the N agents. For all sufficiently large N , the risk of these N simultaneous losses will outweigh the positive expected net benefit, i.e., $E(b) - c - NkVar(b)$ will be negative even though $E(b) - c$ is positive. (The Arrow-Lind theorem showing that government investments should be risk-neutral does not apply here, due to the correlated losses.) Agencies that focus only on expected benefits can therefore undertake activities or impose regulations whose risk-adjusted values are negative. This is most likely for regulations with widely distributed but uncertain benefits (i.e., large N), such as air pollution or food safety regulations. In essence, ignoring risk aversion to correlated losses if the causal hypothesis that the regulation will create its intended benefits turns out to be wrong encourages reckless expenditures and costly regulations with large net losses if the expected benefits do not occur.

Warnings from Behavioral Economics and Decision and Risk Psychology: The Tyranny of Misperceptions

Regulatory agencies, as well as courts and corporations, are staffed by human beings with a full array of psychological foibles—heuristics and biases such as those discussed in Chap. 12—that shape their beliefs and behaviors in addressing uncertain risks. Well-documented weaknesses in individual and group decision-making under uncertainty include forming opinions and take actions based on too little

information (related to Kahneman's "what you see is all there is" heuristic) (Kahneman 2011); making one decision at a time in isolation rather than evaluating each in the context of the entire portfolio or stream of decisions to which it belongs (narrow framing); believing what it pays us to believe, what it feels most comfortable to believe, or what fits our ideological world view (motivated reasoning, affect heuristic); seeking and interpreting information to confirm our existing opinions while failing to seek and use potentially disconfirming information (confirmation bias); and being unjustifiably confident in both our judgments and our level of certainty about them (overconfidence bias) (Kahneman 2011; Schoemaker and Tetlock 2016; Thaler 2015). In short, judgments and choices about how to manage risks when the effects of actions are highly uncertain are often shaped by emotions and psychology ("System 1") far more than by facts, data, and calculations ("System 2"). Such decisions can be passionately advocated, strongly felt to be right, and confidently approved of without being causally effective in producing desired results. These common weaknesses of human judgment and opinion formation have most opportunity to shape policy when the consequences caused by different choices are least certain.

Regulatory agencies face additional challenges to learning to act effectively under uncertainty, stemming from social and organizational psychology. Consider the following selection mechanism by which people with similar views might assemble into a regulatory agency with an organizational culture prone to groupthink (Coglianese 2001, p. 106) and holding more extreme perceptions of the risks that they regulate than most of the population. Suppose that people whose ideologies and beliefs suggest that it is highly worthwhile to regulate substance, activity, or industry X are more likely to work for agencies that regulate X than are people who do not share those beliefs. Then an organizational culture might develop that is inclined to regulate well beyond what most people would consider reasonable or desirable. The result is a *tyranny of misperceptions*, somewhat analogous to the Winner's Curse in auctions, in which those whose perceptions are most extreme are most likely to invest the time and effort needed to stimulate regulatory interventions. In this case, when the true hazards caused by a regulated substance or activity are highly uncertain, those who believe them to be worse than most people judge them to be may be disproportionately likely to shape the regulations that restrict them. If average judgments are typically more accurate than extreme ones (Tetlock and Gardner 2015), such regulations may tend to reflect the misperceptions of those advocating regulation that risks are higher than they actually are.

Of course, the same self-selection bias can function throughout the political economy of a democracy: those who care most about making a change are most likely to work to do so, whether through advocacy and activism, litigation, regulation, legislation, or journalism directed at influencing perceptions and political actions. But regulatory agencies are empowered to take costly actions to promote social benefits even when the benefits caused by actions are highly uncertain, so that decisions are most prone to System 1 thinking. Moreover, empirical evidence suggests that simply recognizing this problem and making an effort to correct for it, e.g., by instituting internal review procedures, is likely to have limited value: the

same heuristics and biases that give rise to a policy are also likely to affect reviews, making misperceptions about risk and biased judgments about how best to manage them difficult to overcome (Kahneman 2011). External review by people who do not share the same information and world view can be far more valuable in flagging biases, introducing discordant information to consider, and improving the effectiveness of predictions and decisions (Kahneman 2011; Tetlock and Gardner 2015).

It is distressing, and perhaps not very plausible *a priori*, to think that people and organizations that devote themselves to promoting the public interest might inadvertently harm it by falling into the familiar pitfalls of System 1 thinking (Chap. 10) when causation is uncertain. After all, do not well-run organizations anticipate and correct for such limitations by using relatively explicit and objective criteria and rationales for their decisions, well-documented reasoning and data based on peer-reviewed publications, multiple rounds of internal review, and invited critiques and reviews by external experts and stakeholders? Indeed, all of these steps play important roles in modern rule-making procedures and regulatory procedures in the United States and elsewhere. Yet, there is evidence that they are not sufficient to guarantee high-quality regulations, or to block regulations for which there is no good reason to expect that the predicted benefits will actually occur. Such regulations are too often “arbitrary and capricious” in the sense that there is no rational connection (although there may be many irrational ones) between the facts presented to support projected benefits of regulations and the belief that these benefits will actually occur, and hence that regulations are worthwhile. The following examples illustrate the real-world importance of these concerns. Possible explanations and remedies will then be explored, including judicial review that insists on more rigorous and trustworthy standards of evidence for causality than regulatory agencies customarily use. We will argue that courts are often the “cheapest misperception corrector” and are best positioned to correct regulatory excesses by enforcing a higher standard of causal inference before uncertain benefits are accepted as justifying costly actions.

Example: The Irish Coal-Burning Bans

Recall from Chap. 1 and several subsequent discussions that, between 1990 and 2015, coal-burning was banned by regulators in many parts of Ireland, based on beliefs that the local government summarized as follows:

“Benefits of a smoky coal ban include very significant reductions in respiratory problems and indeed mortalities from the effects of burning smoky coal. The original ban in Dublin has been cited widely as a successful policy intervention and has become something of an icon of best practice within the international clean air community. It is estimated that in the region of 8,000 lives have been saved in Dublin since the introduction of the smoky coal ban back in 1990 and further health, environmental and economic benefits (estimated at 53m euro per year) will be realised, if the ban is extended nationwide” (Department of Housing, Planning, Community, and Local Government 2016).

The underlying scientific studies that led to these beliefs (Clancy et al. 2002), still widely and approvingly cited by regulators and activists, clearly showed that particulate matter levels from coal smoke and mortality rates dropped significantly after the bans. For over a decade, activists, regulators, and the media have celebrated such findings as showing a clear causal link between coal burning and mortality and a clear opportunity to reduce mortality by reducing coal burning, leading to substantial estimated health and economic benefits from extending the bans nation-wide and substantial unnecessary deaths from delaying (Kelly 2016).

Yet, as emphasized in Chaps. 1 and 2, there is a clear logical fallacy at work here. Although the claimed successes of the bans in reducing mortality appear might appeal to common sense, wishful thinking, and confirmation bias, no potential *disconfirming* evidence that might conflict with this causal conclusion was sought or used in the studies that led to the claim. For example, the original study (Clancy et al. 2002) did not examine whether the drop in mortalities following the bans had causes unrelated to the ban, nor whether it also occurred in other countries and in areas unaffected by the bans (Wittmaack 2007). When a team including some of the original investigators examined these possibilities a decade later, long after successive waves of bans had already been implemented, they found no evidence that the bans had actually caused the reductions in total mortality rates that had originally been attributed to them: mortality rates had fallen just as much in areas not affected by the bans as in areas affected by them (Dockery et al. 2013). The bans had no detectable effect on reducing total mortality rates. Rather, mortality rates had been declining over time throughout Ireland and much of the developed world since long before the bans began, and they continued to do so without interruption during and following them. Thus, mortality rates were indeed lower after the bans than before them, even though the bans themselves had no detectable effect on them. (Data-dredging that sought associations of pollution reductions with reductions in cause-specific mortality rates, but without controlling for multiple testing bias, found a few associations, but disconfirmed the original claim of significant reductions in cardiovascular mortality associated with the bans.) If the ban left more elderly people cold in the winter, and thereby increased their mortality rates—a possibility that was not investigated—then this effect was masked by the historical trend of improving life expectancy and reduced mortality risks.

In this example, it appears that confirmation bias led to widespread and enthusiastic misinterpretation of an ongoing historical trend of declining elderly mortality rates—brought about largely by improved prevention, diagnosis, and treatment of cardiovascular diseases and reduced cigarette smoking—as evidence that the coal-burning bans were causally effective in protecting public health. This meme has continued to drive regulatory and media accounts and regulatory policy until the present as Ireland pushes to extend the bans nation-wide (Kelly 2016). The finding that the bans actually had no detectable effect in reducing all-cause or cardiovascular mortality (Dockery et al. 2013) continues to be widely ignored. This example illustrates how regulations can be enthusiastically supported and passed based on unsound reasoning about causality, such as neglecting to use control groups in assessing the effects of bans. It also shows how they can be perceived and evaluated

favorably in retrospect by regulators, activists, environmental scientists, and the media as having been highly successful in creating substantial public health benefits, even if they actually had no beneficial effects.

Example: Estimated Benefits of Fine Particulate Matter (PM2.5) Regulation in the United States

An analogous process is currently unfolding on a much larger scale in the United States. The majority of total estimated benefits from all Federal regulations in the U.S. are attributed to the effects of Clean Air Act regulations in reducing fine particulate matter (PM2.5) air pollution and thus reducing estimated elderly mortality risks. The United States Environmental Protection Agency (EPA) credits its regulation of fine particulate matter with creating nearly two trillion dollars per year of health benefits (EPA 2011a, b). Yet, notwithstanding widespread impressions and many published claims to the contrary in scientific journals and the news media, it has never been established that reducing air pollution actually *causes* these benefits, as opposed to being associated with them in a historical context where both air pollution levels and mortality rates have been declining over time. As the EPA's own benefits assessment states in a table, their "analysis assumes a causal relationship between PM exposure and premature mortality based on strong epidemiological evidence... However, epidemiological evidence alone cannot establish this causal link" (EPA 2011a, b, Table 5-11) (EPA 2011b). The reason that the epidemiological evidence cannot establish the assumed causal link is that it deals only with association, and not with causation, as detailed in Chap. 2.

In the absence of an established causal relation, historical data showing that both PM2.5 and mortality rates are both higher than average in some places and times (e.g., during cold winter days compared to milder days, or in earlier decades compared to later ones), and thus that they are positively correlated, is widely treated as if it were evidence of causation. This again illustrates the practical importance of confirmation bias in shaping perceptions and economic evaluations of the benefits attributed to (but not necessarily caused by) major regulations by those advocating them. Potential disconfirming evidence, such as that mortality risks declined just as much in cities and counties where pollution levels increased as where they decreased (see Chap. 10), has been neither sought nor used by advocates of PM2.5 reductions in attributing health and benefits to such reductions. As pointed out recently, "Many studies have reported the associations between long-term exposure to PM2.5 and increased risk of death. However, to our knowledge, none has used a causal modeling approach" (Wang et al. 2016). The relatively rare exceptions that report positive causal relations rest on unverified modeling assumptions to interpret associations causally, as discussed in greater detail later. Approaches that seek to avoid making such assumptions by using nonparametric analyses of whether changes in exposure concentrations predict changes in mortality rates have concluded that

“A causal relation between pollutant concentrations and [all-cause or cardiovascular disease] mortality rates cannot be inferred from these historical data, although a statistical association between them is well supported” (Cox and Popken 2015) and that, for 100 U.S. cities with historical data on PM2.5 and mortality, “we find no evidence that reductions in PM2.5 concentrations cause reductions in mortality rates” (Cox et al. 2013).

On the other hand, hundreds of peer-reviewed articles and media accounts claim that reducing PM2.5 causes reductions in mortality risks (Wang et al. 2016). These often present sensational conclusions such as that “An effective program to deliver clean air to the world’s most polluted regions could avoid several hundred thousand premature deaths each year” (Apte et al. 2015). Similar to the original mistaken claims about effects of coal-burning bans on all-cause mortality risks in Ireland, such conclusions conflate correlation and causation. This confusion is facilitated by the increasing use of computer models to project hypothetical benefits based on assumptions of unknown validity. For example, the EPA provides a free computer program, BenMAP, to enable investigators to quantify the human health benefits attributed to further reductions in criterion air pollutants such as ozone (O3) and fine particulate matter (PM2.5) based on embedded expert opinions about their concentration-response correlations. Activist organizations such as the American Thoracic Society (ATS) have used BenMAP simulations to develop impressive-looking and widely publicized estimates of health benefits from further reductions in air pollution, such as that “Approximately 9320 excess deaths (69% from O3; 31% from PM2.5), 21,400 excess morbidities (74% from O3; 26% from PM2.5), and 19,300,000 adversely impacted days (88% from O3; 12% from PM2.5) in the United States each year are attributable to pollution exceeding the ATS-recommended standards” (Cromar et al. 2016). But the concentration-response relations assumed in the computer simulations are not established causal relations. To the contrary, as clearly and repeatedly stated in the technical documentation for BenMAP (March 2015, Table E-1, Health Impact Functions for Particulate Matter and Long-Term Mortality, pp 60–61), there is “no causality included” in BenMAP’s summary of health impact functions based on expert judgments. In more detail, the documentation explains that “Experts A, C, and J indicated that they included the likelihood of causality in their subjective distributions. However, the continuous parametric distributions specified were inconsistent with the causality likelihoods provided by these experts. Because there was no way to reconcile this, we chose to interpret the distributions of these experts as unconditional and ignore the additional information on the likelihood of causality.” Similar caveats hold for other instances of the increasingly prevalent practice of predicting reductions in mortality caused by reductions in exposure concentration by applying previously estimated concentration-response associations and slope factors, without any independent effort to establish whether they are causal. For example, Lin et al. (2017) “estimate the number of deaths attributable to PM2.5, using concentration-response functions derived from previous studies” and conclude “that substantial mortality reductions could be achieved by implementing stringent air pollution mitigation measures” without noting that the previous studies referred to only assessed associations and not causation.

In summary, similar to the case of the coal-burning bans in Ireland, substantial health benefits are attributed to tighter Clean Air Act regulations in the United States, with many calls for further reductions being voiced by activists, regulators, public health researchers and advocacy groups, and the media. Yet, it has not been shown that the regulations actually cause the benefits that are being attributed to them, and causal analysis approaches that do not make unverified modeling assumptions do not find any detectable beneficial effect of reducing current ambient concentrations of PM2.5 or ozone in recent decades, despite a voluminous scientific and popular literature projecting substantial health benefits that should be easily detectable if they were occurring (Cox and Popken 2015).

Example: Food Safety Regulation Based on Assumed Causation

Between 2000 and 2005, the Food and Drug Administration's Center for Veterinary Medicine (FDA-CVM), in conjunction with activist and advocacy organizations such as the Alliance for Prudent Use of Antibiotics (APUA) and the Union of Concerned Scientists, successfully pushed to ban the antibiotic enrofloxacin, a fluoroquinolone antibiotic, from use in chickens, on the grounds that its use might select for antibiotic-resistant strains of the common bacterium *Campylobacter*, potentially causing cases of antibiotic-resistant food poisoning that would be more difficult to treat than non-resistant cases. This concern certainly sounds plausible. It received extensive media coverage via stories that usually linked it to frightening statistics on the tens of thousands of cases per year of “superbug” infections with multi drug resistant bacteria occurring in the United States. Few stories explained that those cases were from different bacteria, not from *Campylobacter*; that campylobacteriosis was specifically associated with consuming undercooked chicken in fast food restaurants, and not with chicken prepared at home or in hospitals; or that molecular fingerprinting showed that superbug infections overwhelmingly were caused by hospital use of antibiotics in people, rather than by animal antibiotic use on the farm. A quantitative risk assessment model used by the FDA simply *assumed* that reducing use of enrofloxacin in chickens would proportionally reduce the prevalence of fluoroquinolone-resistant cases of campylobacteriosis food poisoning: “A linear population risk model used by the U.S. Food and Drug Administration (FDA) Center for Veterinary Medicine (CVM) estimates the risk of human cases of campylobacteriosis caused by fluoroquinolone-resistant *Campylobacter*. Among the cases of campylobacteriosis attributed to domestically produced chicken, the fluoroquinolone resistance is assumed to result from the use of fluoroquinolones in poultry in the United States” (Bartholomew et al. 2005). This assumption swiftly made its way into risk numbers cited in activist reports and media headlines, where it was treated as a fact.

Industry and animal safety experts argued that this causal assumption was amply refuted by real-world data showing that the strains of fluoroquinolone-resistant campylobacter found in people were hospital-acquired and not the same as those from animals; that campylobacteriosis was usually a self-limiting disease that caused diarrhea and then resolved itself, with no clear evidence that antibiotic therapy made any difference; that in the rare cases of severe infections, typically among AIDS patients or other immunocompromised people, physicians and hospitals did not treat campylobacteriosis with fluoroquinolones but generally prescribed a different class of antibiotics (macrolides); that even when fluoroquinolones (specifically, ciprofloxacin) was prescribed as empiric therapy, resistance did not inhibit its effectiveness because therapeutic doses are high enough to overcome the resistance; that, evidence from earlier antibiotic bans for farm animals in Europe showed that reducing use in animals increased illnesses in animals (and hence total bacterial loads on meat) but did not benefit human health; that fluoroquinolone-resistant strains of campylobacter occur naturally whether or not enrofloxacin is used and would persist even if it were withdrawn; and that the main effect of continued use of enrofloxacin was to keep animals healthy and well-nourished, reducing risks of foodborne bacteria, both resistant and non-resistant. These arguments were heard by an Administrative Law Judge (ALJ), a career FDA employee with a record of deciding cases in favor of his employer. The ALJ found the industry arguments unpersuasive, and the FDA withdrew approval of enrofloxacin use in poultry in 2005. Meanwhile, during the run-up to this decision from 2000 to 2005, consumer advocacy groups scored major successes in persuading large-scale food producers and retailers to reduce or eliminate use of antibiotics in chickens. Advocates for bans on animal antibiotics, from the Centers for Disease Control and Prevention (CDC), APUA, and elsewhere, many of whom had testified for FDA, quickly declared the enrofloxacin ban a “public health success story” in both the press and in scientific journals (Nelson et al. 2007).

After more than a decade, the causal aspects of this case are easier to see clearly. Already by 2007, some of the researchers who had most strongly advocated the ban were beginning to observe that the original FDA assumption that withdrawing enrofloxacin would reduce fluoroquinolone-resistant *Campylobacter* proportionally now appeared to be mistaken: the resistant strains persisted, as industry had warned (Price et al. 2007). By 2016, it was clear that the dramatic improvements in food safety and reductions in campylobacteriosis risk in the population that had been taking place prior to the voluntary cessations of antibiotic use in farm animals and the enrofloxacin ban, including a nearly 50% reduction in risk between 1996 and 2004, had stopped and reversed course, as shown in Fig. 14.1. Advocates who had been vocal between 2000 and 2005 in explaining to Congress and the public why they thought that banning enrofloxacin would protect public health moved on to advocate banning other antibiotics. No *post mortem* or explanation has yet been offered for the data in Fig. 14.1.

Yet, understanding why the enrofloxacin ban failed to produce the benefits that had been so confidently predicted for it (or, if benefits did occur, why they are not more apparent) might produce valuable lessons that would help future efforts to protect public health more effectively. Such lessons remain unlearned when the

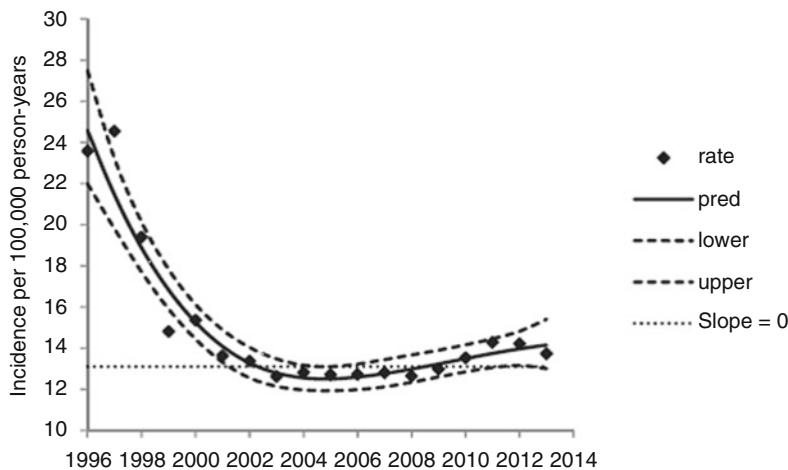


Fig. 14.1 Reductions in campylobacteriosis risk stopped and reversed around 2005. *Source:* Powell (2016)

process for passing new regulatory actions relies on unproved causal assumptions for purposes of advocacy and calculation of hypothetical benefits of regulation—essentially, making a prospective case for regulation—with no need to revisit assumptions and results after the fact to assess how accurate they were or why they failed, if they prove inaccurate. In this example, it appears that the FDA's causal assumption that risk each year is proportional to exposure (Bartholomew et al. 2005) was simply mistaken (Price et al. 2007). But there is no formal regulatory process at present for learning from such mistakes, for correcting them, or for preventing them from being made again in future calls for further bans.

Lessons from the Examples

The foregoing examples illustrate that regulators and administrative law judges sometimes deal with uncertainties about the benefits caused by a regulation by making large, unproved, simplifying assumptions. This may be done with the best of intentions. Uncertainty invites Rorschach-like projection of beliefs and assumptions based on the rich panoply of System 1 (“Gut”) thinking (see Chap. 12), genuinely felt concerns about the currently perceived situation, and hopes to be able to improve it by taking actions that seem sensible and right to System 1. Such projection often feels like, and is described as, careful and responsible reflection and deliberation followed by formation of considered expert judgments based on careful weighing of the totality of the evidence. The resulting judgments are typically felt to be valuable guides to action under uncertainty, not only by those who provide them, but also by those who receive them (Tetlock and Gardner 2015). Beliefs suggested

by System 1 (“Gut”) in the absence of adequate data or opportunity for System 2 (“Head”) analysis are often confidently held, easy to reinforce with confirming evidence, and difficult to dislodge with disconfirming evidence—but they are also often objectively poor guides to what will actually happen (*ibid*, Gardner 2009).

For air pollution and food safety alike, one such large assumption about causation is that *risk of adverse health effects decreases in proportion to reductions in exposure* to a regulated substance or activity. This is easy to understand and appeals to intuition. It leads to readily calculated predictions based on aggregate data: simply divide estimated cases of adverse health outcomes per year by estimated average annual exposure, if exposure is assumed to be the sole cause of the adverse health effects, as in the FDA example. Otherwise, regress adverse health outcomes against exposure, allowing for an intercept and other hypothesized contributing factors to explain any cases not attributed to exposure, as in air pollution health effects modeling. Either way, there is no need for complex modeling of effects of different combinations of risk factors for different individuals, or of interactions and dependencies among the hypothesized explanatory variables, as in the causal DAG models of Chap. 2. Such simplicity is often seen as a virtue (Bartholomew et al. 2005) rather than as omitting the very details that are essential for correctly understanding and quantifying effects caused specifically by exposure, and not by other factors with which it is associated. Assuming that all-cause or cardiovascular disease mortality risks will decrease in direct proportion to reduction of ambient concentrations of PM2.5 in air, or that drug-resistant foodborne illness counts will decrease in proportion to reduction of antibiotic used on the farm, provides simple, plausible-seeming slope factors for calculating hypothetical benefits of further regulation without the difficulty of building and validating models of a more complex reality.

Historically, the resulting numbers have been sensational enough to garner prominent coverage in both scientific journals and popular media, where they are usually presented as if they were facts rather than assumptions (e.g., Cromar et al. 2016). Such coverage attracts the anxiety and resolution of activists to take action to reduce exposures and encourages funding from agencies and other stakeholders to support further, similar assumption-driven research on how large the benefits of regulation might be. This cycle, and associated phenomena such as the social amplification of perceived risks as concern attracts more concern, are well documented in the social science and psychology of risk (Gardner 2009). They are well served by simple assumptions and large risk numbers. By contrast, more complex and nuanced System 2 calculations suggesting that the quantitative difference in public health made by reducing exposures is at most vanishingly small (e.g., on the order of at most one extra case of compromised treatment of campylobacteriosis per 100 million person-years, and plausibly zero (Hurd and Malladi 2008)) typically attract far less attention, and may be viewed with suspicion because they require more detailed data and calculations (Bartholomew et al. 2005).

The “risk reduction is proportional to exposure reduction” formulation of regulatory benefits encourages another System 1 habit that makes life seem simpler and more manageable (Kahneman 2011): narrowly focusing on just what one cares about and what one can do about it. For example, in banning enrofloxacin, the FDA

focused exclusively on preventing cases of drug-resistant food poisoning by controlling what they could control—use of an animal antibiotic. The historical evidence from Europe that such control caused no detectable reductions in human illness risks was irrelevant for this focus, as FDA's risk-is-proportional-to-exposure model assumes no other possibilities. That adequate cooking of meats prior to consumption is the only known control measure that demonstrably reduces illness risks was likewise irrelevant for an agency that does not oversee food preparation. It was excluded from the FDA risk assessment by considering only the ratio of drug-resistant illnesses to drug use on the farm. Similarly, estimates of human health benefits from reducing PM2.5 have seldom inquired about other effects, such as whether cleaner air promotes warmer temperatures, with consequent man-made climate change implications for economic and human health risks.

In summary, although a sound BCA approach unambiguously requires assessing the *total* costs and benefits caused by a regulation, regulatory agencies, like most of us, cope with uncertainty and complexity in the causal impacts of actions by adopting narrowly focused agendas, restricted jurisdictions, and greatly simplified causal models that focus on just a few things. These typically include some actions that we can take (never mind that other, less costly, actions by us or others might work much better); the consequences we want them to produce (never mind their unintended, unanticipated, or out-of-scope consequences); and at most a very few other factors (never mind the large, complex, and uncertain outside world that may make the consequences of our actions quite different from what was intended or predicted). It is much easier to understand and make predictions with these simplified models than to develop and validate more complex and realistic causal models (Kahneman 2011). Disregarding or downplaying most of the causal web in which our potential actions and desired consequences are embedded makes the effects of our own actions, and their potential benefits, loom larger in our calculations than they really are. This very human tendency to substitute simplified causal models for fuller and more realistic ones in the face of uncertainty and complexity (Kahneman 2011; Thaler 2015) is inconsistent with the requirements of the BCA principle, but may be the best that can be done in the absence of institutions that enforce a higher standard.

Better Causal Inferences and Benefits Estimates via More Active Judicial Review

If regulations are sometimes advocated based on overly optimistic and simplistic causal assumptions and models of their effects and the benefits that they cause, what can and should be done about it—and by whom? How might causal inferences and benefits estimates used in regulatory proceedings be made more accurate and trustworthy? This section develops the following points.

- Once the most relevant concept of causation, as identifying *actions that change the probabilities of preferred outcomes* (manipulative causation) has been clearly defined, such improvements in predicting or assessing the benefits caused by regulations are indeed technically possible, based on experience in a variety of other areas.
- It is (or should be) well within the competence and jurisdiction of courts to help bring about these improvements by exercising more stringent judicial review of the causal reasoning used to project benefits and advocate for regulations. This is especially so if current forms of deference to regulatory agencies are replaced by a more active role, as urged by the recently proposed Separation of Powers Restoration Act amendment to the Administrative Procedures Act (Walker 2016).
- The organizational culture of many regulatory agencies makes it difficult for them to improve their own causal assumptions and benefits assessments without the compelling presence of an active judiciary engaged in questioning and challenging their reasoning. In part, this is because of a tendency to dismiss as non-scientific or non-expert the concerns of laypeople and other stakeholders that regulations will not produce their intended benefits (Wynne 1993). In part it arises because regulators use frameworks that treat causality as a matter for expert judgment rather than as a matter of empirically discoverable and verifiable fact.
- To overcome these obstacles, it is both necessary and practical to inject more data-driven rather than judgment- and assumption-driven concepts and techniques for assessing causation into deliberations over regulations.

Advances in data science and analytics make it technically possible, and even easy, to test whether necessary conditions for causality, such as that a cause should help to predict its effects, hold in available data sets. They enable the shift from judgment-driven to data-driven analyses of causal impacts and benefits from regulations in many important cases where relevant data are available or readily obtainable, as in the examples of air pollution regulation and food safety regulation. But this shift is unlikely to take place within established regulatory cultures that emphasize the judgments of salaried experts and assumption-based modeling as tools for deciding how the world works (Wynne 1993).

By contrast, an adversarial setting in which both those who support a proposed regulation and those who oppose it subject their competing causal analyses and resulting benefits estimates to critical review based on rigorous objective standards provides incentives for production and use of relevant data and analyses. These incentives are lacking when only the regulator is charged with making a credible case for proposed regulations, and when opposing views are addressed only through responses to public comments (which, in practice, can usually be readily dismissed, e.g., by citing the contrary views and expertise of those who side with the regulator). Excessive deference to regulatory science by administrative law courts damps the incentives for others to challenge, and perhaps improve, it. Conversely, more active judicial review can stimulate challenges to improve the factual basis for estimated regulatory benefits. Courts are already positioned as the cheapest providers of review and enforcers of rigorous reasoning about the benefits claimed to be caused by

proposed regulations. Finding regulations to be arbitrary and capricious when the evidence provides no good reason to expect that they will actually cause the benefits claimed for them might create incentives to improve the quality of causal inference in regulatory science and to reduce passage of regulations whose benefits end up being less than was projected, and perhaps less than their costs.

Distinguishing Among Different Types of Causation

In discussing the health and economic benefits caused by regulations, policy makers, regulators, courts, scientists, media, and the public refer to several distinct types of causation, often without clearly distinguishing among them (Dawid 2008). These were articulated and examined at some length in Chap. 2. Here we recapitulate some key points to help make this chapter relatively self-contained; Chap. 2 offers additional explanations and refinements. Each of the following concepts of causality is widely used in discussing the causal implications of associations found in observational data. Each has its own large, specialized technical literature (see Chap. 2), but they are often conflated.

- *Associational and attributive causation.* This is the concept of causation most commonly used in epidemiology and in regulatory risk assessments and benefits estimates for health and safety regulations. It addresses how much of an observed statistical association between an exposure and an adverse outcome will be attributed to the exposure, and how much will be attributed to other factors. This is often interpreted as showing how much of the causation (or blame or liability in legal applications) for an adverse outcome is attributable to exposure, and how much to each of the other causes or factors that produced it. In epidemiology, etiological fractions, population attributable fractions, population attributable risks, burdens of disease, and probabilities of causation are all examples of attributive causal concepts (Tian and Pearl 2000). As commonly used and taught, all are derived from *relative risk*, i.e., the ratio of risks in exposed and unexposed populations, or among more- and less-exposed individuals. Hence, they are all based solely on statistical associations.
- *Predictive causation.* In statistics, economics, physics, and neuroscience, among other fields, it is common to define one variable as being a cause of another if and only if the first helps to predict the second (e.g., Friston et al. 2013; Furqan and Siyal 2016). For example, if exposure helps to predict an adverse health response, then exposure is considered a (predictive) cause of the response. As an important special case, *Granger causality* between an exposure time series and a response time series (Kleinberg and Hripcsak 2011) is based on the principle that causes help to predict their effects. (Technically, X is a Granger-cause of Y if the future of Y is dependent on—or, more formally, is not conditionally independent of—the history of X , given the history of Y .) Thus, nicotine-stained fingers can be a Granger cause of lung cancer, helping to predict it, even if cleaning one's fingers

would have no effect on future lung cancer risk (manipulative causality) (Woodward 2013).

- *Counterfactual causation* (Höfler 2005; Wang et al. 2016) attributes the difference between observed outcomes and predicted outcomes that would have occurred under different conditions, such as if exposure had not been present, to the differences between the real and alternative (“counterfactual”) conditions. This difference in conditions is said to cause the difference in outcomes, in counterfactual causal models.
- *Structural causation and exogeneity*. In constructing a simulation model of a dynamic system, the values of some variables are calculated from the values of others. As the simulation advances, input values may change, and then the values of variables that depend on them may change, and so forth, until exogenous changes in the inputs have propagated through the system, perhaps leading to new steady-state values for the output variables until a further exogenous change in inputs leads to further changes in the values of other variables. The order in which variable values are calculated and updated reflects a concept of causality in which the values of some variables are determined by the values of others that cause them. This computational view of causality considers that the values of effects (or their conditional probabilities, in stochastic models) can be determined from the values of their causes via equations or formulas, representing causal mechanisms, with exogenous changes entering from outside the modeled system propagating through the modeled mechanisms in such a way that values of causes are determined prior to the values of effects that depend on them. It has been formalized in seminal work by Simon (1953) and subsequent work by many others, mainly in economics and econometrics, artificial intelligence, and time series analysis (e.g., Iwasaki 1988; Hendry 2004; Voortman et al. 2010; Hoover 2012).
- *Manipulative causation* is the concept of causality in which changing (“manipulating”) the values of controllable inputs to a system changes the values of outputs of interest (Woodward 2013; Voortman et al. 2010; Hoover 2012). In detailed dynamic models, the changes in inputs might propagate through a system of algebraic and differential equations describing a system to determine the time courses of changes in other variables, including outputs. If such a detailed dynamic model is unavailable, the relation between changes in values of controllable inputs and changes on the values of variables that depend on them may instead be described by more abstract models such as functions (“structural equations”) relating the equilibrium values, or by Bayesian Networks specifying conditional probability distributions of outputs conditioned on values of inputs. Manipulative causation is the type of causality of greatest interest to decision-makers and policy-makers seeking to make preferred outcomes more likely by changing the values of variables that they can control.
- *Mechanistic/explanatory causation* describes how changes in the inputs to a system or situation propagate through networks of causal laws and mechanisms to produce resulting changes in other variables, including outputs.

These different concepts of causality are interrelated, but not equivalent. For example, attributional causality does not imply counterfactual, predictive, manipulative, or mechanistic causality and is not implied by them. There is no guarantee that removing a specific exposure source would have any effect on the risks that are attributed to it, nor is there any requirement than no more than 100% of a risk be attributed to the various factors that are said to cause it. For example, in the diagram $X_1 \leftarrow X_0 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow \dots \rightarrow X_n$, if exogenously changing the value of a variable at the tail of an arrow from 0 to 1 causes the value of any variable into which it points to change from 0 to 1, and if the value of X_0 is changed from 0 (interpreted as “unexposed”) to 1 (interpreted as “exposed”), then not only would these measures attribute 100% of the blame for X_1 becoming 1 to this change in X_0 , but also they would attribute the same 100% of the blame to changes in each of X_2, X_3, \dots and X_n , even though those are simply other effects of the change in X_0 . Relative risks are the same for all of these variables, and so attributional risk measures derived from relative risk assign the same blame to all (in this case, a “probability of causation” of 1).

Tort law, by contrast, uses an attributional concept of “but-for” causation that attributes harm to a cause if and only if the harm would not have occurred in its absence, i.e., “but for” the occurrence of the cause. This concept would single out the change in X_0 as the only but-for cause of the change in X_1 . On the other hand, $X_0, X_2, X_3, X_4 \dots$ would all be causes of X_n by this criterion. Thus, but-for causation can lead to promiscuous attribution of harm to remote causes in a chain or network, for example, by attributing responsibility for a smoker’s lung cancer not only to the practice of smoking, but also to the retailer who sold the cigarettes, the manufacturer of the cigarettes, the grower of the tobacco, the media that advertised the brand smoked, the friends or family or culture that encouraged smoking, the schools that failed to intervene, genes that predisposed the smoker to addiction, and so on. All can be considered but-for causes of smoking. Tort law also provides standards such as more-likely-than not and joint and several liability for cases where causation is uncertain or is distributed among multiple causes.

Predictive causality does not necessarily imply manipulative causality unless other conditions hold, such as that no omitted confounders are present. This is illustrated by the standard counter-example, mentioned above and in Chap. 2, of nicotine-stained fingers being a predictive but not a manipulative cause of lung cancer, where smoking is the omitted confounder. Often in public health and safety regulations, it is not known whether these other conditions hold, and hence it is not clear whether predictive causality implies manipulative causality. On the other hand, predictive causation can very often be established or refuted (at a stated level of statistical confidence) based on data by applying statistical tests to determine whether predictions of outcomes are significantly improved by conditioning on information about their hypothesized causes. Such tests examine what *did* happen to the effects when the hypothesized causes had different values, rather than requiring speculations about what *would* happen to effects under different conditions, as in counterfactual causal modeling. Therefore, even though predictive causality does not necessarily imply manipulative causality, it provides a useful data-driven screen

for potential manipulative causation, insofar as manipulative causation usually implies predictive causation (since changes in inputs help to predict the changes in outputs that they cause).

For counterfactual causation, what the outcomes would have been under different, counterfactual conditions is never observed. Therefore, the estimated difference in outcomes caused by differences between real and counterfactual conditions must be calculated using predictive models or assumptions about what would have happened. These predictions may not be accurate. In practice, they are usually simply assumed, but are difficult or impossible to validate. Counterfactual models of causation are also inherently ambiguous, in that the outcome that would have occurred had exposure been absent usually depends on *why* exposure would have been absent, which is seldom specified. For example, nicotine-stained fingers would be a counterfactual cause of lung cancer if clean fingers imply no smoking, but not if they arise only because smokers wear gloves when smoking. In the case of air pollution, especially if exposure and income interact in affecting mortality rates, assuming that the counterfactual condition without exposure occurs because everyone becomes wealthy enough to move to unpolluted areas might yield quite different estimates of counterfactual mortality rates than assuming that lack of exposure was caused by the onset of such abject poverty and economic depression that pollution sources no longer operate. Counterfactual models usually finesse any careful exposition of specific assumptions about why counterfactual exposures would occur by using statistical models to predict what would have happened if exposures had been different. But these models are silent about why exposures would have been different, and hence the validity of their predictions is unknown. (Economists have noted a similar limitation of macroeconomic models derived from historical data to predict the effects caused by future interventions that change the underlying data-generating process. This is the Lucas critique of causal predictions in macroeconomics policy models mentioned in Chap. 1.)

Although manipulative causality usually implies predictive causality, neither one necessarily implies attributive causality. For example, if consuming aspirin every day reduces risk of heart attack in an elderly population, but only people with high risks take daily aspirin, then there might be both a positive association (and hence a positive etiologic fraction, probability of causation, and population attributable risk) between aspirin consumption and heart attack risk in the population, but a negative manipulative causal relationship between them, with aspirin consumption reducing risk. Even if aspirin had no effect on risk, it could still be positively associated with risk if people at high risk were more likely to consume it. Thus, manipulative and associational-attributive causation do not necessarily have any implications for each other.

The following example illustrates some of these important distinctions among causal concepts more quantitatively.

Example: Associations Do Not Necessarily Provide Valid Manipulative Causal Predictions

Suppose that in a certain city, daily mortality rate, R and average daily exposure concentration of an air pollutant, C , over an observation period of several years are perfectly described by the following Model 14.1:

$$R = C + 50 \quad (\text{Model 14.1})$$

That is, each day, the number of deaths is equal to 50 deaths plus the average daily concentration of the air pollutant. What valid inferences, if any, do these observations enable about how changing C would change R ? The answer, as stressed in Chaps. 1 and 2, is none: historical associations do not logically imply anything about predictive, counterfactual, structural, or manipulative causation. One reason is that Model 14.1 implies that the same data are also described perfectly by the following Model 14.2, where T is an unmeasured third variable (such as temperature) with values between 0 and 100:

$$\begin{aligned} C &= 50 - 0.5T \\ R &= 150 - C - T \end{aligned} \quad (\text{Model 14.2})$$

(The first equation implies that $T = 100 - 2C$, and substituting this into the second equation to eliminate T yields Model 14.1.) If the equations in Model 14.2 are structural equations with the explicit causal interpretation that exogenously changing the value of a variable on the right side of an equation will cause the value of the dependent variable on its left side to change to restore equality, then the second equation reveals that each unit of reduction in C would increase R by one unit. In this case, if Model 14.1 is only a reduced-form model describing historical associations, then mis-interpreting it as a causal model would mistakenly imply that increasing C would increase R . The associational Model 14.1 is not incorrect as a description of past data. It would be valid for predicting how many deaths would occur on days with different exposure concentrations in the absence of interventions. But only the causal Model 14.2 can predict how changing C would change R , and there is no way to deduce Model 14.2 by scrutiny of Model 14.1.

This review of different concepts of causation has highlighted the following two key conclusions: (a) As emphasized in Chap. 2, policy-makers, regulators, courts, and the general public are (or should be) primarily interested in *manipulative* causation, i.e., in how regulations or other actions that they take affect probabilities of outcomes, and hence the benefits caused by their actions; but (b) Regulatory science and claims about the causal impacts of regulations usually address only associational-attributive causation, and occasionally about other non-manipulative (especially, counterfactual) causal concepts. Judicial review of the causal reasoning and evidence used to support estimates of regulatory benefits can and should close this gap between causal concepts by insisting that arguments and evidence presented must address manipulative causation, and that other forms of causation must not be

conflated with it. There is an urgent need to enforce such clarity, as current practices in epidemiology, public health, and regulatory science routinely confuse associational-attributive causation with manipulative causation.

As documented in Table 2.5, many published articles in peer-reviewed scientific journals move freely between associational and manipulative causal interpretations of exposure-response associations without showing that the presented associations do in fact describe (manipulative) causation. As a consequence, regulatory benefits assessments and calls for further regulation based on these and many similar analyses do not reveal what consequences, if any, further regulations should actually be expected to cause. In this sense, they might be regarded as arbitrary and capricious, as they provide no rational basis for identifying the likely consequences of the recommended regulations.

Can Regulatory Benefits Estimation Be Improved, and, If So, How?

Can more active judicial review truly improve the accuracy of causal inferences and benefits predictions used in deciding which proposed regulatory changes to make and in evaluating their performance? To what extent are improvements constrained by hard limits on what can be reliably predicted and learned from realistically incomplete and imperfect data? The following distinct lines of evidence from very different areas suggest that substantial improvements are indeed possible in practice, but that they are best accomplished with the help of strong external critical review of the evidence and reasoning relied on by regulatory agencies and advocates.

The first line of evidence comes from sociological and organizational design studies (see Chap. 13). These suggest that the organizational culture and incentives of regulatory agencies usually put weight on authoritative, prospective estimates of benefits, with supporting causal assumptions that reflect the entrenched views of the organization that regulation produces desirable results and that the beliefs of the regulators are scientific and trustworthy (Wynne 1993). However, organizational cultures that foster demonstrably high performance in managing risks and uncertainties function quite differently. They typically acknowledge ignorance and uncertainty about how well current policies and actions are working. They focus on learning quickly and effectively from experience, frequently revisiting past decisions and assumptions and actively questioning and correcting current policies entrenched assumptions and beliefs as new data are collected (Dekker and Woods 2009; Weick and Sutcliffe 2001; see Chap. 13). For example, difficult and complex operations under uncertainty, such as managing air traffic coming and going from nuclear aircraft carriers, operating nuclear power plants or offshore oil platforms safely for long periods under constantly changing conditions, fighting wildfires, landing airplanes successfully under unexpected conditions, or performing complex surgery, are carried out successfully in hundreds of locations worldwide every day.

As discussed in Chap. 13, the disciplines and habits of mind practiced and taught in such high reliability organizations (HROs) have proved useful in helping individuals and organizations plan, act, and adjust more effectively under uncertainty. Regulatory agencies dealing with uncertain health and safety risks can profit from these lessons (Dekker and Woods 2009).

Five commonly listed characteristics of HROs are as follows (see Chap. 13): sensitivity to operations—to what is working and what is not, with a steady focus on empirical data and without making assumptions (Gamble 2013); reluctance to oversimplify explanations for problems, specifically including resisting simplistic interpretations of data and assumptions about causality; preoccupation with failure, meaning constantly focusing on how current plans, assumptions, and practices might fail, rather than on building a case for why they might succeed; deference to expertise rather than to seniority or authority; and commitment to resilience, including willingness to quickly identify and acknowledge when current efforts are not working as expected and to improvise as needed to improve results (Weick and Sutcliffe 2001). Of course, regulatory processes that unfold over years, largely in the public sphere, are a very different setting from operations performed by specially trained teams. But it is plausible that many of the same lessons apply to regulatory organizations seeking to improve outcomes in a changing and uncertain environment (Dekker and Woods 2009).

A second line of evidence that the improvements in predicting the effects of regulations can be achieved in practice comes from research on improving judgment and prediction, recently summarized in the popular book *Superforecasting* (Tetlock and Gardner 2015). Although most predictions are overconfident and inaccurate, a small minority of individuals display consistent, exceptional performance in forecasting the probabilities of a wide variety of events, from wars to election outcomes to financial upheavals to scientific discoveries. These “superforecasters” apply teachable and learnable skills and habits that explain their high performance. They remain open-minded, always regarding their current beliefs as hypotheses to be tested and improved by new information. They are eager to update their current judgments frequently and precisely, actively seeking and conditioning on new data and widely disparate sources of data and evidence that might disprove or correct their current estimates. They make fine-grained distinctions in their probability judgments, often adjusting by only one or a few percentage points in light of new evidence, which is a level of precision that most people cannot bring to their probability judgments. The authors offer the following rough recipe for improving probability forecasts: (1) “Unpack” the question to which the forecast provides an answer, e.g., about the health benefits that a regulation will end up causing, into its components, such as who will receive what kinds of health benefits and under what conditions. (2) Distinguish between what is known and unknown and scrutinize all assumptions. For example, do not assume that reducing exposure will cause proportional reductions in adverse health effects unless manipulative causation has actually been shown. (3) Consider other, similar cases and the statistics of their outcomes (taking what the authors call “the outside view”) and then (4) Consider what is special or unique about this specific case in contradistinction to others (the “inside

view”). (5) Exploit what can be learned from the views of others, especially those with contrasting informed predictions, as well as from prediction markets and the wisdom of crowds. (6) Synthesize all of these different views into one (the multi-faceted “dragonfly view,” in the authors’ term) and (7) Express a final judgment, conditioned on all this information, as precisely as possible using a fine-grained scale of probabilities. Skill in making better predictions using this guidance can be built through informed practice and clear, prompt feedback, provided that there is a deliberate focus on tracking results and learning from mistakes.

A third line of evidence showing that it is possible to learn to intervene effectively even in uncertain and changing environments to make preferred outcomes more likely and frequent comes from machine learning, as reviewed in Chap. 1: the design and performance of reinforcement-learning algorithms that automatically learn decision rules from experience and improve them over time. A very successful class of algorithms called “actor-critic” methods (Konda and Tsitsiklis 2003; Lei 2016; Ghavamzadeh et al. 2016) consist of a policy or “actor” for deciding what actions to take next, given currently available information; and one or more reviewers or “critics” that evaluate the empirical performance of the current policy and suggest changes based on the difference between predicted and observed outcomes. These algorithms have proved successful in learning optimal (net-benefit maximizing) or near-optimal policies quickly in a variety of settings with probabilistic relations between actions and their consequences and with systems that behave in uncertain ways, so that is necessary to adaptively learn how best to achieve desired results.

High-reliability organizations, superforecasters, and successful machine learning algorithms for controlling uncertain systems all apply the following common principles.

- (a) Recognize that even the best current beliefs and models for predicting outcomes and for deciding what to do to maximize net benefits will often be mistaken or obsolete. They should be constantly checked, improved, and updated based on empirical data and on gaps between predicted and observed results.
- (b) Relying on any single model or set of assumptions for forecasting and decision-making is less effective than considering the implications of many plausible alternatives.
- (c) Seek and use potential disconfirming data and evidence from many diverse sources to improve current beliefs, predictions, and control policies.
- (d) Use informed external critics to improve performance by vigilant review, frequent challenging of current assumptions, predictions and policies, and informed suggestions for changes based on data.

Applying these principles to regulatory agencies suggests that a mindset that seeks to identify and defend a single “best” model, set of assumptions, or consensus judgment about the effects caused by proposed regulations will be less likely to maximize uncertain net social benefits than treating effects as uncertain quantities to be learned and improved via experience and active learning from data. A *judgment-driven culture* in which selected experts form and defend judgments about causation and estimated regulatory benefits is less beneficial than a *data-driven culture* in

which the actual effects of regulations are regarded as uncertain, possibly changing quantities to be learned about and improved by intelligent trial and error and learning from data. A data-driven regulatory culture expects to benefit from independent external challenges and reviews of reasoning and assumptions before regulatory changes are approved and from frequent updates of effects estimates based on data collected after they are implemented. Strong judicial review can provide the first part, external reviews of reasoning, by maintaining a high standard for causal reasoning based on data and manipulative causation.

Working against the establishment of a data-driven culture is a long tradition in medicine, public health, and regulatory science of treating causation as a matter of informed judgment that can only be rendered by properly prepared experts, rather than as a matter of empirically discoverable and independently verifiable fact that can be determined from data. The difficulties and skepticism that have faced proponents of evidence-based medicine and evidence-based policies, emanating from traditions that emphasize the special authority of trained experts (Tetlock and Gardner 2015), suggest the barriers that must be overcome to shift more toward data-driven regulatory cultures.

The following sections discuss the contrasting technical methods used by proponents of the causation-as-judgment and causation-as-fact views, and then suggest that a modern synthesis of these methods provides practical principles for defining and using informative evidence of manipulative causation in administrative law to achieve better results from regulations.

Causation as Judgment: The Hill Considerations for Causality and Some Alternatives

As discussed in more detail in Chap. 2, one of the most influential frameworks for guiding consideration and judgments about causality is that of Sir Austin Bradford Hill, who in 1965 proposed nine aspects of an exposure-response association that he recommended “we especially consider before deciding that the most likely interpretation of it is causation” (Hill 1965, quoted in Lucas and McMichael 2005). This original formulation reflects a view in which causation is dichotomous: an association is either causal or not. Modern statistics and machine learning approaches to causal inference take a more nuanced view in which the total association between two quantities can be explained by a mix of factors and pathways, including some causal impacts and some confounding, sample selection and model selection biases, coincident historical trends, omitted variables, omitted errors in explanatory variables, model specification errors, overfitting bias, p-hacking, and so forth.

The expressed goal of the Hill considerations is to help someone make a qualitative judgment, “deciding that the most likely interpretation of [an exposure-response association] is causation,” rather than to quantify how likely this interpretation is, and what the probability is that the association is not causal after all, even if

causation is decided to be the most likely interpretation. Thus, Hill's considerations were never intended to provide the quantitative information that is essential for BCA evaluations of uncertain regulatory benefits. Consistent with the culture of many medical and public health organizations over a long history (Tetlock and Gardner 2015), they instead portray causality as a matter for informed subjective qualitative judgment by expert beholders, not as a fact to be inferred (or challenged) by rigorous, objective, and independently reproducible analysis of data.

The Hill considerations themselves—briefly referred to as strength, consistency, specificity, temporality, biological gradient, plausible mechanism, coherence, experimental support (if possible), and analogy for exposure-response associations—are discussed in more detail later in the context of showing how they can be updated and improved using ideas from current data science. Chapter 2 provides a much more thorough discussion. Hill himself acknowledged that his considerations are neither necessary nor sufficient for establishing causation, but suggested that admittedly fallible subjective judgments based on these considerations may be the best that we can hope for. This line of thinking continues to dominate many regulatory approaches to causal inference. For example, the US EPA has formulated and adopted modified versions of the Hill considerations as principles for making weight-of-evidence determinations about causation for ecological, carcinogen, and other risks. Neither the original Hill considerations nor more recent weight-of-evidence frameworks based on them distinguish between associational-attributive, predictive, manipulative, and other types of causation. Thus, the enormous influence of these considerations has tended to promote judgment-based cultures for making and defending causal assertions while conflating different concepts of causation, without providing a sharp focus on objective evidence and quantification of the manipulative causal relationships needed for rational choice among alternatives based on BCA calculations.

Of course, methodologists have not been blind to the difficulties with associational and attributive methods. The fact that the sizes and signs of associations are often model-dependent and that different investigators can often reach opposite conclusions starting from the same data by making different modeling choices has long been noted by critics of regulatory risk assessments, finally leading some commentators to conclude that associational methods are unreliable in general (Dominici et al. 2014). Recognizing such criticisms, there has been intense effort over the past decade to develop and apply more formal methods of causal analysis within the judgment-oriented tradition. This has produced a growing literature that replaces the relatively crude assumption that appropriately qualified and selected experts can directly judge associations to be causal with more sophisticated technical assumptions that imply that associations are causal without directly assuming it. Human judgment still plays a crucial role, insofar as the key assumptions are usually unverifiable based on data, and are left to expert judgments to accept. The most important of these assumption-driven causal inference frameworks, and their underlying assumptions, are as follows (Cox 2017).

- *Intervention studies* assume that if health risks change *following* an intervention, then the change is (probably) *caused by* the intervention. This assumption is often mistaken, as in the Irish coal burning ban studies (Dockery et al. 2013): both exposures and responses may be lower after an intervention than before it simply because both are declining over time, even if neither causes the other. Construing such coincidental historical trends as evidence of causation is a form of the *post hoc ergo propter hoc* logical fallacy.
- *Instrumental variable* (IV) studies assume that a variable (called an “instrument”) is unaffected by unmeasured confounders and that it directly affects exposure but not response (Schwartz et al. 2015). The validity of these assumptions is usually impossible to prove, and the results of the IV modeling can be greatly altered by how the modeler chooses to treat lagged values of variables (O’Malley 2012).
- *Counterfactual “difference-in-differences”* and *potential outcomes* models assume that differences between observed responses to observed exposure concentrations and unobserved model-predicted responses to different hypothetical “counterfactual” exposure concentrations are caused by the differences between the observed and counterfactual exposures (e.g., Wang et al. 2016). However, they might instead be caused by errors in the model or by systematic differences in other factors such as distributions of income, location, and age between the more- and less-exposed individuals. The assumption that these are not the explanations is usually untested, but is left as a matter for expert judgment to decide.
- *Regression discontinuity* (RD) studies assume that individuals receiving different exposures or treatments based on whether they are above or below a threshold in some variable (e.g., age, income, location, etc.) triggering a government intervention are otherwise exchangeable, so that differences in outcomes for populations of individuals above and below the threshold can be assumed to be caused by differences in the intervention or treatment received. The validity of this assumption is often unknown. In addition, as noted by Gelman and Zelizer (2015), RD models “can overfit, leading to causal inferences that are substantively implausible...” For an application to air pollution health effects estimation based on differences in coal burning in China, they conclude that a “claim [of a health impact], and its statistical significance, is highly dependent on a model choice that may have a data-analytic purpose, but which has no particular scientific basis.”
- As discussed in Chap. 2, *associational, attributable-risk, and burden-of-disease studies* assume that if responses are *greater* among people with higher exposures, then this difference is *caused* by the difference in exposures, and could be removed by removing it (manipulative causation). Typically, this assumption is made without careful justification. It simply assumes that association reflects causation. Conditions such as the Hill considerations of strong and consistent association are commonly misconstrued as evidence for manipulative causation in such studies (e.g., Fedak et al. 2015; Höfler 2005), without testing potential disconfirming alternative hypotheses such as that strong and consistent modeling assumptions, biases, confounders, effects of omitted variables, effects of omitted

error terms for estimated values of predictors, model specification errors, model uncertainties, coincident historical trends, and regression to the mean, might account for them (Greenland 2005).

These methods all make assumptions that, if true, could justify treating associations as if they indicated manipulative causation. Whether they are true, however, is usually not tested based on data, but is left to expert judgment to decide. As succinctly noted by Gelman and Zelizer (2015) in presenting their own critique of regression discontinuity [RD] studies, “One way to see the appeal of RD is to consider the threats to validity that arise with five other methods used for causal inference in observational studies: simple regression, matching, selection modeling, difference in differences, and instrumental variables. These competitors to RD all have serious limitations: regression with many predictors becomes model dependent...; matching, like linear or nonlinear regression adjustment, leans on the assumption that treatment assignment is ignorable conditional on the variables used to match; selection modeling is sensitive to untestable distributional assumptions; difference in differences requires an additive model that is not generally plausible; and instrumental variables, of course, only work when there happens to be a good instrument related to the causal question of interest.” Something better than unverified assumption-driven methods is needed.

Causation as Discoverable Empirical Fact: Causal Inference Algorithms and Competitions

At the opposite pole from Hill’s perspective that determination of causation cannot be reduced to a recipe or algorithm is a rich body of literature and computational approaches to causal inference, introduced in Chap. 2, that seek to do exactly that by providing algorithms for automatically drawing reliable causal inferences from observational data (e.g., Aliferis et al. 2010; Kleinberg and Hripcsak 2011; Hoover 2012; Rottman and Hastie 2014; Bontempi and Flauder 2015). The best-known modern exponent of causal inference algorithms may be the computer scientist Judea Pearl (e.g., Pearl 2009, 2010), although, as discussed in Chap. 2, this analytic tradition extends back to work by economists and social statisticians since the 1950s (e.g., Simon 1953) and by biologists, geneticists, and psychologists since the invention of path analysis by Sewell Wright a century ago (Joffe et al. 2012). Most causal inference algorithms use statistical tests to determine which variables help to predict effects of interest, even after conditioning on the values of other variables (Pearl 2010). Thus, they mainly detect predictive causation, although some also explicitly address implications for causal mechanisms, structural causation, and manipulative causation (Iwasaki 1988; Voortman et al. 2010). Their emphasis on predictive causation allows causal inference algorithms to benefit from well-developed principles and methods for predictive analytics and machine learning (ML).

Key technical ideas of causal inference algorithms can be used more generally to guide human reasoning about causal inference. Here, we very briefly summarize some of the key ideas explained much more thoroughly in Chap. 2. An idea used in many causal inference algorithms is that in a chain such as $X \rightarrow Y \rightarrow Z$, where arrows denote manipulative or predictive causation (so that changes in the variable at the tail of an arrow change or help to predict changes in the variable that it points into, respectively), each variable should have a statistical dependency on any variable that points into it, but Z should be conditionally independent of X given the value of Y , since Z depends on X only through the effect of X on Y . Algorithms that test for conditional independence and that quantify conditional probability dependencies among variables are now mature (Frey et al. 2003; Aliferis et al. 2010) and are readily available to interested practitioners via free Python and R packages for ML, such as the *bnlearn* package in R, which learns probabilistic dependencies and independence relations (represented via Bayesian network (BN) structures and conditional probability tables) from data. A second, related idea is that in the chain $X \rightarrow Y \rightarrow Z$, Y should provide at least as much information as X for predicting Z . A third idea, introduced in path analysis for linear relationships among variables and generalized in BNs to arbitrary probabilistic dependencies, is that the effect of changes in X on changes in Z should be a composition of the effect of changes in X on Y and the effect of changes in Y on Z . Such ideas provide constraints and scoring criteria for identifying causal models that are consistent with data.

Modern causal inference algorithms offer dozens of constructive alternatives for assessing predictive causal relations in observational data without relying on human judgment or unverified modeling assumptions. The field is mature enough so that, for over a decade, different causal inference algorithms have been applied to suites of challenge problems for which the underlying data-generating processes are known to see how accurately the algorithms can recover correct descriptions of the underlying causal models from observed data. Competitions are now held fairly regularly that quantify and compare the empirical performance of submitted causal inference algorithms on suites of test problems (e.g., NIPS 2013 Workshop on Causality; Hill 2016). Results of recent causal inference competitions suggest the following principles for causal inference from observational data as common components of many of the top-performing algorithms.

- *Information principle:* Causes provide information that helps to predict their effects and that cannot be obtained from other variables. This principle creates a bridge between well-developed computational statistical and ML methods for identifying informative variables to improve prediction of dependent variables, such as health effects, and the needs of causal inference (Pearl 2009). To the extent that effects cannot be conditionally independent of their direct manipulative causes, such information-based algorithms provide a useful screen for potential manipulative causation, as well as for predictive causation.
- *Propagation of changes principle:* Changes in causes help to explain and predict changes in effects (Friston et al. 2013; Wu et al. 2011). This applies the information principle to changes in variables over time. It can often be visualized in

terms of changes propagating along links (representing statistical dependencies) in a BN or other network model.

- *Nonparametric analyses principle.* Multivariate non-parametric methods, most commonly, classification and regression trees (CART) algorithms, can be used to identify and quantify information dependencies among variables without having to make any parametric modeling assumptions (e.g., Halliday et al. 2016). CART trees can also be used to test for conditional independence, with the dependent variable being conditionally independent of variables not in the tree, given the variables that are in it, at least as far as the tree-growing algorithm can discover (Frey et al. 2003; Aliferis et al. 2010).
- *Multiple models principle.* Rather than relying on any single statistical model, the top-performing causal analytics algorithms typically fit hundreds of nonparametric models (e.g., CART trees), called *model ensembles*, to randomly generated subsets of the data (Furqan and Siyal 2016). Averaging the resulting predictions of how the dependent variable depends on other variables over an ensemble of models usually yields better estimates with lower bias and error variance than any single predictive model. This is reminiscent of the principle in high-reliability organizations and among superforecasters of considering many theories, models, and points of view, rather than committing to a single best one. Computational statistics packages such as the *randomForest* package in R automate construction, validation, and predictive analytics for such model ensembles and present results in simple graphical forms, especially partial dependence plots (Chap. 2) that show how a dependent variable is predicted to change as a single predictor is systematically varied while leaving all other variables with their empirical joint distribution of values. If this dependency represents manipulative causality, then the partial dependence plot indicates how the conditional expected value of an outputs such as mortality in a population is expected to change when a variable such as exposure is manipulated, given the empirical joint distribution of other measured predictors on which the output also depends. Otherwise, it quantifies a predictive relation.

High-performance causal inference algorithms for observational data usually combine several of these principles. Interestingly, none of them uses the Hill considerations or associational-attributional methods such as probability of causation or attributable risk formulas from epidemiology. A counterfactual-potential outcomes causal modeling approach was entered in a recent competition (Hill 2016), but performed relatively poorly, with roughly 20 times larger bias, 20 times larger mean square prediction error for estimated causal effects, and wider uncertainty intervals than tree-based algorithms incorporating the above principles. This presumably reflects the fact that the counterfactual approach depends on models of unknown validity. In short, causal inference and discovery algorithms that assume that causal relationships are empirical facts that can be discovered from data have made great progress and yield encouraging performance in competitive evaluations (Bontempi and Flauder 2015), but none of them uses the methods usually relied on by regulatory agencies in making judgments about causation. Such methods,

including weight-of-evidence schemes for evaluating and combining causal evidence, were tried and evaluated as approaches to automated assessment of causality in expert systems research in the 1980s (e.g., Spiegelhalter 1986; Todd 1992), but they have been out-competed by modern causal inference algorithms incorporating the above principles, and are no longer used in practical applications. That they continue to play a dominant role in causal inference in many regulatory agencies invites the question of whether these agencies could also dramatically improve their performance in predicting and assessing causal effects of regulations by applying modern causal inference algorithms and principles instead.

Synthesis: Modernizing the Hill Considerations

The enduring influence and perceived value of the Hill considerations and of judgment-centric methods for causal inference in regulatory agencies shows that they fill an important need. Despite Hill's disclaimers, this is largely the need to have a simple, intuitively plausible checklist to use in assessing evidence that reducing exposures will reduce risks of harm. At the same time, the successes of data-centric, algorithmic methods of causal inference and causal discovery in competitive evaluations suggests the desirability of a synthesis that combines the best elements of each. This section describes each of the Hill considerations, its strengths and limitations, and possibilities for improving on Hill's original 1965 formulation using contemporary ideas. The same list of considerations is also discussed in Chap. 2, using the technical concepts and terminology of modern causal analytics introduced there. Here, we reconsider them without assuming that technical background.

- *Strength of association:* Hill proposed as the first consideration that larger associations are more likely to be causal than smaller ones. One possible underlying intuition to support this is that causal laws always hold, so they should produce large associations, but conditions that generate spurious associations only hold sometimes, e.g., when confounders are present, and thus they tend to generate smaller associations. Whatever the rationale, objections to this consideration are that
 - The existence, direction, and size of an association is often model-dependent (Dominici et al. 2014; Gelman and Zelizer 2015). Recall the example of Models 14.1 and 14.2 with $R = C + 50$ and $R = 150 - C - T$, respectively. In Model 14.1, C is positively associated with R while in Model 14.2, C is negatively associated with R . More generally, whether an association is large or small may reflect modeling choices rather than some invariant fact about the real world that does not depend on the modeler's choices.
 - Associations are not measures of manipulative causation.
 - There is no reason in general to expect that a larger association is more likely to be causal than to expect that it indicates stronger confounding, larger

modeling errors or biases, stronger coincident historical trends, or other non-causal explanations.

On the other hand, there is a useful insight here that can be formulated more precisely and correctly in more modern terminology. In a causal network such as the chain $W \leftarrow X \rightarrow Y \rightarrow Z$, where arrows signify predictive or manipulative causation or both, it must be the case that Y provides at least as much information about Z as X or W does, and typically more (Cover and Thomas 2006). (Technically, the information that one random variable provides about another, measured in bits, is quantified as the expected reduction in the entropy of the probability distribution of one variable achieved by conditioning on the value of the other.) Thus, if Y is a direct manipulative or predictive cause of a dependent variable Z , it will provide as much or more information about Z than indirect causes such as X or non-cause variables such as W that are further removed from it in the causal network. The same is not necessarily true for correlations: if $Y = X^2$ and $Z = Y^{1/2}$, then X and Z will be more strongly correlated than Y and Z , even though Z depends directly on Y and not on X . Thus, replacing *association* in Hill's formulation with *information* yields a useful updated principle: the direct cause(s) of an effect provide more information about it than indirect causes and variables to which it is not causally related. Thus, *variables that provide more information about an effect are more likely to be direct causes or consequences of it than are variables that provide less information*. Modern causal discovery algorithms incorporate this insight via the information principle that effects are not conditionally independent of their direct causes and via CART tree-growing algorithms that identify combinations of predictor values that are highly informative about the value of an effect dependent variable.

- *Consistency:* Hill proposed that if different investigators arrive at consistent estimates of an exposure-response association in different populations, then this reproducibility provides evidence that the consistently found association is causal. Against this, as noted by Gelman and Zelizer (2015), is the recognition that, “once researchers know what to expect, they can continue finding it, given all the degrees of freedom available in data processing and analysis.” Modern ensemble modeling methods for predictive analytics pursue a somewhat similar criterion—but avoid the potential bias of knowing what to expect and using p-hacking to find it—by partitioning the data into multiple randomly selected subsets (“folds”), fitting multiple predictive models (e.g., CART trees, see Chap. 2) to each subset, and then evaluating their out-of-sample performance on the other subsets. Averaging the predictions from the best-performing models then yields a final prediction, and the distribution of the top predictions characterizes uncertainty around the final prediction. Such computationally intensive methods of predictive analytics provide quantitative estimates of predictive causal relations and uncertainty about them. The Hill consideration that consistent associations are more likely to be causal is replaced by a principle that consistency of estimates across multiple models and subsets of available data implies less uncertainty about predictive relationships. In addition, conditions and

algorithms have been developed for “transporting” causal relations among variables inferred from interventions and observations in one population and setting to a different population and setting for which observational data are available (Bareinboim and Pearl 2013; Lee and Honavar 2013). These transportability algorithms have been implemented in free R packages such as *causaleffect* (Tikka 2018). They capture the idea that causal relationships can be applied in different situations, but that differences between situations may modify the effects created by a specified cause in predictable ways. This is a powerful generalization of the consistency consideration envisioned by Hill.

- *Specificity*: Hill considered that the more specific an association is between an exposure and an effect is, the more likely it is to be causal. This consideration is seldom used now because it is recognized that most exposures of interest, such as fine particulate matter, might have more than one effect and each effect, such as lung cancer, might have multiple causes. Instead, modern causal inference algorithms such as those in the R package *bnlearn* discover causal networks that allow multiple causes and effects to be modeled simultaneously.
- *Temporality*: Hill considered that causes must precede their effects. This was the only one of his nine considerations that he held to be a necessary condition. Modern causal inference algorithms agree, but refine the criterion by adding that causes must not only precede their effects, but must also help to predict them. Methods such as Granger causality testing specify that the history (past and present values) of a cause variable must help to predict the future of the effect variable better than the history of the effect variable alone can do.
- *Biological gradient*: This consideration states that if larger exposures are associated with larger effects, then their association is more likely to be causal than if such monotonicity does not hold. This is closely related to the strength-of-association criterion, since many measures of association (such as correlation) assume a monotonic relationship. Just as a strong confounder can explain a strong exposure-response association in the absence of manipulative causation, so it can explain a monotonic relation between exposure and response even in the absence of manipulative causation. Since 1965, research on nonlinear and threshold exposure-response relations has made clear that many important biological processes and mechanisms do not satisfy the biological gradient criterion. Modern methods of causal discovery, including CART trees and Bayesian Networks (Chap. 2), can discover and quantify non-monotonic relationships between causes and their effects, so the biological gradient criterion is unnecessary for applying these methods.
- *Plausibility*: Hill considered that providing a plausible mechanism by which changes in exposure might change health effects would make a causal relationship between them more likely, while acknowledging that ignorance of mechanisms did not undermine epidemiological findings of associations. The converse is that ignorance of mechanisms can make many proposed mechanisms seem superficially plausible. Fortunately, modern bioinformatics methods allow principles of causal network modeling to be applied to elucidate causal mechanisms and paths, as well as to describe multivariate dependencies among population

level variables. Thus, proposed causal mechanisms and paths linking exposure to harm can now be tested using the principles already discussed and data on the relevant variables in bioinformatics databases. For example, a mechanistic path such as “Exposure X increases biological activity Y , which then increases risk of adverse effect Z ” might sound plausible when proposed, but might then be shown to be not plausible after all if changes in Z turn out to be independent of changes in Y , or if changes in Z are still dependent on changes in X even when the value of Y has been conditioned on. The same causal discovery and inference algorithms can be applied to both epidemiological and biological data. No new principles or algorithms are required to develop causal network models and dynamic causal simulation models from data collected at the levels of populations, individuals, organ systems, tissues and cell populations, or intracellular processes, as witnessed by the explosive growth of causal discovery and inference algorithms and network modeling in systems biology.

- *Coherence:* Similar to plausibility, coherence of a manipulative causal exposure-response with current scientific understanding, which Hill considered to increase the likelihood that a causal relationship exists, is can be addressed by modern causal diagram methods (Joffe et al. 2012) without introducing any new principles or algorithms. Causal network inference and modeling algorithms can be applied to variables at different levels in the biological hierarchy, allowing coherence among causal networks at different levels to be determined from data. Coherence of knowledge at different levels is then an output from these algorithms, rather than an input to them. Alternatively, if knowledge is sufficient to allow some arrows in a causal diagram to be specified or forbidden, then these knowledge-based constraints can be imposed on the network-learning algorithms in programs such as *bnlearn*, assuring the coherence of discovered networks with these constraints.
- *Experiment:* If interventions are possible for a subset of controllable variables, then setting them to different values and studying how other variables respond can quickly elucidate manipulative causality (Voortman et al. 2010). Causal network discovery algorithms add to this consideration specific techniques for designing experimental manipulations to reveal manipulative causal relationships and algorithms for “transporting” the resulting causal knowledge to new settings with different values of some of the variables (Tikka 2018).
- *Analogy.* The last of Hill’s considerations is that it is more likely that an association is causal if its exposure and response variables are similar to those in a known causal relationship. A difficulty with this is that what constitutes relevant “similarity” may not be known. For example, are two mineral oils “similar” for purposes of predicting causation of dermal carcinogenicity if they have similar viscosities, or similar densities, or similar polycyclic aromatic hydrocarbon (PAH) content, or some other similarities? The theory of transportability of causal relationships across different settings (Bareinboim and Pearl 2013; Lee and Honavar 2013; Tikka 2018) provides a more precise and rigorous understanding of what conditions must be satisfied for a causal relationship identified and quantified in one system to hold in another. The variables (e.g.,

viscosity, density, PAH content, etc.) that are relevant for letting a causal relationship be transported define the relevant similarities between systems, and thus allow the analogy consideration to be made precise.

This comparison of Hill considerations with principles used in current causal network learning algorithms suggests that real progress has been made since 1965. The considerations of strength, consistency, and temporality can be refined and made more precise using modern concepts and terminology. The considerations of specificity, plausibility, and biological gradients incorporate restrictions that are no longer needed to draw sound and useful causal inferences, since current causal inference algorithms can simultaneously handle multiple causes and effects, multiple causal pathways, and nonlinear and non-monotonic relationships. The somewhat vague considerations of coherence and analogy can be made more precise, and experimental and observational data can be combined for purposes of causal inference, using the recent theory of transportability of causal relationships (Bareinboim and Pearl 2013; Tikka 2018). These technical advances suggest that it is now practical to use data-driven causal inference methods and concepts to clarify, refine, and replace earlier judgment-based approaches to causal inference. They provide concrete criteria that can be implemented in software algorithms or applied by courts to make more objective and accurate determinations of manipulative causality than has previously been possible. This provides a technical basis for expanding the role of judicial review to include encouraging and enforcing improved causal inference.

Summary and Conclusions: Potential Roles of Judicial Review in Transforming Regulatory Causal Inference and Prediction

This chapter has argued that more active and stringent judicial review of the causal reasoning and claims advanced in support of regulations can increase the net social benefits from regulations by correcting problems that currently promote unrealistically large estimates of the benefits caused by regulations. Among these are the following:

1. *Ignoring risk-aversion and risk premiums for correlated losses.* When it is not certain that reducing exposure to a regulated substance or activity will actually cause the expected health, economic, or other benefits attributed to such reductions, and when the regulation affects a large number of economic agents, then the risk-adjusted value of the uncertain benefits can be much less than their expected value. This difference, called the risk premium in decision analysis, is due to risk-aversion, which penalizes large numbers of correlated losses. This reduction in benefits due to uncertainty about causation is not accounted for in benefits assessments and BCA calculations that focus on expected net benefits while ignoring risk aversion.

2. *Tyranny of extreme perceptions.* Regulatory agencies may attract and retain employees who believe that the uncertain net benefits caused by regulation are higher than most other people do. If so, these relatively extreme perceptions are likely to shape agency beliefs and benefits assessments for regulations.
3. *Use of unvalidated and simplistic models of benefits caused by regulations.* Confronted with uncertainty and complexity in the causal networks that link regulations to their consequences (both intended and unintended), regulators, like other people, often adopt simplistic, inaccurate, or unproved modeling assumptions, such as that adverse health effects will decrease in proportion to reductions in regulated exposures, or that positive exposure-response associations represent manipulative causation. These assumptions can lead to large but inaccurate predictions of the benefits from regulation. Such estimates are then amplified by media reports and public concerns in which the assumption-based numbers are treated as facts, without discounting them for uncertainty.
4. *Failure to focus on manipulative causality.* The epidemiological evidence of harm caused by regulated exposures, and estimates of the presumed benefits of reducing exposures, are based almost entirely on associational-attributive causal findings in important real-world examples such as the Irish coal-burning bans, the US EPA Clean Air Act Amendments, and the FDA ban of animal antibiotics. As previously discussed, such findings have no necessary implications for predictive or manipulative causation. They do not provide a logically sound basis for risk assessments or benefits estimates for proposed future changes in regulations to reduce exposures. Moreover, associational causation can almost always be found by making modeling choices and assumptions that create a statistically significant exposure-response association (“p-hacking”), even in the absence of predictive or manipulative causation. Thus, conflating evidence of associative and attributive causation with evidence of manipulative causation can lead to routinely exaggerated estimates of the benefits caused by regulations.
5. *Failure to learn effectively from experience.* Health, safety, and environmental regulations are usually evaluated during the rule-making process based on prospective modeling and prediction of the desirable effects that they will cause. This prospective view does not encourage learning from data via retrospective evaluation, or designing regulations to be frequently modified and improved in light of experience. But such adaptive learning and policy-refinement have been found to be essential for effective decision-making and forecasting under uncertainty in other areas such as high-reliability organizations, superforecasting, and control of systems under uncertainty. As illustrated by the example of the Irish coal burning bans, the relatively rare retrospective evaluations of the effectiveness of regulatory interventions that are currently conducted are prone to unsound design and confirmation bias. Rigorously designed data collection, evaluation, and modification based on performance feedback are not routinely incorporated into the implementation of most regulations. Thus, estimates of the benefits caused by a costly but ineffective regulation may remain exaggerated for years or decades, leading to widespread perceptions that it was effective and to adoption of similar

measures elsewhere, as in the case of the Dublin coal-burning bans that are now being advocated for nation-wide adoption.

The preceding problems have a single root cause: reliance on fallible and overconfident human judgments about causation. Such judgments tend to overestimate the benefits of regulations and neglect or underestimate uncertainties about them, thus promoting more regulation than needed to maximize net social benefits. We have argued that, fortunately, more objective and trustworthy data-driven estimates of the effects actually caused by regulations and of uncertainties about those effects are now technically possible, and that they are also organizationally possible and practical if judicial review of causal reasoning and claims is strengthened. Advances in data science have yielded demonstrably useful principles and algorithms for assessing and quantifying predictive causation from data. Stronger judicial review that incorporates lessons from these methods into the review and application of causal reasoning used to support contested regulations can help to correct the preceding problems and to obtain many of the benefits of more accurate and trustworthy estimates of the impacts caused by regulations.

The following recommendations suggest how courts can promote better regulatory benefits assessment, impact evaluation, and adaptive learning to increase the net social benefits of regulations.

1. *Insist on evidence of manipulative causation.* Rules of evidence used in determining whether it is reasonable to conclude that a proposed regulation will probably cause the benefits claimed for it should admit only evidence relevant for manipulative causation. This includes evidence of predictive causation, insofar as manipulative causation usually implies predictive causation. It also includes evidence on causal pathways and mechanisms whereby changes in exposures in turn change harm, based on well-validated and demonstrably applicable causal laws, mechanisms, processes, or paths in a causal network. Reject regulatory actions proposed without evidence of manipulative causation. Insofar as they provide no sound reason to believe that the proposed actions will actually bring about the consequences claimed for them, they should be viewed as arbitrary and capricious.
2. *Exclude evidence based on associational and attributive causation.* Such evidence is not a logically, statistically, or practically sound guide for predicting effects of regulatory interventions.
3. *Encourage data-driven challenges to current benefits estimates.* Producing relevant (manipulative causation or predictive causation) information about the impacts caused by regulations can improve risk assessments and benefits estimates, but is expensive for those who undertake to develop such information. To the extent that it can increase the net social benefits of regulation by more accurately revealing the impacts of changes in regulations, such information has a social value—a positive externality—and its production and use to improve regulations should therefore be encouraged. One way to do so might be to grant legal standing to parties who seek to challenge current estimates of regulatory impacts based on new information or analyses of manipulative causation (at least

if they also bear either costs or predicted benefits of proposed regulations). A second way might be to emphasize that the burden of proof for changing a regulation can be met by any stakeholder with standing who can show that doing so will increase net social benefits.

4. *Discourage reliance on expert judgments of causation.* Do not defer to regulatory science and expertise based on professional or expert judgements. Instead, insist on data-driven evidence of manipulative causation (including tests for predictive causation and elucidation of causal pathways or mechanisms) as the sole admissible basis for causal claims and estimates of the impacts caused by regulations.

A joint regulatory and legal system that encourages data-driven challenges to the assumptions and benefits estimates supporting current regulatory policies can create incentives for stakeholders—whether advocates or opponents of a change in regulation—to develop and produce the information needed to improve the effectiveness and net social benefits of regulation. It can also create incentives for regulators to adopt more of the habits of high-reliability organizations, regarding current policies as temporary and subject to frequent change and improvements based on data. Setting expectations that judicial review will provide independent, external, rigorous review of causal claims in light of data whenever stakeholders with standing insist on it may also encourage development of lighter-weight regulations that are less entrenched and difficult to change and that are more open to learning from experience and revising as needed to maximize net social benefits.

Of course, there is a large overhead cost to changing regulations that makes too-frequent change undesirable (Stokey 2008). However, the threat of rigorous judicial review and court-enforced revisions when data show that estimates of benefits caused by regulations are either unsound or obsolete would encourage regulators to develop sounder initial causal analyses and more modest and defensible estimates of the benefits of actions when manipulative causality—and hence the true benefits from regulation—are uncertain. This provides a useful antidote to the above factors that currently promote over-estimation of uncertain regulatory benefits, with little penalty for being mistaken and little opportunity for stakeholders to correct or improve estimates based on the judgments of selected experts.

In summary, more active judicial review of causal claims about regulatory impacts, with data-driven evidence about manipulative causation being the price of entry for affecting decisions, creates incentives to expand the socially beneficial role of stakeholders as information collectors. Simultaneously, active and rigorous judicial review of causal claims provides a mechanism to help regulators learn to perform better. It does so both by serving as an external critic and reviewer of causal reasoning and predictions on which contested actions are predicated, and also by providing an opportunity for new information and different data-informed views to be brought to bear in assessing the actual effects being caused by current contested policies. An adversarial system allows different stakeholders to produce relevant information, both confirming and disconfirming, for evaluating the hypothesis that current regulatory policies are performing as predicted in causing desired effects. Active judicial review of causal claims supporting contested regulations by a court that is known to apply BCA or law-and-economics principles provides incentives for

the stakeholders to produce such information with the intent of reinforcing, revising, or overturning current regulations as needed to increase net social benefits. Doing so always coincides with increasing the net benefits to at least some of the stakeholders, since increasing the sum of net benefits received by all affected individuals implies increasing the net benefits received at least some of them. Thus, judicial review can promote production and use of causally relevant information and help regulators to learn from experience how to make regulations more beneficial. This is not a role that can easily be played by other institutions.

If courts develop, maintain, and routinely apply expertise in dispassionate, data-driven causal inference, both the threat and the reality of judicial review will help to overcome the significant drawbacks of current judgment-based approaches to causal inference for regulatory benefits assessment. Such review will also provide both regulators and stakeholders affected by regulations with incentives and ability to improve the net social benefits from regulations over time.

References

- Aliferis CE, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XS (2010) Local causal and Markov Blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *J Mach Learn Res* 11:171–234
- Apte JS, Marshall JD, Cohen AJ, Brauer M (2015) Addressing global mortality from ambient PM2.5. *Environ Sci Technol* 49(13):8057–8066
- Bareinboim E, Pearl J (2013) Causal transportability with limited experiments. In: Proceedings of the 27th AAAI conference on artificial intelligence. AAAI Press, Palo Alto, pp 95–101
- Bartholomew MJ, Vose DJ, Tollefson LR, Travis CC (2005) A linear model for managing the risk of antimicrobial resistance originating in food animals. *Risk Anal* 25(1):99–108
- Bontempi G, Flauder M (2015) From dependency to causality: a machine learning approach. *J Mach Learn Res* 16:2437–2457. <https://arxiv.org/abs/1412.6285>
- Clancy L, Goodman P, Sinclair H, Dockery DW (2002) Effect of air-pollution control on death rate in Dublin, Ireland: an intervention study. *Lancet* 360:1210–1214
- Coglianese C (2001) Is consensus an appropriate basis for regulatory policy? In: Orts EW, Deketelaere K (eds) Environmental contracts: comparative approachesto regulatory innovation in the United States and Europe. Kluwer Law International, London, pp 93–114
- Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, Hoboken, NJ. ISBN: 13 978-0-471-24195-9, ISBN: 10 0-471-24195-4. <https://archive.org/details/ElementsOfInformationTheory2ndEd>. Accessed 1 Nov 2018
- Cox LA Jr (2017) Do causal concentration-response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality. *Crit Rev Toxicol* 47 (7):603–631. <https://doi.org/10.1080/10408444.20>
- Cox LA Jr, Popken DA (2015) Has reducing fine particulate matter and ozone caused reduced mortality rates in the United States? *Ann Epidemiol* 25(3):162–173
- Cox LA Jr, Popken DA, Ricci PF (2013) Warmer is healthier: effects on mortality rates of changes in average fine particulate matter (PM2.5) concentrations and temperatures in 100 U.S. cities. *Regul Toxicol Pharmacol* 66:336–346
- Cromar KR, Gladson LA, Perlmutt LD, Ghazipura M, Ewart GW (2016) American Thoracic Society and Marron Institute report. Estimated excess morbidity and mortality caused by Air Pollution above American Thoracic Society-Recommended Standards, 2011–2013. *Ann Am Thorac Soc* 13(8):1195–1201

- Dawid PA (2008) Beware of the DAG! *J Mach Learn Res* 6:59–86. Workshop and conference proceedings
- Dekker SWA, Woods DD (2009) The high reliability organization perspective. In: Human factors in aviation. 2nd edn., pp 123–143
- Department of Housing, Planning, Community, and Local Government (2016) <https://www.dccae.gov.ie/en-ie/environment/topics/air-quality/smoky-coal-ban/Pages/default.aspx>
- Dockery DW, Rich DQ, Goodman PG, Clancy L, Ohman-Strickland P, George P, Kotlov T, HEI Health Review Committee (2013) Effect of air pollution control on mortality and hospital admissions in Ireland. *Res Rep Health Eff Inst* 176:3–109
- Dominici F, Greenstone M, Sunstein CR (2014) Science and regulation. Particulate matter matters. *Science* 344(6181):257–259
- EPA (2011a) The benefits and costs of the clean air act from 1990 to 2020: summary report. U.S. EPA, Office of Air and Radiation, Washington, DC. www.epa.gov/air/sect812/aug10/summaryreport.pdf
- EPA (2011b) The benefits and costs of the clean air act from 1990 to 2020. Full report. U.S. EPA, Office of Air and Radiation, Washington, DC. <http://www.epa.gov/oar/sect812/feb11/fullreport.pdf>
- Fedak KM, Bernal A, Capshaw ZA, Gross S (2015) Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol* 12:14
- Frey L, Fisher D, Tsamardinos I, Aliferis CF, Statnikov A (2003) Identifying Markov Blankets with decision tree induction. In: Proceedings of the third IEEE international conference on data mining, Melbourne, FL, 19–22 November 2003. pp 59–66
- Friston K, Moran R, Seth AK (2013) Analysing connectivity with Granger causality and dynamic causal modelling. *Curr Opin Neurobiol* 23(2):172–178
- Furqan MS, Siyal MY (2016) Random forest Granger causality for detection of effective brain connectivity using high-dimensional data. *J Integr Neurosci* 15(1):55–66
- Gamble M (2013) 5 Traits of high reliability organizations: how to hardwire each in your organization. Becker's Hospital Review, 29 Apr 2013. <https://www.beckershospitalreview.com/hospital-management-administration/5-traits-of-high-reliability-organizations-how-to-hardwire-each-in-your-organization.html>
- Gardner D (2009) The science of fear: how the culture of fear manipulates your brain. Penguin Group, New York, NY
- Gelman A, Zelizer A (2015) Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Res Polit*:1–7. http://www.stat.columbia.edu/~gelman/research/published/rd_china_5.pdf
- Greenland S (2005) Multiple-bias modelling for analysis of observational data. *J R Stat Soc A Stat Soc* 168(Part 2):267–306
- Halliday DM, Senik MH, Stevenson CW, Mason R (2016) Non-parametric directionality analysis—extension for removal of a single common predictor and application to time series. *J Neurosci Methods* 268:87–97
- Hammond PJ (1992) Harsanyi's utilitarian theorem: a simpler proof and some ethical connotations. In: Selten R (ed) Rational interaction: essays in honor of John Harsanyi. Springer, Berlin
- Harsanyi JC (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *J Polit Econ*:309–321
- Hendry DF (2004) Causality and exogeneity in non-stationary economic time-series. In: Welfe A (ed) Contributions to economic analysis, vol 269. Centre for Philosophy of Natural and Social Science, London, pp 21–48
- Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58:295–300
- Hill J (2016) Atlantic causal inference conference competition: is your SATT where it's at? <http://jenniferhill7.wixsite.com/acic-2016/competition>
- Höfler M (2005) The Bradford Hill considerations on causality: a counterfactual perspective. *Emerg Themes Epidemiol* 2:11

- Hoover KD (2012) Causal structure and hierarchies of models. *Stud Hist Phil Biol Biomed Sci* 43 (4):741–830. <https://doi.org/10.1016/j.shpsc.2012.05.007>
- Hurd HS, Malladi S (2008) A stochastic assessment of the public health risks of the use of macrolide antibiotics in food animals. *Risk Anal* 28(3):695–710
- Iwasaki Y (1988) Causal ordering in a mixed structure. In: Proceedings of the 27th AAAI conference on artificial intelligence. AAAI Press, Palo Alto
- Joffe M, Gambhir M, Chadeau-Hyam M, Vineis P (2012) Causal diagrams in systems epidemiology. *Emerg Themes Epidemiol* 9:1
- Kahneman D (2011) Thinking, fast and slow. Farrar, Straus, and Giroux, New York
- Kelly O (2016) How the coal ban dealt with Dublin's burning issue. The prohibition of 'smoky' coal in 1990 resulted in 350 fewer annual deaths in city. *The Irish Times*. www.irishtimes.com/news/environment/how-the-coal-ban-dealt-with-dublin-s-burning-issue-1.2367021. Accessed 26 Sept 2015
- Kleinberg S, Hripcsak G (2011) A review of causal inference for biomedical informatics. *J Biomed Inform* 44(6):1102–1112
- Lee S, Honavar V (2013) Causal transportability of experiments on controllable subsets of variables: z-transportability. In: Proceedings of the 27th AAAI conference on artificial intelligence. AAAI Press, Palo Alto
- Lin H, Liu T, Fang F, Xiao J, Zeng W, Li X, Guo L, Tian L, Shootman M, Stamatakis KA, Qian Z, Ma W (2017) Mortality benefits of vigorous air quality improvement interventions during the periods of APEC Blue and Parade Blue in Beijing, China. *Environ Pollut* 220:222–227
- Lucas RM, McMichael AJ (2005) Association or causation: evaluating links between "environment and disease". *Bull World Health Organ* 83:792–795
- Nelson JM, Chiller TM, Powers JH, Angulo FJ (2007) Fluoroquinolone-resistant *Campylobacter* species and the withdrawal of fluoroquinolones from use in poultry: a public health success story. *Clin Infect Dis* 44(7):977–980
- O'Malley AJ (2012) Instrumental variable specifications and assumptions for longitudinal analysis of mental health cost offsets. *Health Serv Outcomes Res Methodol* 12(4):254–272
- Pearl J (2009) Causality: models, reasoning and inference, 2nd edn. Cambridge University Press, New York, NY
- Pearl J (2010) An introduction to causal inference. *Int J Biostat* 6(2):7
- Powell MR (2016) Trends in reported foodborne illness in the United States; 1996–2013. *Risk Anal* 36(8):1589–1598
- Price LB, Lackey LG, Vailes R, Silbergeld E (2007) The persistence of fluoroquinolone-resistant *Campylobacter* in poultry production. *Environ Health Perspect* 115(7):1035–1039
- Rottman BM, Hastie R (2014) Reasoning about causal relationships: inferences on causal networks. *Psychol Bull* 140(1):109–139. <https://doi.org/10.1037/a0031903>
- Schoemaker PJH, Tetlock PE (2016) Superforecasting: how to upgrade your company's judgment. *Harv Bus Rev* 94:72–78. <https://hbr.org/2016/05/superforecasting-how-to-upgrade-your-companys-judgment>
- Schwartz J, Austin E, Bind MA, Zanobetti A, Koutrakis P (2015) Estimating causal associations of fine particles with daily deaths in Boston. *Am J Epidemiol* 182(7):644–650
- Simon HA (1953) Causal order and identifiability. In: Hood WC, Koopmans TC (eds) Studies in econometric method. Cowles Commission Monograph. Wiley, New York, pp 49–74
- Spiegelhalter DJ (1986) Computers, expert systems, and ADRs: can causality assessment be automated? *Drug Inf J* 20:543–550
- Stokey NL (2008) The economics of inaction: stochastic control models with fixed costs. Princeton University, Princeton
- Tetlock PE, Gardner D (2015) Superforecasting: the art and science of prediction. Penguin Random House LLC, New York, NY
- Tian J, Pearl J (2000) Probabilities of causation: bounds and identification. *Ann Math Artif Intell* 28:287–313

- Tikka S (2018) Package “causal effect”: deriving expressions of joint interventional distributions and transport formulas in causal models. The Comprehensive R Archive Network 1.3.6. <https://cran.r-project.org/web/packages/causaleffect/index.html>
- Thaler R (2015) Misbehaving: the making of behavioral economics. W. W. Norton and Company, New York
- Todd BS (1992) An introduction to expert systems. Oxford University, Oxford
- Voortman M, Dash D, Drudzsel MJ (2010) Learning causal models that make correct manipulation predictions with time series data. *J Mach Learn Res* 6:257–266
- Walker C (2016) Courts regulating the regulators. Oxford Business Law Blog, 1 May. <https://www.law.ox.ac.uk/business-law-blog/blog/2016/05/courts-regulating-regulators>
- Wang Y, Kloog I, Coull BA, Kosheleva A, Zanobetti A, Schwartz JD (2016) Estimating causal effects of long-term PM2.5 exposure on mortality in New Jersey. *Environ Health Perspect* 124 (8):1182–1188
- Weick KE, Sutcliffe KM (2001) Managing the unexpected—assuring high performance in an age of complexity. Jossey-Bass, San Francisco, CA, pp 10–17
- Wittmaack K (2007) The big ban on bituminous coal sales revisited: serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black-smoke exposure. *Inhal Toxicol* 19:343–350
- Woodward J (2013) Causation and manipulability. In: Zalta EN (ed) The stanford encyclopedia of philosophy. <http://plato.stanford.edu/archives/win2013/entries/causation-mani/>
- Wu MH, Frye RE, Zouridakis G (2011) A comparison of multivariate causality based measures of effective connectivity. *Comput Biol Med* 41(12):1132–1141
- Wynne B (1993) Public uptake of science: a case for institutional reflexivity. *Public Underst Sci* 2 (4):321–337

Chapter 15

Intergenerational Justice in Protective and Resilience Investments with Uncertain Future Preferences and Resources



This final chapter considers the challenging question of how much each generation should invest in building resilient infrastructure to protect against possible future natural disasters. If such disasters are infrequent, members of each generation may be tempted to defer investments in resilience and protective infrastructure (e.g., in building or improving dams and levees; retrofitting office and residential buildings; creating more robust transportation, power, and communications networks; etc.) in favor of consumption or growth. Succumbing to this temptation imposes risks on future generations of needlessly large losses or disproportionate need to invest in resilience. Yet, even the most dutiful and altruistic present generation has limited obligations to invest to protect future ones, especially if present investments in resilience reduce growth and future prosperity, or if the preferences, priorities, resources, and capabilities of future generations are highly uncertain. This chapter presents several different frameworks for clarifying how much each generation should invest in protection, drawing on and extending ideas and methods for collaborative and adaptive decision-making discussed in earlier chapters. It introduces optimal economic growth models, which provide a well-developed technical framework for maximizing average or minimal expected social utility over time, but require consistency and cooperation over time that may not be psychologically or politically realistic. If investment decisions are viewed as a form of dynamic “dictator game” in which earlier generations choose how to allocate benefits between themselves and later generations, then insights from behavioral economics, risk psychology, and moral psychology suggest that cues related to deservingness and trustworthiness powerfully affect what is perceived as fair and right in such settings. We propose that a Rawlsian concept of justice (what investment decision rules would people choose from behind a veil of ignorance, in which no one knew what generation he or she would be born into?) can address problems of over-discounting long-delayed and uncertain consequences that have frustrated some previous efforts to apply cost-benefit analysis to ethically charged issues involving intergenerational justice.

This chapter closes by suggesting several principles for applying insights from these different frameworks to investments in building resilient communities and mitigating natural disaster risks across generations. Principles of prescriptive, collaborative, and learning analytics introduced in Chap. 1 and developed in the last several chapters—adaptive optimization, collaboration among multiple decisionmakers, and learning from experience—are shown to be useful for dynamic optimization of societal risk management decisions across generations, as well as for the individual, group, and organizational decisions.

Introduction: How Much Care Does Each Generation Owe to Future Ones?

Each of us lives with decisions that our former selves have made—choices about what to read and study, whom to marry, where to live, what to work at, how hard to work, how much to consume, and what to do with any savings. Our satisfaction with the choices made by our prior selves may be mixed, but it is not unusual to find that our own current choices are only modestly responsive to the imagined preferences and evaluations of our future selves (Kahneman 2011). And so we may let slide New Year's resolutions that our future selves will predictably wish we had abided by; spend more on current consumption and less on savings or prudent investments or retirement plans than our censorious future selves will, predictably, deem optimal; and succumb to tempting but risky prospects in the heat of the present moment knowing that, with high probability, we will soon regret them as imprudent choices that we will (predictably) wish we had made differently.

Researchers investigating individual preferences and choices over time and under uncertainty have long noted, and created analytic models of, the dynamic inconsistency of individual plans and intentions formed under realistic (hyperbolically discounted) preferences for future rewards (*ibid*). More usefully, they have developed a large array of devices for improving the time-consistency of present decisions to help us make present choices that are less regrettable to our future selves. These devices range from freezing credit cards in block of ice to changing defaults and nudges (e.g., to opt-in or opt-out of an employer-sponsored savings or retirement plans such as Save More Tomorrow) to encouraging “pre-mortem” thinking about how well-intended plans and projects might come to be seen in retrospect as predictable failures. Such imposition of dynamic consistency and rationality constraints and effortful (“System 2”) deliberation to restrain and modify the choices urged by immediate intuitive and impulsive (“System 1”) responses to the opportunities and temptations before us, is at the forefront of current efforts to improve real-world decision-making for individuals by better integrating and coordinating the recommendations from Systems 1 and 2 (Kahneman 2011). It also corresponds to venerable traditions in personal ethics that extol the virtues of temperance, patience, prudence, thrift, steadfastness, and the like.

Social decision-making, too, raises questions of how best to trade-off the needs and preferences of current vs. future agents. In many cases, these agents include members of present and future generations, rather than only our own selves at earlier and later dates. Crucial practical risk management policies and social investment decisions depend on how we trade-off their interests. For example,

- *How large a tsunami, hurricane, earthquake or other natural disaster should nuclear power plant designers, levee designers, sea wall builders, and other construction and infrastructure engineers design for?* Following experiences with extreme events such as the Fukushima tsunami or Hurricane Sandy, it is common for news stories, editorials, and politicians to urge that more should have been done and should be done now to provide for facilities that would better withstand the stresses of wind and water under such extreme conditions. Hind-sight bias may make such criticisms and recommendations inevitable. But should engineers henceforth design for the most extreme events that are expected to occur on average once every 50 years, or 100 years, or 1000 years, or for some other level of extremity? Designing to protect against more extreme (rarer) events typically costs more, but the future risk reduction benefits purchased by these present investments in safety may be very uncertain, depending on part on future climate and weather.
- *How much should we invest in developing and stockpiling antibiotics and other countermeasures in case of an anthrax (or other) outbreak?* The preventive measures cost money now. Whether they will be used before they expire, and how much good they will do if they are used, are typically quite uncertain. If the decision problem is stationary (i.e., looks the same starting from any point on time), then it may be reduced to a choice between (A) Creating and maintaining some level of stockpiled countermeasures, costing a certain amount per year; or (B) Foregoing the costs and benefits of doing so. In either case, some people alive now and others not yet born may share in the costs, risks, and benefits resulting from whatever decision is made.
- *How much should we spend now in seeking to postpone or mitigate future losses from climate change?* Is it better to commit a substantial fraction of present GDP to climate change mitigation efforts, or to invest in current economic growth and technological and institutional progress, that so that future generations will be better equipped to deal with any problems that materialize?
- *How much should we spend per year on nuclear fusion reactor research?* If there is substantial uncertainty about whether and when fusion-generated power production might become practical (e.g., at least producing more power than it consumes), then how much should the current generation invest in fusion R&D? Under what conditions (if any) should it stop investing in favor of more promising alternative uses of current resources? If future generations were allowed to vote on or have input to current decisions, how (if at all) would that change the answers?

These examples, and countless others in the realms of public finance and investments, consumption of non-renewable resources, costly exploration and discovery initiatives, storage of radioactive wastes, and even military investments in long-term security and readiness, illustrate the challenges and importance of learning to address preferences and trade-offs in decisions with consequences that affect multiple generations. All of them can be posed generically as decisions about *how much care (whether measured in prevention, investment, or other expenditures) current decision-makers should take in order to benefit future recipients of the benefits of these decisions*. Making such decisions wisely requires constructive frameworks for determining both *what* we should decide to do now, and *how* we should decide what to do now, given that some of those who bear the consequences may not be alive yet to participate in the decisions.

Optimizing trade-offs between present costs of taking care (or investing in risk reduction) and uncertain future risk-reduction benefits requires an approach to decision-making different from traditional subjective expected utility (SEU)-maximizing decision analysis. The fact that there are multiple stakeholders, not all of whom are alive at present, moves the decision problem from the domain of a single decision-maker to the less well-developed domain of multiple agents, for whom justice and cooperation over time in the absence of opportunities for explicit collaboration and agreement may be important. Even defining clearly what “optimize” means in such settings is difficult. At the individual level, what seems the best choice in retrospect may be very different from what seemed best in prospect, when decisions had to be made and when many possible futures were being considered. In retrospect, evaluations of alternatives can be powerfully affected by hindsight bias, regret, blame-seeking, availability bias, and other potent psychological influences on how we judge previous decisions (Kahneman 2011). In prospect, our evaluations and judgments about what course of action seems best are influenced by different biases, including over-optimism, narrow framing, and over-confidence (*ibid*). Defining a single concept of “optimal” choice that satisfies the biases of both forward-looking and backward-looking evaluations may be impossible. At the group level, what most people prefer to do now may differ predictably from what most of them (or future generations) will later wish had been done now—and from what they may praise or blame current policy makers for doing or failing to do. Such differences can arise because of predictable shifts in beliefs and information, or from changes in tastes and preferences, as well as because of predictable differences in prospective and retrospective preferences and evaluations of alternatives.

The following sections present several frameworks for thinking constructively about how much we should invest in safety, resilient infrastructures and communities, precautionary measures against attacks or natural disasters, sustainable production and consumption practices and processes, and other instances of costly present care taken to create potential future risk-reducing benefits that will affect multiple generations. To simplify and clarify key issues, we first introduce simple, idealized models of multi-generational conflict and cooperation in the gradual consumption of a desirable resource, generically called “capital stock” or, more colloquially, “pie.” This can be variously interpreted to illuminate issues such as savings vs. investment

in growth of capital stock in economic growth models; optimal consumption of a scarce resource that many generations would like to share (e.g., oil, or susceptibility of microbial pathogens to current antibiotics, or military reputation with allies and with enemies, etc.); investment in a stock of “safety” in infrastructure (e.g., fraction of bridges replaced or renewed no more than a target number of years ago, or height of a levee or sea wall to protect against flooding during extreme weather events); or investments in precaution against future disasters (e.g., stockpiling of antibiotics or other countermeasures). We consider how one might answer the ethical questions of how big a piece of the remaining pie each successive generation should take, and what duties each generation has to add to the stock of pie or capital, if doing so is possible but costly. This multi-generational sharing of a pie, and generalizations in which the size of the pie can change due to the production and consumption decisions of each generation, and perhaps to random events, provide fruitful metaphors for clarifying both economic and ethical aspects of intergenerational justice that arise in many practical applications, including decisions about how much each generation should spend to help reduce risks from natural disasters or terrorist attacks for future generations. We compare several principles and constructive analytic and ethical frameworks for deciding what each generation should do, addressing both what decisions should be made and how they should be made. Finally, the main generalizable insights and results from applying these frameworks to problems of intergenerational cooperation and justice are summarized, with attention to how they might be applied to improve practical decisions in which intergenerational consequences and justice are key concerns. In particular, we discuss implications for how to make ethically defensible investments in protecting against natural hazards, as well as implications for proposed principles of sustainable consumption and production of safety benefits.

Simple Models of Intergenerational Justice: Sharing a Pie Over Time

An idealized model of intergenerational cooperation and decision-making is the following pie-division problem. At the start of the problem (conventionally labeled the beginning of period 1), the first generation is endowed with a pie of fixed size, conventionally normalized to be 1, or 100%. Members of that generation must decide how much of the pie to consume themselves, and how much to leave as a bequest to the next generation. Each successive generation in turn must then decide how much of the endowment of pie that it received from previous generations it will consume, and how much it will pass on to later generations. In the simplest case, the pie does not grow or shrink, except by the consumption decisions of each generation. Of course, generations may overlap, but the simplest models treat them as discrete; this does not substantially affect the analysis.

In variations of the problem, pie not consumed grows at some rate (analogous to savings and investment in economic growth models); or additions to the stock of pie may occur at random times (analogous to discoveries of new deposits of a non-renewable and valuable resource such as oil); or the pie may be perishable or may randomly shrink or disappear (analogous to depletion or exhaustion of a resource, or to random destruction of property by disasters). In each variation, the key question of how much to consume now remains. Especially instructive variations make explicit the possibility of two different types of investment for unconsumed pie: (a) Invest in growth (i.e., save some pie, which may then grow at a certain rate); or (b) Invest in safety, which reduces the amount of pie lost when disasters occur.

Analytic Frameworks for Managing Cooperation Across Generations

This section briefly introduces several different analytic frameworks for addressing the question of how much each generation should consume, save, and invest, allowing for the possibilities of investing in either growth (as in traditional economic models of optimal consumption and savings/investment over time) or in safety.

Economic Growth Models

For the past century, classical and neoclassical economic models of optimal capital accumulation and growth, such as the Ramsey (1928), Solow (1956), Swan (1956), and Phelps (1961) models, have been used to answer questions about how best to manage a society's consumption, savings, and investment over time. These models describe the changing stock of capital (or "pie") over time by differential equations in continuous time or by difference equations in discrete time, such as

$$K(t+1) = (1 + g)(1 - c)[1 - d(t)]K(t)$$

where $K(t)$ is the stock of capital at the start of period t ; g is the fractional growth rate for unconsumed capital (typically between about $g = 0.02$ and $g = 0.10$ per year); c is the consumption fraction (and therefore $(1 - c)$ is the savings fraction, often denoted by s); and $d(t)$ is the random fraction of capital lost in period t due to disasters. In classical economics, the capital good, K , is often interpreted as something like "corn" that can be either consumed, thereby producing immediate utility from consumption; or saved and invested ("planted"), in which case it will grow at rate g and expand future opportunities for consumption and savings/investment. In more elaborate models, the values of c and g may also vary over time. Traditional deterministic economic models of optimal consumption, savings, and growth do not

model losses from disasters (in effect, assuming that all $d(t) = 0$), but do model labor supply and population growth, in addition to capital growth. Utility in period t is usually assumed to be based on the amount consumed then, $cK(t)$. A central planner who wishes to maximize the net present value of utility (or, in some variations, the steady-state sustainable level of utility) can solve a dynamic optimization problem, constrained by the differential or difference equations for capital accumulation (and for growth in the population and labor force, if they are modeled), to obtain recommended levels of consumption, savings, and investment for each period (Ramsey 1928). Extensions to stochastic growth models, in which the growth rate of unconsumed capital, g , varies randomly over time, have been developed and analyzed using techniques such as stochastic dynamic programming to illuminate the trade-offs between present and future consumption implied by different growth rates (or distributions of growth rates, if they are random) and policies (Olson 2005).

Useful qualitative insights flow from such economic models, which characterize the set of possibilities for intergenerational choices and their probable consequences over time. Models of capital consumption and accumulation specify the physics, or rules of the game, within which intergenerational sharing of resources, consumption, and production take place. They also reveal qualitative properties of optimal policies. For example, analysis of growth models reveals conditions under which all optimal growth trajectories (i.e., sequences of states, consisting of consumption and savings/investment levels in each period, that jointly maximize discounted utility or similar objective functions) approach each other and stay close to each other along most of their lengths, regardless of the initial state. Such optimal growth path “turnpikes” exist for many deterministic and stochastic growth models. In simple models, the optimal policies have simple and intuitively appealing economic interpretations, with each period’s consumption and investment being optimized by adjusting consumption levels to equate the marginal utility from further current consumption to the discounted expected marginal value of further investment (i.e., the product of the marginal productivity of investment and the marginal utility from consuming the incremental output next period) (*ibid*).

Perhaps most interestingly, stochastic growth models reveal that there may be a critical threshold level of the initial capital stock above which it is guaranteed (with probability 1) that the optimal policy will never exhaust the capital stock, but below which there is a risk (or, if the growth rate g is variable enough, a certainty) that even optimal management will eventually end with $K = 0$ (economic collapse). Such results have implications for the management of fisheries and other renewable resources. If extinction is a realistic possibility, then keeping the stock above the threshold needed to avoid extinction might well be an overriding priority, trumping all other considerations about what each generation owes to future ones. Including the possibility of disasters in growth models ($d(t) > 0$) can modify optimal policies, e.g., by providing reason to restrict current consumption to provide an adequate margin of safety. For a non-renewable resource (i.e., growth rate $g = 0$), the optimal policy for sharing the initial stock among generations may shift toward more consumption earlier on if disaster might destroy some or all of the remaining stock. Whether such generalizations hold depends on details of the model

considered, such as whether utility of consumption exhibits increasing or decreasing marginal utility (or some of each, perhaps being S-shaped, e.g., if consuming very little oil per generation has zero or little marginal utility compared to consuming more, but consuming a lot also generates less utility per barrel than consuming less). Likewise, incorporating disaster mitigation opportunities into a detailed model requires specifying the cost curve or technology possibilities for spending $K(t)$ to reduce (shift leftward the distribution of) $d(t)$, and how long the effects of such expenditures last. For example, investing in higher sea walls or levees consumes more of the current capital stock that might otherwise have been spent on consumption or invested in economic growth, but reduces the probable losses from floods during the useful life of the walls or levees. Understanding the relation between present costs and future benefits, as modeled by the leftward shift in the loss terms $d(t)$ in future periods purchased by a current investment in safety, provides the essential technical information about possible costs and benefits of disaster risk reduction needed to decide what to do in a multi-period optimization model.

If the model includes population sizes and labor forces, and if a value function for reducing lives lost in the event of a disaster is included in the objective function (thus inviting the usual vexed questions about how to value statistical lives saved), then economic optimization models can deliver recommendations for consumption and investments (in growth and safety) in each period that take into account this evaluation of lives saved. The effects on optimized current consumption of allowing for potential disasters depend on model details; in various specific models, they range from consuming more now (“Get it while it lasts!”) to consuming less now in order to protect and expand future opportunities (“Safety first,” “Make hay while the sun shines,” i.e., invest while it is still productive to do so). If the objective function is smooth and exhibits diminishing marginal returns, then optimizing multi-period consumption, investment, and savings (e.g., via stochastic dynamic programming) typically requires equating the marginal returns from incremental expenditures of $K(t)$ on present consumption, on savings and capital growth, and on disaster mitigation, assuming that future decisions will also be made optimally.

For our purposes of comparing different frameworks for making intergenerational consumption, growth, and safety (i.e., disaster mitigation) investment decisions, the point of including the random disaster term $d(t)$ is not to study optimal growth policies in specific models, but only to point out that standard economic methods for studying optimal consumption and investment over time in models of capital and growth can easily be modified to investigate how the possibility of capital-destroying disasters—and of safety investments that mitigate them—changes optimal policies. Overall, the possibility of investing in precautions that stochastically reduce the damage done by disasters (sacrificing some present consumption or investment in growth by instead spending some of $K(t)$ to shift the distribution of $d(t)$ leftward, toward smaller values) provides an alternative to savings and investment as a way to increase capital stock and production and consumption possibilities over time. Concerns for intergenerational justice are addressed only implicitly, by making decisions in each period, or for each generation, to maximize the objective function (e.g., expected discounted utility) for all. A fundamental limitation of all such

models is that no single objective function may correctly represent the preferences of different generations, or even of social planners living at different times. Not only might future tastes and preferences for consumption vs. safety trade-offs and future societal attitudes toward accepting or mitigating disaster risks differ from present ones, but also future generations whose wellbeing is discounted in present objective functions might wish that a different objective function had been optimized. (Worse, if current choices about wealth vs. safety affect the existence or sizes of future generations, then the hypothetical preferences of potential future individuals might be considered to matter in deciding what should be done now. But the hypothetical preferences of as-yet non-existent individuals provides at best a speculative basis for making present choices.) Allowing different generations to have, and to act on, different objective functions from the current generation's requires shifting our analytic perspective from multi-period economic optimization to game-theory to better understand how the choices of different generations interact over time.

Behavioral Game Theory Framework for Strategic Interactions Among Generations

Game-theoretic frameworks for deciding how much of its inherited stock of goods each generation should allocate to consumption and how much it should bequeath to its successors via investments in disaster mitigation measures and economic growth differ radically from optimal economic growth models. They drop the fiction of a single dispassionate social planner who is willing and able to make decisions for each generation to maximize some overall objective function. The difference can be illustrated as follows. Suppose that the first generation, perhaps motivated by compassion or ethical considerations, intends to consume only a small share of the initial endowment of a non-renewable resource ("pie"), leaving the rest for posterity. If they somehow learned that the second generation plans to consume the entire remainder, passing nothing on, then generation 1 might be inclined to revise its initial generous intent, consuming more itself. But, if it turns out that generation 2's intent to consume everything is based on discovering that generation 3 plans to consume whatever is bequeathed to it, passing nothing on, then generation 1 might feel that generation 2 is not so undeserving after all. In short, what each generation concludes that it should do might well depend on what it expects subsequent generations to do. If a generation trusts that a plan that it initiates for enriching the future will be followed faithfully by at least the next few generations, then it may bequeath more than if it lacks such trust.

How such trust arises is better illuminated by behavioral game theory, experimental economics, psychology, and descriptions of what is sometimes called "social capital" (trustworthiness of others with whom one participates in dealings) than by the logical prescriptions of formal game theory models, in which trust plays no role.

Real people are often far more altruistic and cooperative than models of purely rational behaviors and interactions predict or explain. For example, in the much-studied dictator game, one player is given an amount of money (or other desirable good) to divide between himself and a second player. The recipient has all the power in this game, and might be expected to keep everything for himself. But this is not what is observed in many experimental dictator games and variations: most dictators choose to bequeath substantial shares (e.g., 20% or more) of the initial endowment to the other, powerless players, depending on what social norms are evoked by the contextual cues of the experiment, such as earning, sharing, pure giving, etc. (List 2007). To what extent generous sharing is observed in practice depends on many contextual factors, such as whether participants view their interactions in a market frame or in a gift frame; on whether taking as well as giving is included in the feasible set of actions; on whether the player selected as the dictator believes that the choice reflects his own skill or luck in a fair contest; on how often the situation is repeated. But purely selfish behavior is seldom observed (List 2007). The multi-generation division of an initial endowment can be viewed as an expansion and generalization of the dictator game in which each generation is in the position of the dictator in deciding how much to share with powerless future generations.

Although behavioral game theory and experiments provides insights into realistic behaviors that formal non-cooperative game theory (e.g., based on the understanding that all rational players will use subgame perfect equilibrium (SPE) strategies if they exist and are unique) cannot, neither type of analysis is concerned primarily with clarifying what choices are most morally correct. Behavioral game theory and behavioral economics recognize that people (and some other primates) have intuitive and emotional responses to perceived fairness, equity, injustice, reciprocity, and altruism that are important drivers of decisions about when and how much to share in a variety of settings, including the dictator game and its variants. These personal moral intuitions and impulses are tremendously important in helping real people to cooperate more successfully than purely rational (SPE-implementing) agents can in many situations (e.g., one-shot and iterated Prisoner's Dilemma, the stag game, the centipede game, the trust game, and other staples of modern game theory). Yet, they do not provide a coherent normative account of applied social or moral decision-making that can be used to obtain reliable moral guidance on important policy questions, such as how best to share the burdens and benefits of investments in disaster mitigation and in economic growth across generations. To the contrary, our moral intuitions are easily dumbfounded when situations appeal to several competing moral intuitions (as in the trolley track problem, or in alternative framings of tax breaks for dependents as providing disproportionate benefits to the wealthy, if they are used; or as imposing disproportionate costs on the wealthy, if they are not). Likewise, the prescriptions from formal mathematical non-cooperative game theory models of interacting rational players (or generations, in our setting) are not intended to convey any moral authority: SPE solutions and refinements only guarantee that no player can get more of what it wants by changing its strategy, and not that what any player wants is morally worthwhile.

Axiomatic Solution Concepts from Cooperative Game Theory

A different branch of game theory concerns itself directly with normative principles that participants in a cooperative enterprise might adopt to decide how to share the benefits from cooperation. This is axiomatic game theory, which defines and studies the logical relations among solution concepts in cooperative games, such as those involving bargaining or fair division of one or more goods. For example, the Shapley value, which assigns to each player the expected incremental value that the player creates by joining a coalition of other players (when the order in which the players join is random), is the unique way to allocate gains from cooperation that satisfies certain axioms, such as that players be treated symmetrically, without favoritism (each player's allocation depends only on what it contributes, and not on who the player is) and that the allocation procedure should be Pareto-efficient, allocating 100% of the potential gains from cooperation. The Nash Bargaining solution, which maximizes the product of the utility gains of the players compared the no-agreement outcome, is likewise the unique solution concept satisfying these two conditions and the additional two that the outcome should not depend on the arbitrary choice of scales for expressing player utilities, and that expanding the opportunity set of utilities jointly available to the players by introducing new options should either leave the original outcome unchanged, or change it to one of the newly added options. Each of these and many other proposed solution concepts for sharing in the cooperative production and allocation of desired outcomes, can be justified as the unique concept implied by a set of more-or-less reasonable-seeming normative axioms, many of which directly represent principles of fairness such as symmetry (i.e., no favoritism), or maximizing the minimum payoff among players, as well as Pareto efficiency.

However, axiomatic cooperative game theory is vulnerable to its own version of moral dumbfounding: different proposed normative principles can conflict logically. This leads to impossibility theorems showing that no possible decision procedure can satisfy all the appealing principles (normative axioms) that one might want to require. For example, the Shapley value solution and the Nash Bargaining solution can prescribe different outcomes for the same situation, so that no outcome satisfies the normative principles proposed for both. In such cases, one might try to decide which principles should be sacrificed so that other (mutually consistent) ones can be preserved, or, equivalently, choose among alternative solution concepts. But there is no meta-ethical framework within axiomatic game theory to prescribe or justify which normative principles to keep and which to abandon when there are logical conflicts among them. Axiomatic theories may also under-determine outcomes in practice. For example, the Nash bargaining solution requires knowing what the disagreement outcome is, but it is not always clear how it should be determined, e.g., as the present *status quo*, which may incorporate the results of many historical injustices, or as an idealized initial position that treats all participants symmetrically. Such questions about how one should, even in principle, implement the prescriptions of normative axiomatic solution concepts, open a gap between their mathematical

implications and the information needed to act on them. Theories of justice developed by Rawls and his successors can help to close this gap by specifying a particular initial position from which (idealized, hypothetical) deliberation and selection of principles proceeds. They also provide a meta-ethical framework for reasoning about how societies should choose among rival normative principles (e.g., among different axiomatic cooperative game-theoretic solution concepts) to guide their subsequent applied choices and the rights, duties, and principles of fair distribution or redistribution over time that they should impose on themselves and on each other. These theories are explained next.

Intergenerational Justice

Philosophical discussions of intergenerational justice since Rawls (2001) have considered an idealized form of social contracting across generations. Participants from all generations are imagined to jointly agree on policies for sharing resources and investments over time, such as social savings rates, from a hypothetical “original position” behind a “veil of ignorance,” in which no participant knows which generation (or other position within society) he or she will be born into (e.g., Stanford Encyclopedia of Philosophy 2008; Manzini et al. 2010). Principles that would be agreed to from behind this veil of ignorance are defined as just principles, to which real policy-makers should adhere if they see to make just decisions. This concept can be applied to inform each generation’s consumption and investment decisions, including decisions about how much to consume and how much to invest in disaster mitigation and community resilience, or directly in economic growth, at different times. In the simple case of sharing a non-renewable resource (“pie”) over time for a finite number of generations (which may be random, if the duration of the human race is uncertain), the allocation of the resource across generations recommended by such a Rawlsian criterion typically coincides with the allocation that would be achieved by social utility maximization in growth economics, but the two may differ when production is possible (Llavadora et al. 2010).

Identifying just policies as those that would result from social contracting if all stakeholders started from a symmetric original position solves the problem of having earlier generations exploit their asymmetrically powerful position to the detriment of later generations, e.g., by consuming all of the initial stock of pie immediately. Because participants in the multi-generation social contract arrived at from the original position do not know which generation they will occupy, they are motivated to treat all generations fairly. This framework for intertemporal justice also provides an alternative to discounting, thus avoiding the ethical problem posed by conventional discounting in cost-benefit analysis of under-weighting the costs and benefits borne by far-future generations compared to those borne by present or near-future generations (Van Liedekerke 2004; Parfit 1982). From the original position, the interests of different generations are valued equally, and so any discounting would

reflect only real asymmetries, such as in production opportunities (e.g., earlier investments in growth pay off over more years), and not a bias against later generations.

Other proposed features of intergenerational justice have been erected on these foundations by speculating about what people in the original position would agree to. Rawls himself argued that the primary duty of each generation is to bequeath just institutions to the next; once this has been fulfilled (via an “accumulation phase”), a frequently proposed secondary duty is that each generation should pass on to the next an endowment at least equivalent (in size or productivity) to the one it received, so that no generation’s consumption reduces the opportunities or wellbeing of those that follow. However, such proposed principles of *sustainability* in consumption and production invite scrutiny and skepticism (e.g., Wolf 2007), especially if they are asserted without careful qualification and reference to specific underlying economic growth models. For example, in the case of an initial endowment of pie that can only be consumed, and not produced, requiring that no generation’s consumption should diminish the stock bequeathed to future generations would imply that none of the pie would ever be consumed—a Pareto-inefficient outcome that would not necessarily appeal to anyone, even behind the veil of ignorance. Similarly, for a model of multi-generational sharing of a renewable resource, Krautkraemer and Batina (1999) show that imposing a sustainability constraint of non-decreasing utility over time creates Pareto-inefficient stockpiling of the resource: everyone would prefer a usage pattern that allowed later generations to have lower utilities than earlier ones. Such conflicts between various proposed criteria for sustainability and what all members of all generations would prefer arise in many other models of intergenerational sharing (e.g., Hoberg and Baumgärtner 2011), although not in all if there is no uncertainty and if property rights, taxes, and transfer payments among generations are dexterously deployed to allow earlier generations to, in effect, purchase resource usage rights from later ones (Howarth and Norgaard 1990). But the frequent conflicts between sustainability principles and economic efficiency (e.g., what is unanimously preferred by all members of all generations) are perhaps also unsurprising from the perspective of a Rawlsian account of distributive justice, insofar as social contracting from an original position in which no participant knows what generation he or she will occupy removes any reason to favor the utility or opportunities of later generations over those of earlier ones.

Despite their considerable intellectual appeal, theories of intergenerational justice based on implicit social contracting behind a veil of ignorance are not free of philosophical and logical difficulties. For example, in such theories, it is not always clear how potential future people whose very existence may be affected by present production and consumption decisions should be treated (Stanford Encyclopedia of Philosophy 2008). Llavadora et al. (2010) present models in which the possible extinction of humanity is considered as a key uncertainty about the future. They prove that it is optimal, in economic growth models with a policy objective of maximizing the minimum welfare across generations, weighted by the sizes of future populations (so that potential individuals, rather than entire discrete generations, are treated as the participants in the social contract) to ignore this uncertainty about

continued survival in deciding how best to allocate consumption and investment over time. This result holds when the economy is sufficiently productive. In effect, the appropriate discount rate due to uncertainty about continued existence is then zero.

Likewise, if future people will have habits, expectations, and preferences that are shaped in part by current production and consumption decisions, then it may not be clear what preferences should be assumed in modeling what they would agree to in the original position. For example, if early generations derive high utility from oil consumption, and if later generations will regard oil consumption as an unattractive addiction that thankfully became obsolete when depletion of oil forced discovery of more attractive alternatives, then a Rawlsian social contract that allowed members of different generations to bring these preferences with them behind the veil of ignorance (but not knowledge of who would be in which generation) might lead to the conclusion that early consumption of oil was just. But if more gradual consumption of oil would lead to all generations putting a high value on it, then these different assumed preferences might imply that gradual consumption was the just pattern. If each alternative choice about patterns of production, consumption, and investment over time induces the generation-specific preferences needed to justify it (by making it the alternative that would be selected from the original position), then the original position loses its prescriptive power.

Sustainability, Protective Principles, and Fundamental Trade-Offs

Sustainability principles, usually requiring that resources or opportunities or utilities be non-decreasing over time (Wolf 2007), are intended to make sure that current policies do not unjustly sacrifice the interests of powerless future generations in favor of the interests of present generations who currently have the power to make choices. Other approaches have the same goal of protecting the interests of future generations in current decision-making; these range from variations on Rawls's idealized social contracting (Van Liedekerke 2004) to idealized Nash bargaining solutions in which each generation is assumed to be given the power to veto plans that it considers unacceptable (Manzini et al. 2010). In addition, some optimal growth models imply a sustainability condition, in which, in steady state, each generation passes on to the next the capital that it received from its predecessor, and this stock is maintained at a level that maximizes the utility from consumption per capita for members of each generation (Phelps 1961). However, when realistic uncertainties about future preferences and choice sets and about the consequences of current decisions are taken into account, the capacity of current decision-makers to protect the interests of future generations effectively may be very limited. For example, Krysiak and Collado (2009) present models in which there is a trade-off between taking actions to protect future generations against risks and taking actions

that all generations prefer, while Hoberg and Baumgärtner (2011) demonstrate similar trade-offs between sustainability and efficiency when irreversible policy decisions are made by earlier generations, trying to protect the interests of later ones, but later generations have better information about the (perhaps unforeseen and unintended) consequences of earlier policies.

More generally, sustainable production and consumption, economic efficiency (i.e., Pareto optimality), Rawlsian justice, and free democratic (non-dictatorial) choice procedures have all been proposed as desirable principles for guiding and constraining how societies should make decisions, both within and across generations. But careful analysis indicates that any two of these principles, when appropriately formalized for specific models relating choices to economic growth, can conflict. For example, free democratic choice procedures may lead to outcomes that no one favors, e.g., if different people have different beliefs about the probable consequences of alternative choices, and these probabilistic beliefs are used to help select policies (Nehring 2007). Similarly, conflicts between sustainability criteria and Pareto-efficiency (Wolf 2007; Hoberg and Baumgärtner 2011) and trade-offs between Pareto-efficiency and various measures of intergenerational equity or justice in resource allocations arise for many intergenerational decision processes (e.g., Krysiak and Collado 2009). No matter how well intended, efforts to protect the interests of future generation using the information available today risks creating outcomes that, in retrospect, no one favors; this is especially likely if today's choices have uncertain, irreversible consequences and if future preferences are uncertain (Hoberg and Baumgärtner 2011). Thus, fundamental tradeoffs must be made among these proposed desirable characteristics of collective choice procedures for managing the distribution of pie or other goods over time and generations. Equivalently, impossibility theorems expressing the logical incompatibility of different sets of principles under stated conditions limit the possibilities for a satisfactory approach to intergenerational cooperation.

Investing in Building Resilient Communities and Societies: An Emerging Framework

If the normative frameworks and principles for intergenerational justice we have considered so far—growth economics, behavioral game theory, axiomatic solution concepts for cooperative games, philosophical theories of intergenerational justice, and proposed principles of sustainability and protection of the interests of future generation—all lead to contradictions or unresolved difficulties, then what type of analysis might be used instead to provide practical guidance on how much to spend on consumption, investments in disaster mitigation, and investments in economic growth in each period? One emerging approach avoids such mathematical and theoretical arguments, instead emphasizing building the capacity of communities and societies to respond quickly and competently to new information and

circumstances as they arise. This is the framework of *resilience* for communities and societies; it is still under development by many investigators. Key concepts are that resilient communities should prepare to manage disasters effectively by accumulating the physical and social capitals needed to adapt and respond effectively and to manage mitigation and recovery efforts when needed (Tierney 2013). Physical capitals include transportation, telecommunications, power, and emergency medical infrastructures. Social capitals include training and preparation, ability of communities to organize effectively and act competently and autonomously when needed, self-reliant communities, and high trust and individual trustworthiness in cooperating to respond to disasters. Proponents of resilience often argue that communities and societies benefit in multiple ways from developing the capacity, responsibility, and self-reliance needed to improvise and take appropriate actions to deal with possibly unforeseen events. From this perspective, the obligation of each generation to the next may be viewed as bequeathing at least a minimal level of resilience, including the needed physical and social capitals.

Resilience frameworks are still a work in progress. They are sometimes conflated with proposed principles of sustainability, participatory decision-making, environmental justice, environmentalism, putative rights to safety and prosperity, and intergenerational equity for managing interlinked social-economic-ecological systems, often with little explicit discussion of the limitations, trade-offs, and contradictions among the principles espoused. Even without such overlays, however, it is useful to keep in mind the basic idea that investment in building community resilience may be a valuable alternative or complement to investments in economic growth or direct protective measures (e.g., better levees) when the possibility of occasional disasters is present.

Ethical Frameworks

Questions about how people should treat each other—including others remote in time or in space who are affected by our choices—have long been addressed by ethical theories. Several are reflected in the preceding frameworks. In brief, utilitarianism and consequentialism are embodied in optimal growth models that choose each period's consumption and investment levels (in growth, safety, resilience, etc.) to maximize a multi-generation social utility function. Deontological ethical theories are reflected in prescriptive models of intergenerational justice, where the implied duty for each generation is to do what is just, such as making investments to maximize the average or minimum utility over all current and future generations (when all others do the same), depending on what rules one believes would be adopted behind the veil of ignorance. Some advocates of sustainability suggest that each generation has a moral duty to invest enough to provide for non-decreasing social utility, opportunities, or resources for future generations. In some simple deterministic models, utilitarian, just duty-based, and intergenerational equity-based prescriptions coincide, as when all three prescribe treating the next generation

(via investment in capital stock) as one would want to be treated by the previous generation—a formulation of the “golden rule” for optimal growth that maximizes steady-state (sustainable) utility from consumption (Phelps 1961). Virtue ethics is reflected in emerging ideas about resilient communities that identify development of resilience with cultivation of private and social virtues such as preparation and prudence, self-reliance, responsibility, trustworthiness even under stress, and capacity to respond to crises competently, cooperatively, and creatively when needed. Thus, the frameworks that we have discussed may be viewed as building on major ethical principles, perhaps helping to move them closer to practical applications in managing intergenerational risks and investment decisions.

Discussion: Principles for Applying the Frameworks to Improve Decisions and Policies

How might an engineer, planner, or policy maker apply insights from the preceding frameworks to improve practical disaster protection and mitigation decisions, such as how high to build a costly levee or sea wall, or how large (and rare) a tsunami, earthquake, flood, or hurricane to plan for in the design of nuclear power plants or other facilities, or how much to invest in a proposed community resilience or civil defense program? The following suggested principles seek to distil from the frameworks implications to help guide practical decision-making when current choices have long-lasting consequences that may affect risks and benefits to future generations.

1. ***Use wide framing. Consider a wide range of alternatives to optimize benefits produced for resources spent, taking into account opportunity costs. Exploit different ways to reduce risk.*** The optimal economic growth framework implies that each method for maximizing a social objective function—whether by investing in economic growth, in reducing potential disaster-related losses, in less costly and more rapid and resilient recovery following disasters, or in other improvements that reduce risk or increase wellbeing—should be funded optimally in each period in competition and combination with the other approaches. In simple settings with diminishing marginal returns, for example, this typically requires funding each alternative up to the point where a different one starts to yield larger marginal returns in improving the objective function. Thus, a planner wondering whether to invest in a higher sea wall or barrier against flooding should ask not only “Is the extra cost of a taller barrier justified by the extra benefit from reduced risk?” but also “Could a larger reduction in risk (or, more generally increase in objective function) be achieved by not making it taller, and instead applying the resulting cost savings to other opportunities, such as relocating people or improving local early warning and transportation systems?” More generally, optimal provision of safety and other good requires considering opportunity costs and optimizing economic trade-offs, while avoiding narrow

framing (Kahneman 2011) that considers only one approach at a time (e.g., investment in levees, but not change in zoning or land use). The optimization problems to be solved can be viewed as allocating each period's limited resources to a portfolio of alternative ways to increase the objective function, with one of those ways being to bequeath more to the next generation, which may have different opportunities.

2. ***Follow golden-rule consumption and investment principles. Do not over-invest (or under-invest) in protecting or benefitting future generations compared to the present one.*** Biases such as the affect heuristic (Kahneman 2011) can encourage simplistic thinking that equates current consumption with selfishness and greed (bad affect) and equates current investment to protect or benefit future generations with benevolence, virtuous self-restraint, and generosity (good affect). Optimal growth models, including ones with ethical and justice constraints, tell a more nuanced story. Under-consumption and over-accumulation of capital stocks to pass on to the future violate the golden-rule maxim of doing in each generation what one would want other generations to do to maximize sustainable utility (Phelps 1961). From this perspective, increasing saving and investment on behalf of the future is not necessarily always better. Instead saving and investing at the golden-rule rate, and not more, maximizes the wellbeing of present and future generations. Thus, optimal economic growth theory weans us from a multi-generation zero-sum perspective, in which increased current consumption necessarily comes at a cost to future generations. Instead, it encourages a cooperative perspective in which members of different generations collaborate in maximizing the sustainable level of wellbeing.
3. ***Use simple rules to help optimize current decisions. Exploit qualitative properties of optimal policies to simplify practical decisions.*** The economic growth perspective can be implemented in detail if trustworthy mathematical or computational models are available representing the causal relation between choices and the probabilities of their consequences (immediate and delayed). Techniques such as stochastic dynamic programming can then be used to decide what to do in each period to maximize a social objective function. Mathematical and computational techniques and resulting solutions can become quite sophisticated and complex, but, in many settings, the optimal solutions have qualitative properties that can inform and improve practical decision-making with simple rules that take into account future effects, even when detailed models and numerical optimization results are not available. For example, both optimal growth and Rawlsian justice might require first boosting economic productivity as quickly as possible to a level where desirable institutions can be sustained and passed on from one generation to the next. Once there, optimal growth policies often have simple characteristics, such as saving and investing just enough so that the marginal productivity of additional capital stock offsets (equals) its effective depreciation rate due to aging, population growth (which dilutes the capital-per-worker), and other causes, including occasional disasters or catastrophes (Phelps 1961). Risk management to jointly optimize consumption and investments in growth, disaster prevention and mitigation to maximize average utility of consumption per capita

per period might require keeping capital stocks of renewable resources at or above certain threshold levels to avoid risk of collapse, which could reduce or eliminate their availability to subsequent generations (Olson 2005). Such simple characterizations of optimal growth and risk management policies can help to focus practical policy-making analysis and deliberation on a few key questions, such as whether the current savings and investment rate is clearly above or below the socially optimal rate (e.g., the golden rule rate, in a Solow growth model (Phelps 1961)); or whether stocks of renewable resources are currently above or below safety-stock thresholds. The answers then suggest directions for remedial actions, such as increasing or reducing investment, respectively. Pragmatic constraints may limit how much adjustment can be made how quickly. In short, knowledge of the qualitative properties of optimal policies, such as the existence of thresholds or of optimal rates of capital accumulation or investment, can produce simple decision rules (e.g., take action to increase investment or stock of a renewable resource if we are below the optimal level, or to decrease it if we are above the optimal level, where the optimal level is estimated from data on depreciation rates or renewal rates, respectively). Such simple rules can often help to answer the practical policy question of what to do next, even without explicit formulation, estimation, and solution of sophisticated multi-period optimization models.

4. ***Do not discount the utility from future benefits. Count lives saved in different generations equally and count increases in utility received in different generations equally.*** In particular, do not discount future lives saved or future utility from benefits received simply because they are in the future. This follows from Rawlsian justice models that treat the interests of future participants in an extended multi-generational social contract symmetrically with present ones. It implies that benefits such as improvements in quality-of-life per person per year due to increased resilience and reduced anxiety, or greater consumption utility per capita-year, should not be discounted. In making cost-benefit comparisons, lives saved or life-years improved that accrue over the lifetime of a facility should all be counted equally according to such models of justice over time. The practical effect of this recommendation is to increase the present evaluation of benefits that flow from current decisions into the future, such as the benefits from risk reductions obtained via current investments in levees or in other protective or resilient infrastructure. Although multi-period optimization methods such as stochastic dynamic programming can still be used to decide what to do in detail, if credible models are available to support the required calculations, concern for intergenerational justice will modify the usual objective function of expected discounted social utility to give equal weights to life-saving or other intrinsically valued benefits received at different times.
5. ***Consider the value of waiting. Do not commit prematurely to expensive present actions with long-lasting or irreversible consequences. Trust future generations to help decide what is best.*** This principle requires current decision-makers to consider the potential value of seeking and using better information to improve decisions before committing resources or foreclosing other options. For example,

it may be worthwhile for Federal regulators to let individual states experiment with new measures first, and to learn from the consequences, before deciding on a Federal policy that all states must follow. This cautious principle, of seeking to learn more before betting large-scale investment decisions with lasting consequences on what currently seems to be the best choice, follows from studies of fundamental trade-offs in making protective investments under uncertainty, such as the trade-off between investing in proposed measures to protect future generations against possible future harms vs. investing in other ways that, in retrospect, all members of all generations might prefer (Krysiak and Collado 2009; Hoberg and Baumgärtner 2011). Acknowledging realistic uncertainties about future costs, benefits, risk attitudes, preferences, technology alternatives, and opportunity costs highlights the potential value of seeking more information before committing to decisions with long-lasting or irreversible consequences, such as about the height of a levee, enactment of enduring regulation of carbon dioxide emissions, diminishment of economic growth rates in order to invest in protective measures, or consumption of non-renewable resources.

Our first principle above, wide framing, encourages planners confronted with a proposed costly measure to reduce risks to future generations to ask not only “*Is it worthwhile?*” in the sense that the proposed measure’s benefits exceed its costs, but also “*Is there a cheaper way to achieve the same benefits?*” The latter question is typically a matter for engineers and economists. The answer is often that no one knows yet. If further research can reduce current uncertainties, then value-of-information (VOI) calculations from decision analysis can address the question of whether the benefits of that research, in terms of improving decisions and their probable outcomes, exceed the costs of doing it, including potential costs from delay. (Indeed, such VOI considerations are automatically included in stochastic dynamic programming whenever acquiring more information is a possible choice.) Thus, the planner should also ask a third question: “*Is it worthwhile to pay for better information before deciding whether to approve the proposed risk-reducing measure?*” When decisions have long-lasting consequences that affect future generations, the value of information acquired to help make the best decision—the decision that will be preferred in retrospect when future information becomes available—may be especially great.

Likewise, there may be a value to keeping options open, recognizing that the best choice based on current information may not still be seen as the best one when evaluated using future information. There can be a “real option” value to keeping options open until better information is available on which to act, even if delay is costly. Again, stochastic dynamic programming considers such real option values as well as VOI, and optimizes information collection and the timing of decisions, including ones with irreversible or long-lasting. However, the guiding principle of stochastic dynamic programming is the Bellman optimality principle: that each generation’s (or period’s) decisions are made optimally, assuming that all future generations’ (or periods’) decisions will likewise be made optimally (Olson 2005). Practical application of this principle across generations requires decision-makers in

different generations to collaborate in implementing it consistently, with each generation acting accordingly, but having to trust other generations to do the same. Behavioral game theory suggests that such cooperation is far more likely than would be expected based on purely rational (System 2) responses, in part because of moral psychology and pro-social impulses (System 1 responses) that make us eager to reciprocate the generosity of earlier generations by being, in our turn, equally generous to our successors. However, System 1 responses are notoriously non-quantitative, and are not designed to identify and optimize quantitative trade-offs (Kahneman 2011). Thus, deliberate investments in social capital, a culture of trustworthiness and effective cooperation, and building resilient communities, may help generations to collaborate more effectively over time in implementing long-term plans that benefit all of them.

Conclusions

Making decisions well over time is challenging for societies as well as for individuals. Our moral intuitions often deliver altruistic and benign impulses toward others, including strangers separated from us in time or by geography. But they usually do not render finely calculated decisions about how to optimize trade-offs between our benefits and theirs. Nor do they identify the most efficient use of protective and other investments (e.g., in growth of economic prosperity, or in building community resilience) to accomplish desired trade-offs, or to carry out the prescriptions of ethical and justice theories to maximize the average or minimum wellbeing of members of present and future generations. Methods of multi-period optimization that have long been used in optimal economic growth models can accomplish these quantitative trade-off and optimization tasks. They can be adjusted to incorporate principles of intergenerational justice, such as assuring that the lives and utilities of future people are not discounted relative to those of people now living. In simple models, including the multi-generation pie-charting example that we started with and in golden-rule optimal growth models (Phelps 1961), multi-period optimization leads to simple consumption and investment rules that also satisfy equity and sustainability conditions. In this way, System 2 methods can be used to help identify multi-generation investment plans to achieve ethical goals that System 1 might approve of. The results can often be expressed as simple decision rules that are useful for informing practical policy-making, such as taking actions to adjust levels of investments or of renewable resources toward desired target levels based on estimated marginal rates of return or renewal rates, respectively, perhaps with adjustments for the value of information and of keeping options open.

However, even when clear and simple rules can be identified for maximizing a social objective function, such as the average or minimum utility per capita in present and future generations, it takes cooperation across generations to implement them. In turn, this may require just and effective institutions, high social capital, and community resilience, as prerequisites for effective multi-generational cooperation

in managing losses due to natural disasters or other causes. These insights suggest that successful efforts to improve intergenerational justice and efficiency must be rooted in a deep understanding of human social nature and cooperation over time, and of the possibilities for designing and maintaining effective cultures and institutions for promoting and sustaining such cooperation. They also require clear understanding of the goals that we seek to achieve in collaboration with other generations, and of the trade-offs that we want to make when those goals conflict.

The frameworks and principles discussed in this chapter make a start at clarifying possible goals and trade-offs among them, and the implications of technical principles (especially, stochastic dynamic programming distributed over multiple generations) for achieving them. How to develop institutions and cultures that promote effective cooperation over time and across generations, as well as within them, without necessarily assuming that future generations will share our preferences and values, remains a worthy problem for both theoretical and applied investigation.

Epilog: A Vision for Causal Analytics in Risk Analysis

This book has set forth a particular vision of how individuals, organizations, and societies can use analytics to identify choices that make their preferred outcomes more likely and predictably regrettable decisions less likely. It uses causal models to describe how a relevant part of the world works by describing how outcome probabilities change in response to different choices of actions, policies, or interventions. Chapters 1–9 discussed and illustrated technical methods for using data to learn, validate, and document causal models using techniques such as Bayesian networks, structural equations, and probabilistic simulation modeling. Chapters 10–15 discussed various ways in which causal understanding of the relation between current actions and probabilities of future consequences can be used in risk management.

Causal models, in turn, are used to quantify how probabilities of outcomes change as decision variables or policies are changed. Policies can be thought of as decision rules that map data or observations—what a decision-maker sees or knows—to choices of actions or controllable inputs to a system or situation. Consequences result. Markov decision processes, influence diagrams, optimal control, reinforcement learning, multi-arm bandit problems, and probabilistic simulation models allow quantitative modeling of the causal relationship between actions or policies and outcome probabilities. Given such a causal model and a utility function or social utility function expressing preferences for outcomes and attitude toward risk, optimization algorithms solve for the best actions or policies—those that maximize expected utility or reward or net present value; or, in some formulations, minimize expected loss or regret. Sensitivity analyses and value of information (VOI) analyses characterize the robustness of recommended decisions to remaining uncertainties and help decide whether to acquire more information before acting, given the potential costs and benefits of delay. For decisions made sequentially over time,

dynamic optimization techniques can be used to optimize the timing of interventions. It then remains to implement the recommended decisions or policies, evaluate how well they are working (Chapters 10 and 11), and adjust them over time as conditions change or simply as better information is collected.

Implementing this vision of causal modeling, decision optimization, and ongoing evaluation and adaptive learning requires causal analytics methods to meet the following five technical challenges:

- *Learning*: How to infer correct causal models (e.g., causal Bayesian networks or optimal control models or simulation models) from observational data?
- *Inference*: How to use causal models to infer probable values of unobserved variables from values of observed variables? This subsumes *diagnosis* of observed symptoms in terms of probable underlying causes; *prognosis* or forecasting of future observations from past ones; *detection* of anomalies or changes in data-generating processes; and identification of the most probable *explanation* for observed quantities in terms of the values of unobserved ones.
- *Causal prediction*: How to predict correctly how changing controllable inputs would change outcome probabilities or frequencies, e.g., how reducing exposures would change mortality rates? The answer to this manipulative-causation question is needed to optimize controllable inputs to make preferred outcomes more likely.
- *Attribution and evaluation*: How to use a causal model to attribute effects to their cause(s), meaning quantifying how the effects would have been different had the causes been different? Defining and estimating direct, total, controlled direct, natural direct, natural indirect, and mediated effects of one variable on another are variations on attribution that are enabled by causal graph methods. Evaluation of the effects of an intervention or policy amounts to attributing a change in outcomes or their probabilities to it.
- *Generalization*: How to generalize answers from study populations(s) to other populations?
- *Decision optimization*: How to find combinations of controllable inputs to maximize the value (e.g., the expected utility) of resulting outcome probabilities? Simulation-optimization, influence diagram algorithms, dynamic optimization algorithms such as stochastic dynamic programming, and reinforcement learning provide constructive methods for answering this question.

Chapters 1 and 2 discussed technical methods for meeting each of these challenges for causal Bayesian networks (BNs), dynamic Bayesian networks (DBNs), influence diagrams, and other models equivalent to them. Subsequent chapters applied causal modeling and risk analytics methods, especially probabilistic simulation and Bayesian networks, to several diverse real-world risk analysis and policy challenges. They illustrate how a combination of descriptive, predictive, prescriptive, evaluation, learning, and collaborative analytics principles and methods can be applied to clarify a wide range of public and occupational health risk analysis problems and more general questions of policy analysis for both the short and long runs. Throughout, causal analysis and modeling of risk have played central

roles by quantifying the changes in outcome probabilities, frequencies, or expected values caused by changes in controllable inputs.

The central lessons of this book are that (1) Methods of causal analytics are now sufficiently well developed to support both the theory and practice of risk analysis; and (2) It is important to use them in practice. In particular, causal analytics methods such as structural equation models, causal Bayesian network and directed acyclic graph (DAG) models, and probabilistic simulation models are useful for addressing structural and mechanistic causation, as well as predictive causation. They can be used to model manipulative causation, which is essential to guide effective decision-making. By contrast, currently widely used methods of association-based analytics, such as statistical regression modeling of observational data, are usually much less useful for understanding, predicting, and optimizing the effects of policies and interventions on outcomes of concern, because they do not describe manipulative causation.

Our hope is that the methods for causal analytics explained and illustrated in this book will increasingly be applied by policy analysts, risk analysts, and other analytics practitioners to make data more useful in guiding causally effective decisions—that is, decisions that will better accomplish what they are intended to, while more successfully avoiding unintended and unwanted consequences. We believe that better understanding of the causal relation between actions or policies and their probable outcomes is the key to more effective decision recommendations and to more useful evaluations of their effects. Modern causal analytics methods and software for developing such understanding are now much more readily available than ever before. Using them can transform the practice of applied risk analysis, decision analysis, and policy analysis, enabling practitioners to provide more effective and reliable advice on what to do to achieve desired outcomes.

References

- Hoberg N, Baumgärtner S (2011) Irreversibility, ignorance, and the intergenerational equity-efficiency trade-off. University of Lüneburg Working Paper Series in Economics No. 198. www.leuphana.de/institute/ivwl/publikationen/working-papers.html
- Howarth BR, Norgaard BR (1990) Intergenerational resource rights, efficiency, and social optimality. *Land Econ* 66(1):1–11
- Kahneman D (2011) Thinking fast and slow. Farrar, Straus, and Giroux, New York, NY
- Krautkraemer AJ, Batina GR (1999) On sustainability and intergenerational transfers with a renewable resource. *Land Econ* 75(2):167–184
- Krysiak FC, Collado IG (2009) Sustainability and its relation to efficiency under uncertainty. *Econ Theor* 41(2):297–315
- List JA (2007) On the interpretation of giving in dictator games. *J Polit Econ* 115(3):482–493
- Llavadora H, Roemer JE, Silvestre J (2010) Intergenerational justice when future worlds are uncertain. *Mathematical economics: special issue in honour of Andreu Mas-Colell, part 1. J Math Econ* 46(5):728–761

- Manzini P, Mariotti M, Veneziani R (2010) Intergenerational justice in the hobbesian state of nature. Economics Department Working Paper Series, Paper 108. http://scholarworks.umass.edu/econ_workingpaper/108
- Nehring K (2007) The impossibility of a paretian rational: a Bayesian perspective. *Econ Lett* 96 (1):45–50
- Olson LJ (2005) Theory of stochastic optimal economic growth. <http://faculty.smu.edu/sroy/olson-roy-handbook.pdf>
- Parfit D (1982) Reasons and persons. Oxford University Press, Oxford
- Phelps E (1961) The golden rule of accumulation: a fable for growthmen. *Am Econ Rev* 51 (4):638–643. <http://www.jstor.org/stable/1812790>
- Ramsey FP (1928) A mathematical theory of savings. *Econ J* 38(152):543–559
- Rawls J (2001) Justice as fairness. Harvard University Press, Cambridge, MA
- Stanford Encyclopedia of Philosophy (2008) Intergenerational justice. <http://plato.stanford.edu/entries/justice-intergenerational/>
- Solow RM (1956) A contribution to the theory of economic growth. *Q J Econ* 70(1):65–94. <https://doi.org/10.2307/1884513>
- Swan TW (1956) Economic growth and capital accumulation. *Economic Record* 32(2):334–361. <https://doi.org/10.1111/j.1475-4932.1956.tb00434.x>
- Tierney K (2013) “Only Connect!” Social capital, resilience, and recovery. *Risk Hazards Crisis Public Policy* 4(1):1–5
- Van Liedekerke L (2004) Discounting the future: John Rawls and Derek Parfits’ critique of the discount rate. *Ethical Perspect* 11(1):72–83. <http://www.sehn.org/tccpdf/tccdiscounting%20future.pdf>
- Wolf C (2007) Chapter 21: intergenerational justice. In: Frey RG, Wellman CH (eds) *A companion to applied ethics*. Wiley, New York. http://www.public.iastate.edu/~jwcwolf/Papers/Wolf_Intergenerational_Justice.pdf

Index

A

- Adaptive optimization, 80, 81, 89, 90, 457, 458, 558
Adjustment set, 167, 169, 173, 176, 191, 194, 197, 199, 212, 214, 215
Affect heuristic, 471, 473, 474, 487, 520, 574
Air pollution, 9, 17, 49, 63, 98, 146, 165, 193, 200–203, 207, 218–219, 282, 417–439, 459, 460, 468, 470, 481, 505, 519, 523, 524, 528, 534, 541
Air pollution health effects research, 98, 165, 207–209, 424
Air pollution regulation, 513, 530
Analytica, 49, 51, 52, 55, 135
Analytics
causal analytics, viii, ix, 91, 98, 99, 146, 178, 229, 232, 238, 258, 356, 443, 457, 544, 545, 578–580
collaborative analytics, 4, 8, 82–87, 579
descriptive analytics, 3, 4, 8–18, 70, 81, 282, 309, 333, 427–429
evaluation analytics, 4, 55–75, 82, 237, 240, 375, 439, 443–453, 457
learning analytics, 4, 48, 75–82, 457, 458, 558
predictive analytics, 3, 4, 17–22, 32, 59, 76, 80, 82, 154–184, 236, 258, 355, 375, 376, 542, 544, 546
prescriptive analytics, 3, 4, 9, 17, 22–55, 82, 83, 85
Anomaly detection, 21–22
Antibiotics, 8, 41–42, 193, 252, 313, 314, 317, 322, 324, 329, 334, 348, 367, 371, 525, 526, 550, 559, 561

B

- Arrow-Lind theorem, 519
Asbestos, 200, 204, 397–399, 401, 407, 408, 412
Associative causation, 207–209, 216
Asthma, 8, 99, 167, 168, 252, 253, 257, 259–275, 277–282, 419
Attributable risk (AR), 100, 216–218, 240, 355, 417, 531, 534, 544
Attribution fraction, 355, 363
Attributive causal modeling, 397
Attributive causation, 100, 216–219, 240, 531, 534–536, 550, 551
Australasian Bayesian network modeling society, 135
- Bandit problem, 48, 81, 86, 578
Barriers, 142, 398, 476, 539
Bayes server™ software, 120, 134
Bayes' rule, 40, 43, 46, 102–111, 113, 115, 118, 124, 273, 304, 337
Bayesian networks
causal Bayesian networks, 82, 91, 126–132, 135, 157, 189, 223, 579, 580
dynamic Bayesian networks, 76, 132–135, 143–145, 148, 153, 227, 236, 237, 421, 579
Bayesian statistical decision theory, 46
Behavioral economics, 462, 514, 519–521, 557, 566
Behavioral game theory, 565–566, 571, 577
Belief and decision networks, 134

- Benefit-cost analysis, 457, 458, 463, 485, 513, 514
- Benzene
benzene metabolites, 251, 286, 290, 297, 299–304, 308
- Berkson’s bias, 125
- Beta-Poisson model, 318
- Bias in decision making, *see* Decision biases
- Biological gradient, 200–201, 540, 547, 549
- Bistability, 200, 397, 398, 404–408, 410, 411
- bnlearn*, 160, 161, 168, 176, 186, 189, 191, 199, 214, 228, 236, 239, 255, 273, 274, 294, 296, 304, 543, 547, 548
- Boolean networks, 144
- Bow-tie diagram, 142–143
- Bradford-Hill, 195, 230, 539
- Burden of disease (BoD), 202, 203, 216–219, 417
- But-for causation, 101, 240, 533
- C**
- Campylobacter, 328, 525, 526
- Cardiovascular disease (CVD), 418, 420, 421, 423, 425, 427, 428, 430, 432–435, 437–439, 461, 522, 524, 528
- Caret package, 154, 178
- CARET trial, 202, 204, 221
- Causal analytics, 3–9, 12, 17–19, 53, 62, 75, 81, 82, 84, 98, 99, 233, 579, 580
- Causal Analytics Toolkit (CAT), ix, 7, 12, 14, 145–148, 155, 160, 162–165, 167–171, 173, 175, 176, 178–180, 182, 186, 199, 205, 206, 208, 214, 228, 233, 236, 259, 297, 304, 306, 439
- Causal completeness, 130
- Causal discovery, 55, 57–59, 82, 98, 130, 132, 159, 160, 168, 176, 177, 184–230, 235, 238–240, 258, 545–548
- Causal loop diagram, 226
- Causal Markov Condition (CMC), 129, 130, 132, 148, 185
- Causal sufficiency, 130, 132
- Causation
associative causation, 207–209, 216
attributive causation, 100, 216–219, 240, 531, 534, 535, 550, 551
but-for causation, 101, 240, 533
counterfactual causation, 8, 219, 233, 235, 238, 240, 532, 534
explanatory causation, 8, 98, 194, 235, 236, 239, 532
manipulative causation, 99, 101, 127, 130, 148, 171, 190, 201, 202, 204, 209, 212, 213, 217–218, 220, 232, 233, 235, 237–240, 457, 530, 532, 534, 535, 537, 539, 541–543, 545–547, 550–552, 580
- mechanistic causation, 101, 148, 218, 240, 580
- predictive causation, 100, 144–148, 185, 190, 193, 214, 235, 237, 239, 533, 534, 542, 543, 551, 552, 580
- structural causation, 101, 149–150, 191, 236, 532, 542
- Change-point, 44, 67–70, 424
- Characteristic function, 83–84
- Cheapest Cost-Avoider Principle, 516–518
- Chinese factory workers, 285
- Classification and regression tree
(CART) tree, 112, 148, 155–158, 160, 161, 176, 179, 185, 186, 194, 209, 228, 258, 285, 297, 306, 544, 546, 547
- Coal burning ban, 221, 234, 541, 550
- Coherence, 191, 195, 197–199, 218, 474, 499, 540, 548, 549
- Collaborative analytics, 4, 8, 82–87, 579
- Collective choice
collective choice mechanisms, 84, 470, 486
collective choice paradoxes, 487
collective choice theory, 462, 470, 517
- Collider, 124–126
- Collider bias, 214, 221, 234
- Colliders, 167
- Collider stratification bias, 125
- Common cause, 124, 130, 132, 139, 147, 149, 167, 496, 502
- Common effect, 124–126, 313
- Comparative statics, 229
- Compartment model, 228
- Competing risk, 209–211, 238
- Concentration-response curve, 165–178, 279, 404
- Conditional independence, 52, 55, 126, 130, 132, 157–161, 167, 175, 176, 185, 189, 193, 194, 199, 214, 230–232, 236, 237, 240, 255, 258, 281, 294, 418, 421, 422, 424, 425, 430, 438, 543, 544
- Conditional probability table (CPT), 29, 42, 44, 50, 51, 55, 57–61, 63, 71–74, 82, 87, 101, 112–115, 120, 124, 126–131, 133, 138, 139, 143, 144, 150–153, 159–161, 165, 171, 184–186, 189, 194, 197, 221, 228, 230, 235, 237, 238, 240, 543
- Confirmation bias, 196–201, 204–210, 417, 458, 467, 481, 487, 496, 520, 522, 523, 550
- Confounding, 122–126, 169, 192, 203, 213, 252, 265, 269, 271, 272, 274, 278, 280, 282, 418, 425, 430, 438, 539, 545

Consistency of association, 195–197
 Controlled direct effect, 129
 Cooperative games, 84, 567–568, 571
 Cooperativity in oligomerization, 402–404, 410
 Counterfactual causal methods, 219–223, 238
 Counterfactual causation, 8, 219, 233, 235, 238,
 240, 532, 534
 Counterfactuals, 8, 64–65, 75, 76, 82, 88, 98,
 100, 101, 194, 219–223, 231–236, 238,
 239, 281, 422, 532–535, 541
 Crystalline silica, 9, 210, 401, 412, 443–453
 Customer satisfaction, 9–11

D

DAGitty, 12, 166, 167, 169, 173, 175, 185, 191,
 194, 197–199, 212, 214, 215, 236
 Dashboards, 16–17, 21, 41
 Decision analysis, 3, 8, 22–28, 35, 49, 50, 84,
 102, 135, 151, 152, 463, 477, 483, 488,
 518, 549, 560, 576, 580
 Decision biases, 458, 471–473, 476, 479, 480,
 482, 487, 488
 Decision-making
 group decision, 9, 469, 476, 519
 individual decision, 83, 462, 466, 484
 on-line decision, 79
 organizational decision, 457–488, 500,
 513, 558
 Decision table, 23–24, 76, 84, 88
 Decision theory, 42–44, 46, 76, 77, 88
 Decision traps, 476
 Decision tree, 48, 84, 89, 153, 188
 Deep learning, 14, 21, 76, 81, 90, 191
 Deepwater Horizon, 494, 497, 506
 Descriptive analytics, 3, 4, 8–18, 70, 81, 282
 Deterministic optimal control, 35–38, 88
 Differential equation, 37, 239, 344, 532, 562
 Direct effect, 73–74, 126, 128–130, 165, 167,
 178, 191, 194, 217, 230, 280
 Directed acyclic graph (DAG), 8, 49–54, 56,
 58–60, 71, 72, 74, 76, 98, 112–116, 120,
 123–126, 128, 130, 132, 150, 153,
 157–159, 165–168, 170–173, 176, 177,
 184–186, 188, 189, 194, 195, 198, 200,
 205, 206, 217, 227, 228, 230, 232, 233,
 259, 274, 528, 580
 Discrete-event simulation (DES), 8, 33, 34,
 47, 75, 76, 227, 235, 236
 Discrete-time deterministic optimal
 control, 35
 Distributed control, 85, 86, 508
 Dose-response model, 318–321, 330, 397,
 399, 401
 Dose-response threshold, 392, 397–411

Doubly robust estimation, 220
 Dublin coal burning ban, 234
 Dynamic bayesian network (DBN), 76,
 132–135, 143–145, 148, 153, 236,
 421, 579
 Dynamic fault trees, 137–139, 144
 Dynamic optimization, 35–38, 89, 558, 563,
 578, 579
 Dynamic programming
 stochastic dynamic programming, 39, 78,
 89, 563, 564, 574–576, 578, 579

E

Economic growth, 557, 559, 561, 562,
 564–566, 568, 569, 571–574, 576, 577
 Effects, *see* Controlled direct effect; Direct
 effect; Indirect effect; Natural direct
 effect; Total effect
 Efficiency, 9, 33, 52, 84, 86, 120, 153, 186, 290,
 460, 467, 517, 567, 569, 571, 578
 Enrofloxacin, 525, 526, 528
 Ensemble modeling, 34, 546
 Entropy, 145, 148, 155, 188, 189, 191, 195,
 232, 236, 281, 546
 Evaluation analytics, 237, 240, 375, 417–439
 Event tree analysis (ETA), 139–141
 Evolutionary game theory, 84
 Evolutionary operations (EVOP), 80, 81
 Expected utility (EU), 22–24, 28, 41, 43, 46, 47,
 51, 53, 83, 89, 151, 152, 463, 466, 471,
 483, 578, 579
 Exposure estimation error, 201
 Extensive form, 83–84, 88
 External validity, 56, 62

F

Faithfulness, 130, 132
 Fault tree analysis (FTA), 135–139
 Feedback loop
 negative feedback loop, 226, 229, 408
 positive feedback loop, 200, 397, 398, 401,
 404, 405, 407–411
 Fine particulate matter (PM2.5), 12, 13, 156,
 168, 201, 202, 253, 268–271, 418, 422,
 481, 523–525
 Follow-the-perturbed-leader (FPL), 80, 90
 Food and Drug Administration (FDA), 236,
 323, 324, 326, 328, 330, 376, 525–529,
 550
 Food safety, 329, 364, 513, 519, 526, 528, 530
 Forecasting, 17, 45, 76, 80, 145, 227, 537, 538,
 550, 579
 Fukushima, 494, 496, 499, 503, 506, 559

G

- Game theory, 83–84, 88, 464, 467, 486, 488, 565–567
 Generalizability, 56, 376, 394
 Granger causality, 9, 145, 147, 156, 184, 188, 191, 193, 214, 232, 235, 236, 281, 418, 421, 424, 426, 435, 437, 531, 547
 Groupthink, 195, 499, 520

H

- Heuristics and biases, 195, 196, 458, 462, 494, 506, 513, 519, 521
 Hidden markov model (HMM), 38–42, 134, 143, 237
 High throughput screening (HTS), 8, 355, 375, 378–380, 392–394
 Highly reliable organizations (HROs), 506, 537
 Hill considerations, 191, 193, 201, 213, 215, 218, 230, 422, 539–541, 544–546, 549
 Hill criteria, *see Bradford Hill*
Homo economicus, 466, 467
 Homogeneity, 57, 188, 189, 191

I

- Identifiability, 19, 20
 Importance plot, 161–165
 Indeterminacy of counterfactuals, 222–223
 Inference algorithms, 98, 120, 121, 132, 139, 143, 144, 188, 229, 232, 237, 258, 542–545, 547–549
 Inflammasome, 8, 9, 200, 397–405, 407, 409–412
 Influence diagrams (IDs), 49, 53, 76, 82, 85, 102, 151–153, 237, 578, 579
 Information principle, 190, 195, 217, 258, 543, 546
 Information theory, 14, 112, 145, 148, 194
 Insensitivity to probability, 474
 Insight Maker, 225, 227, 236
 Intergenerational cooperation, 561, 571
 Intergenerational justice, 580
 Internal validity, 62, 192, 213, 214, 420
 International Agency for Research on Cancer (IARC), 100, 192, 197, 210, 213, 236, 288, 377
 Intervention time series analysis, 61–64
 Invariance, 54–56, 58, 188, 189, 191, 197

J

- Joint probability, 69, 74, 84, 103–107, 112–114, 120, 132, 143, 150, 166, 468
 Judicial review, 9, 513–553
 Justice
 intergenerational justice, 580

K

- Kalman filters, 45, 237
 Knowledge-based constraints, 98, 168, 171, 172, 199, 235–237, 239, 548

L

- Latent confounder, 130, 132, 438
 Latent variable, 40, 57, 124, 134, 149, 188, 189, 191, 237
 Law-and-economics, 458, 513–518, 552
 Le Chatelier’s principle, 229
 Learned Hand formula, 508, 515–516
 Learning, 82
 adaptive learning, 236, 550, 551, 579
 deep learning, 14–16, 21, 76, 81, 90, 191
 machine learning (*see* Machine learning)
 reinforcement learning, 35, 37, 77, 78, 85, 86, 89, 458, 483, 486, 488, 578, 579
 Learning analytics, 4, 48, 75–82, 457, 458
 Learning aversion, 457–488
 Liability, 508, 515–517, 533
 Loss averse preferences, 474
 Loss function, 44
 Lucas critique, 54–56, 128, 534
 Lung cancer, 9, 99, 148, 165, 200, 204, 210, 211, 234, 235, 397–399, 401, 531–534, 547

M

- Machine learning, 7, 13, 14, 17, 19–21, 35, 37, 46, 59–61, 67, 77, 81, 84, 86, 97, 99, 161, 178, 181, 182, 184, 190, 231, 237, 238, 258, 259, 294, 297, 305, 394, 463, 486, 488, 538, 539, 542
 Manipulative causation, 99, 101, 127, 130, 148, 171, 190, 201, 202, 204, 209, 212, 213, 216, 217, 220, 232,

- 233, 235, 237–240, 457, 530, 532, 534, 535, 537, 539, 541–543, 545–547, 550–552, 580
- Markov chain, 40, 45, 134, 137, 143, 144
- Markov decision processes (MDPs), 28, 29, 31–35, 40, 75–79, 81, 85, 88, 89, 134, 153, 485, 486, 578
- Maximal ancestral graphs (MAGs), 132
- Mechanism design, 84
- Mechanistic causation, 101, 148, 190, 218, 240, 580
- Mediator, 129
- Mesothelioma, 9, 200, 397–399, 401
- Metabolism, 121, 251, 285, 286, 288–290, 295, 296, 299, 301–303, 308, 309
- Metabolites, 8, 285, 287–290, 294, 295, 299, 301–304, 307–309
- Methicillin-resistant *Staphylococcus aureus* (MRSA), 313, 329, 333–349, 371
- Microbial risk assessment, 121
- Model ensemble, 76, 79, 161, 165, 173, 214, 237, 258, 297, 306, 309, 387, 394, 544
- Monte Carlo Tree Search (MCTS), 47, 48, 76, 81, 90
- Multi-agent influence diagrams (MAIDs), 85
- Multi-agent reinforcement learning (MARL), 86
- Multi-agent systems (MAS), 85–88, 90
- Multi-armed bandit (MAB), 46, 78, 79, 86
- Multi-divisional firm, 83
- N**
- Narrow framing, 458
- Natural direct effect, 129, 165, 178
- Negligence, 458, 515–516
- Netica*, 115–121, 124, 125, 131, 133, 135, 139, 151, 152, 160, 188
- Nlrp3 inflammasome, 8, 200, 398–402, 404, 405, 407–410, 412
- Normal form decision analysis, 23, 24, 27
- O**
- Oligomerization, 397, 400–404, 409–412
- On-line decision, 79
- Optimal control
- deterministic optimal control, 35–38, 88
 - discrete-event simulation (DES), 8, 33, 34, 47, 75, 76, 227, 235, 236
 - stochastic optimal control, 37–42, 88
- Optimal harvesting, 37–38
- Optimal stopping, 25–27, 477, 504
- Ordinary differential equation (ODE), 37, 38, 224, 225, 228, 238
- Organizational decision making, 488, 500, 513, 558
- Organizational psychology, 520
- Organizational risk management, 505
- Overconfidence bias, 520
- Overfitting, 539
- Ozone (O_3), 168, 176, 202, 203, 207, 252, 253, 255, 257, 266, 268–275, 278, 418–420, 424–426, 430, 432–434, 437–439, 469, 470, 524, 525
- P**
- Pareto efficiency, 567
- Pareto-efficient decision, 469–470, 517, 567
- Partial ancestral graphs (PAGs), 132
- Partial dependence plots (PDPs), 129, 167, 168, 173, 175, 184, 191, 201, 214, 236, 251, 278, 544
- Partially observable Markov decision process (POMDP), 38–42, 46–48, 76–78, 85, 88–90, 134, 153
- Path analysis, 149–150, 185, 230, 236, 237, 422, 542, 543
- Pattern recognition, 13, 45, 67, 178–184
- Percolation, 408–411
- P-hacking, 196, 197, 199, 214, 539, 546, 550
- Plausibility, 191, 195, 197–199, 218, 288, 322, 547–549
- Pm2.5 (fine particulate matter), 12, 13, 156, 168, 201, 202, 253, 268–271, 418, 437, 468, 470, 481, 503, 523–525, 547
- Power calculations for causal graphs, 177–179, 181, 182, 184
- Predictive analytics, 3, 4, 17–22, 59, 80, 154–184, 236, 258, 355, 376, 546
- Predictive causation, 100, 144–148, 185, 190, 193, 214, 235, 237, 239, 531, 534, 543, 551
- Predictive maintenance, 21–22, 122
- Premortem, 485, 488
- Prescriptive analytics, 3, 4, 9, 17, 22–55, 82, 83, 85
- Prisoner’s Dilemma, 83, 566
- Probabilistic Boolean network (PBN), 144
- Probabilistic graphical models, 149, 232
- Probability of causation, 100, 209–211, 217, 236, 417, 533, 534, 544
- Proportion dominance, 473
- Prospect theory, 467, 474
- Prospective hindsight, 484, 485, 488, 496

Q

Quantitative microbial risk assessment (QMRA), 314–317, 320–321
 Quasi-experiments (QEs), 5, 61, 76, 230, 418, 420, 439

R

Random forest, 112, 162–165, 167, 169, 173, 178–180, 209, 228, 297, 305, 306, 309
 randomForest package, 161–163, 165, 175, 259, 278, 306, 544
 randomForest, 161–163, 165, 169, 214, 259, 278, 297, 305, 306, 544
 Randomized control trials (RCTs), 5, 55–61, 64, 72
 Rational regret, 460, 463, 482–484, 488
 Rawls, 460, 466, 568–570
 Regression trees, 59, 112, 154–161, 231, 258, 259, 275, 277, 282
 Regulatory risk assessment, 289, 513, 531, 540
 Reinforcement learning, 35, 37, 76–78, 85, 86, 89, 458, 483, 486, 578
 Relative risk, viii, 97, 192, 201, 203, 204, 209, 210, 213, 216–218, 234, 238, 240, 344, 363, 367, 417, 419, 439, 479, 531, 533
 Resilience, 86, 506, 537, 557–580
 Respirable crystalline silica (RCS), 9, 210, 397, 399, 443–453
 Response surface methodology (RSM), 81
 Risk premium, 519, 549

S

Safety culture, 493, 495, 505–507
Salmonella, 178, 180, 219, 313, 317–319, 371
 Scope insensitivity, 474
 Selection bias, 122–126, 149, 167, 169, 192, 214, 278, 422, 499, 520, 539
 Semi-Markov decision process, 33
 Sequential detection algorithms, 66–70
 Silicosis, 399, 401, 443, 444
 Simpson’s paradox, 11–12, 54, 128
 Simulation
 continuous simulation, 76, 225, 236
 Simulation-optimization, 76, 90, 579
 Specification error, 204–207
 State-action-reward-state-action (SARSA), 77, 78, 81, 89
 Static optimization, 89
 Statistical decision theory, 76, 88
 Stochastic dynamic programming, 39, 89
 Stochastic optimal control, 38–42

Stratification bias, 124–126

Strength of association, 191, 194, 545

Stroke, 282

Structural equation

Structural equation model (SEM), 59, 129, 149–150, 188, 228, 230, 236, 237

Sustainability, 569–572, 577

Swine, 8, 252, 328, 333, 334, 337, 342, 343, 345, 347, 348, 355–357, 360, 362, 367–372

System 1, 333, 471, 474, 502, 520, 521, 527, 528, 558, 577

System 2, 333, 412, 471, 474, 502, 520, 528, 558, 577

System dynamics, 36, 46, 76, 224–227, 235, 236

T

Team theory, 87

Temporality, 191, 200–201, 540, 547, 549

Thompson sampling, 79, 82, 90

Threshold for dose-response, *see*

Dose-response threshold

Total effect, 73–74, 127–131, 150, 167, 168, 170, 176, 194, 199, 215

Toxoplasma gondii, 355–363, 365, 366, 368, 369

Toxoplasmosis, 8, 355–359, 363–366, 368, 370, 371

Trade-offs, 356, 372, 464, 482, 493, 508, 509, 560, 563, 565, 571–573, 576–578

Transfer entropy, 145

Transport formulas, 5, 57, 76, 82, 197, 232, 237

Transportability, 56–61, 191, 547–549

U

UCB1, 79, 81, 90

UCT, 81

V

Value of information (VOI), 4, 44, 81, 152, 458, 462, 484, 501, 503–504, 576–578

Veil of ignorance, 460, 466, 517, 557, 568–570, 572

W

Wason selection task, 196–201, 204, 205, 207–210

Willingness-to-pay (WTP), 458, 466, 473–480, 487, 517