LEEDS CAUSAL SCHOOL

CAUSAL INSIGHTS LTD

CAUSAL INFERENCE WITH OBSERVATIONAL DATA

The challenges and pitfalls

26-30 April 2022
Leeds

PETER WG TENNANT

GEORGIA TOMOVA

ROBERT LONG

MARK S GILTHORPE

LECTURE NOTES

Table of Contents

DAY 1

1.1 The Need for a Causal Framework

Learning objectives

- To recognise causal inference as a fundamental aspect of human reasoning
- To recognise causal inference methods as a formal framework for accepting and admitting our causal ambitions

Humans are programmed for causal inference

Causal inference is intertwined with the very essence of humanity. From the moment we are born, we begin learning the rules of cause and effect through intense observation and repeated experiment. What happens when a spoon falls off a table? Does the same happen when we push it off ourselves?

In just a few months and years, we learn how objects and people interact with each other and react to us. Before long, this extends beyond reality itself and into our imagination, as we learn to ask what might have been or how things might one day be different.[1]

Accepting and admitting our causal ambitions

In our daily lives, this empirical understanding of cause and effect works very well. But it struggles once we move beyond what we can immediately control, moving into the unknown and the uncertain. Similar problems have plagued quantitative social science research, which has likewise struggled with causal inference outside of experimental contexts. It has consequently become rare for social scientists to declare when they are aiming to answer causal questions and estimate causal effects.[2] Yet such questions are critical to our understanding of the world and, in turn, our ability to change it: how does A cause B, and what would happen to B if we changed A?

By making our causal aims explicit, and embracing the necessary methods, we gain huge insight into the challenges of analysing observational data, and thereby greatly improve our chances of obtaining robust and meaningful estimates.

Causal inference methods

'Causal inference' methods – as they have come to be known – are a unification of **counterfactual** and **probabilistic** theories of causation into a formal algebraic and/or graphical framework.[3]

In focussing on these methods, we make no judgement or comment on the validity and/or utility of other prominent theories of causality (such as regularity, agency/interventionist, or mechanistic[4]) or the alternative methods they advocate. We do however believe that causal inference methods offer a tremendous aid to understanding some otherwise tricky ideas. This does not make them universally appropriate and absolutely does **not** make them a replacement for careful and rigorous thinking, which we believe is the most important requirement for analysing observational data for causal inference.

1.2 Distinguishing Prediction and Causal Inference

Learning objectives

- Describe the three tasks of data science
- Explain the distinction between prediction and causal inference
- Identify the methodological differences in model building and evaluation between prediction and causal inference

What is the research question?

"*Fanaticism consists of redoubling your efforts when you have forgotten your aim.*" So said George Santayana, without (one hopes) an awareness of its relevance to contemporary health and medical science. The cultural pressure to 'publish or perish' – combined with an abundance of data – has catalysed an abundance of research without clear aims or research questions. This is exacerbated by a widespread lack of in-depth epistemological training, and a systemic expectation for novelty and impact that encourages seeking answers rather than evaluating questions. This distinction is key, because the way a question is framed is crucial for informing the approach to solving it – from data collection to analysis and evaluation. More fundamentally, this starts with recognising whether the question is one of **DESCRIPTION**, **PREDICTION**, or **CAUSAL INFERENCE**.

The three tasks of data science

The goal of most quantitative research can be broadly divided into 3 tasks: **DESCRIPTION**, **PREDICTION**, and **CAUSAL INFERENCE**.[5]

The first of these – description – is focussed on summarising, describing, and visualising patterns and trends from the data. This can be useful for gaining contextual understanding about the occurrence of some characteristic, factor, or condition, and of how it may differ between people, areas, or over time. For example, descriptive questions about diabetes might include: *How many people have diabetes?*; *How does the prevalence of diabetes differ between areas?*; or *How has the risk of diabetes changed over time?* Although rarely headline-grabbing, good descriptive research is vital for understanding the occurrence and distribution of disease, and it should therefore be recognised as an important foundation to all that follows.

The second task – prediction – focusses on identifying and discriminating patterns from the data to determine features or to forecast events. Predictive questions about the condition of diabetes may therefore be: *What is the risk that someone has - or will have - diabetes?*; *Can diabetes be detected from retinal images?*; or *How many adults will be living with diabetes in 2050?* Prediction is tremendously useful for screening and forecasting, helping us to prepare for future events. Moreover, the maturation of various machine learning techniques has offered substantial improvements in the efficiency and accessibility of predictive tasks. Prediction cannot, however, explain how or why certain conditions occur, nor what is necessary to prevent or treat them.

In contrast, the third task – causal inference – focusses on understanding how one thing might influence another, and hence on estimating the extent to which changing one factor might change another. Causal questions about diabetes may thus be: *How much do gym prescriptions reduce the future risk of diabetes?*; *To what extent would routine retinal screening widen inequalities in diagnosis and prognosis between social groups?*; or *What effect would a new regulation around the sale of sugar-sweetened beverages likely have on the future prevalence of diabetes?* Causal inference thus encompasses both our intellectual desire for understanding and our practical desire to find ways of altering the world to make it better.

Although most quantitative researchers will be familiar with the differences between descriptive and analytic research, the distinction between prediction and causal inference remains poorly appreciated.[6] Since both use similar analytical tools (e.g. multivariable regression), causal inference is rarely recognised as an alternative philosophy that requires an alternative approach. Causal effects are hence all too often inappropriately inferred from predictive models, which at best creates confusion and at worst creates harm.

Prediction as pattern recognition and forecasting

Prediction (or classification, where the outcome is categorical) is primarily focussed on explaining variation in an outcome. The word 'explanation' is however misleading, as this is a statistical rather than a mechanistic explanation. The goal is not to explain how an outcome came into being, but to simply recognise or forecast what value it is most likely to take. For some predictive challenges, such as diagnosing disease from medical images, this may require limited or no external contextual knowledge beyond that required to develop (i.e. 'train') the algorithm; but for many situations, external knowledge may be essential.

Where the aim is to predict some future event, for example, it would clearly not be useful to include 'post-dictors' (i.e. consequences of the outcome), even if they offer good 'explanation' in the training dataset.[6] Similarly, if the goal is to predict the natural history of a disease, it would be vital to distinguish between those who did and did not receive treatment, else the 'natural' (i.e. untreated) effect of a disease may end up being substantially masked by the benefits of treatment.[7] The selection of candidate predictors may also benefit from outside knowledge, such as whether a variable is commonly available, accurately measured, or directly or tangentially related to the outcome. Beyond these qualitative concerns, it matters little which specific variables end up in the model, since prediction aims to optimise their joint information as a group. Similarly, the amount by which any one variable contributes to the statistical 'explanation' and the apparent 'effect' or 'contribution' of individual predictors in 'explaining' an outcome cannot and should not be interpreted, although these both remain common.

The primary aim, and challenge, for prediction is therefore building a model with the highest generalisability from the smallest subset of predictor variables. This requires a diligent focus on both parsimony and validation. Predictions derived in one dataset (the 'training data') should ideally be evaluated on a different dataset (the 'test data') to reduce the risk of overfitting the model, although the training and test datasets are often subsamples of a single parent dataset that is randomly split multiple times (a process known as cross-validation).[8] Clearly, the more extensive the validation, the more confidence can be assigned. Unfortunately, few prediction models are properly validated,[9] meaning that the true performance is often greatly exaggerated.[9]

Causal inference as identification and estimation

While almost everyone learns at some point that "*correlation does not equal causation*", few learn how to proceed when causal inference is nevertheless of interest. Without the benefit of randomisation, most observational studies feign an indifference to causality, preferring instead to describe 'associations' or 'potential links'.[2] But it is unclear what utility – if any – is provided from estimating non-causal associations. The uneasy truth is that despite these semantic contortions, most observational studies exploring one or more specific associations are implicitly interested in estimating causal effects, as evidenced by widespread efforts to control for 'potential confounders' and interpret subsequent coefficients (i.e. 'associations'). This ambiguous approach, and the limitations thereof, has likely helped maintain a widespread belief that observational studies are fundamentally unsuited for causal inference, which in turn inhibits the uptake of contemporary causal inference methods.

Quantitative causal inference formally involves three stages: (1) identifying the causal effect of interest (the **ESTIMAND**); (2) choosing an approach, or building a model, that can estimate it (the **ESTIMATOR**); and (3) estimating the effect in an appropriate dataset (the **ESTIMATE**). Separating the process of identifying the quantity being sought from the process of conducting the data analysis diverges radically from the 'traditional' (albeit increasingly discouraged) approach of delving into a dataset and conducting ad-hoc analyses without necessarily having agreed on the best modelling strategy. For a start, the three-stage approach ensures that the research question is explicitly framed before analysis begins. Thus, rather than loosely seeking to identify '**RISK FACTORS**' for some outcome (a rather meaningless concept),[10] causal effect estimands should explicitly denote the effects of interest (e.g. the total causal effect of metformin on fasting plasma glucose concentration).

The process of identifying the estimand also facilitates the integration of external knowledge into subsequent analyses – a fundamental feature of any causal analysis. In randomised controlled trials (RCTs) and quasi-random natural experiments, the necessity of this step may not be obvious since it is built into the design and/or context. Where the treatment or - more generally - the '**EXPOSURE**' has not been randomised, however, causal effects can only be estimated by first providing some external knowledge of and/or hypotheses about the data-generating process.

Prediction vs Causal Inference

The main differences between prediction and causal inference are outlined in the table below. We also recommend Arnold *et al.* (2020) for further reading.[6]

Task	
Prediction	**Causal Inference**
Scientific aim	
To predict values of an outcome The purpose of a prediction model is to utilise all available data in order to accurately predict the values of a specific outcome of interest (possibly in the future).	**To estimate a causal effect** The purpose of a causal inference model is to estimate the causal effect of a specific exposure variable on a given outcome variable.
Types of question	
What will happen? Who will be affected? Are people with X more likely to have Y?	What would happen if …? Why were some people affected? If we changed X, how would it change Y?
Modelling aim	
Maximise variance ‘explained’ (R^2) Prediction is concerned with maximising the outcome variance that is jointly ‘explained’ by the set of covariates in a model, i.e, the greater the R^2 the better the model.	Maximise accuracy of the estimates Causal inference aims to maximise the accuracy of the estimates by removing all other hypothesised associations that might distort the focal relationship, so that the estimated causal effect is as close as possible to the true causal effect.
Variable selection principles	
Balancing predictive power & parsimony The covariates of a prediction model are selected based on their joint ability to predict the outcome, with a focus on balancing the predictive power and parsimony, i.e. an ideal model should predict the outcome with accuracy but should also be generalisable across different data.	External knowledge & judgement When the aim is to estimate a causal effect, covariate selection (i.e. adjustment) should be based on external expert knowledge and professional judgement. The hypothesised causal relationships believed to be operating among the variables in the dataset can be depicted using graphical models such as directed acyclic graphs (DAGs).
The availability of variables The availability of variables is key, regardless of their temporal relationship with the outcome, meaning that a predictor variable (i.e. occurring before the outcome) may be as useful as a ‘post-dictor’ variable (i.e. occurring after the outcome), if the context allows.	The roles of the variables In order to select the most appropriate adjustment set, the role of each variable should be considered with respect to the focal exposure-outcome relationship. It is particularly important to differentiate between variables that act as confounders (i.e. they are causes of both the exposure and the outcome), and variables that act as mediators (i.e. they are on the causal path between the exposure and outcome).
Maximising the joint information Covariate selection is not informed by the relative associations between individual predictors and the outcome; instead, the ‘best’ covariate set is the one that maximises the information captured jointly by all variables as a group, in order to predict the outcome most accurately.	Minimising confounding and selection bias When causal inference is sought, the model should control for all spurious associations, not control for any of the causal association of interest, and not create any additional spurious associations. To ensure this, all variables confounding the focal relationship should be appropriately adjusted (i.e. ‘controlled’) for, meaning they should be included as covariates. Extra care should be taken to ensure that no variables are inappropriately included in

	the adjustment set (e.g. mediators) because this can bias the coefficient estimates away from the truth.
Meaning of coefficients / weights - coefficient estimates	
Uninterpretable A prediction model provides information on the expected value of an outcome but does not provide information on how a change in one or more variables would affect that outcome. The coefficients of the predictor variables do not reflect their relationship or relative importance to the outcome, since both their magnitude and sign depend on the overall set of predictors. The coefficients in a prediction model are, therefore, uninterpretable and should not be used to infer any associations, causal or otherwise.	Interpretable The coefficient of the exposure variable in a model can be interpreted as an estimate of the total causal effect of that exposure on the outcome, i.e. the total expected change in the value of the outcome as a result of a change in the value of the exposure. However, the model does not provide information about the causal effects of the other covariates on the outcome, because each exposure-outcome relationship generally requires a separate adjustment set. Erroneously interpreting the coefficients of model variables other than the exposure is referred to as the 'Table 2 Fallacy'.
Possibility for automation	
Favoured Automation of the covariate selection process is favoured; the specific variables of choice (and their parameterisation) have little importance, as the goal is to find the set of predictors that together predict the outcome best according to a pre-specified fit criterion (e.g. AIC, BIC, adjusted R^2). This can be achieved by automatic processes such as all-subsets regression.	Not possible Automation of the covariate selection process is currently not possible, because the external knowledge and judgement required to appropriately select an adjustment set cannot be inferred from data alone.
Approach to evaluation	
'Goodness-of-fit' criteria Prediction models can be assessed via statistical evaluation of the overall model using 'goodness-of-fit' criteria such as root mean squared error of residuals, (adjusted) R^2 and receiver operator characteristic (ROC) curves. The performance of the model is often assessed on a dataset different to the one used to inform covariate selection, to increase generalisability.	Sensitivity analyses To assess and further refine a model used for causal inference, sensitivity analyses may be conducted to explore 1) any statistical independencies implied by the DAG, which can be empirically tested, and 2) to estimate the magnitude of biases potentially arising from residual confounding or collider bias.

1.3 Counterfactuals, Potential Outcomes, and Identifiability

Learning objectives

- Understand the principles of counterfactuals, and how DAGs may be used to evaluate counterfactual contrasts
- Recognise the distinction between unconditional and conditional exchangeability, and its implications for causal inference

The Potential Outcomes Framework

The **Potential Outcomes Framework** (also known as **Rubin's Causal Model**) is a popular algebraic approach to defining and considering causal enquiries, particularly among epidemiologists[11]. Although it lacks the graphical language to Pearl's later **Structural Causal Model**, the two frameworks are nevertheless convergent in their underlying philosophy and mathematics; primarily because of their formal reliance on **counterfactual** and **probabilistic** reasoning.

While most quantitative social scientists should be familiar and comfortable with conceptualising a cause as probabilistic (i.e. that a 'causal event' may increase or decrease the *probability* of a later 'consequence event', even if the consequence cannot be known with certainty), the idea of counterfactuals may be less familiar. Some grasp of **counterfactuals** and the related idea of '**potential outcomes**' are however useful for understanding the inner workings of the Potential Outcomes Framework, Structural Causal Model, and their various conditions and caveats.

Contrasting counterfactuals

According to counterfactual reasoning, an event X may be considered a *cause* of an event Y if, *contrary to fact*, had X not occurred, then Y would not have occurred.

As an example, imagine that an individual, George, is driving to work and comes to a fork in the road. He chooses to go left (X=left) and arrives late for work (Y=late). Upset, George declares '*I should have gone right instead!*'.

What George implies is that his decision to go left at the fork in the road *caused* him to be late for work because had he gone right (X=right) he would not have arrived late (Y=on time). Of course, there is no way to prove such a statement, as it would require that George was able to travel back to the same moment in space and time to observe what would have happened if he took the right turn.

This problem is known as the **fundamental problem of causal inference**, i.e. once we have observed the consequence or **outcome** (Y=late) of one 'causal event' (the **factual**, X=left), we can *never* know what the outcome would have been (i.e. the value of Y) if the 'causal event' had been different (the **counterfactual**, X=right). Put another way, George had three **potential outcomes** (Y=late, Y=on-time, Y=early) when considering which way to turn, only one of which can ever be observed in the real world.

While this may thought-experiment may seem rather impractical, it nevertheless highlights our philosophical aim of comparing how things would have been different in a counterfactual universe, where all else was the same except for the cause we are trying to study. Formally, this comparison is known as a **counterfactual contrast** between **exchangeable units of analysis** – i.e. 'units' that are equivalent *in every way except for the presumed 'causal factor' of interest.*

Exchangeability

Unconditional exchangeability

Because of the fundamental problem of causal inference, it is impossible (at least within a counterfactual framework) to identify and estimate individual causal effects for individual events[5].

Averaged causal effects may however be estimated at the *group* level by comparing the average outcomes between two exchangeable groups of analysis that. This is the aim of randomised controlled trials (RCTs) and explains why they are often considered to be the 'gold standard' for demonstrating causality in social science and medical research.

Suppose you are interested in identifying whether taking aspirin (X) is an effective treatment for headache (Y). To test this, a large, representative sample of individuals with headaches are selected and randomly assigned to take either a dose of aspirin (x = 1) or a placebo (x = 0). These individuals are then followed up two hours later and observed to have either a headache (y = 1) or not (y = 0).

Although individuals within the study likely differ with respect to both measured and unmeasured characteristics (e.g. original headache severity) that may affect whether or not they have a headache two hours later, the randomisation of individuals to treatment groups ensures that the distribution of such factors is roughly equivalent between the two groups so that, *on average*, those who take a dose of aspirin are **exchangeable** with those who do not (ignoring sampling variability). Randomisation to treatment groups in an RCT thus aims to produce two units of analysis (i.e. the treatment groups) that are **unconditionally exchangeable**.

Because of this exchangeability, we can say that what *did* happen in the treatment group (their '**outcome**' Y when X=1; which we might write as $Y_{x=1}$ or Y_1) provides a good estimate of what *would* have happened to the placebo group (their '**potential outcome**') if they had received the treatment. To estimate the causal effect of aspirin, we therefore simply compare the difference between the true observed **outcome** in the placebo group (Y when X=0, or Y_0) and their **potential outcome** (Y_1), as estimated from the treatment group (i.e. causal risk difference = Y_1-Y_0; causal risk ratio = Y_1/Y_0).

Conditional exchangeability

Unfortunately, very few questions can be answered by randomised controlled experiment, due to problems with feasibility, practicality, and ethnics. It would not be feasible, for example, to randomise people to being either obese or not-obese to estimate the effect of obesity on the risk of diabetes over the next 5-years. Similarly, it would not be ethnical to randomise people to either smoking or not-smoking to estimate the effect of smoking on the risk of lung cancer over the next 5-years. Most social science questions must therefore rely on non-randomised observational data.

As outlined earlier, the key to identifying average causal effects is to create two units of analysis that are exchangeable and compare their outcomes. Whilst trivial in a well-conducted RCT, this represents one of the biggest challenges in non-experimental data. For example, if we tried to estimate the causal effect of an influenza vaccine (X) on influenza diagnosis (Y), we would likely find that those individuals who received the vaccine were systematically different from those who did not (e.g. they would probably be older, more affluent, etc.). Therefore, a simple comparison of the outcomes between those who received the vaccine and those who did not would not be sufficient for identifying an average causal effect, as the differences in outcomes might be attributable to other differences between the groups. However, in principle, identification of a causal effect could be achieved by comparing the outcomes between *subgroups* for which the distributions of all relevant factors are equivalent. Such subgroups would therefore be referred to as **conditionally exchangeable** (i.e. they are exchangeable conditional on these factors).

Identification and estimation

One of the biggest philosophical differences between traditional 'theory-free' approaches to observational data analysis and a formal causal inference approach, is the separation of the theoretical stage of defining what you want to know (**identification**) with your practical attempt to get there (**estimation**).

This distinction has largely been ignored because of a collective reliance on the rules and conventions of randomised experiment studies, where the causal effect we seek (the **estimand**) is so explicit from the study design that it has never needed formally defining.

Where unconditional exchangeability is not possible, however, we must think more formally about what we seek and how we might get there, i.e. we must identify our causal **estimand** so we can work out the best way to **estimate** it.

At the most basic level, this means **identifying** which variables we need to measure and condition on to most closely approximate conditional exchangeability. I.e. which variables are necessary to include in our multivariable regression model (the **estimator**) to accurately **estimate** our **estimand**? If you chose the wrong set, then you will inadvertently estimate the wrong estimand, risking spurious interpretations and conclusions.

This process can be conceptualised as a baking exercise where the **estimand** is the cake that you are hoping to make. Once this has been identified, then the recipe (i.e. model or '**estimator**') for baking that cake (i.e. your **estimate**) can be determined using causal theory. One of the key benefits of causal diagrams, which we introduce in the next section, is that they make it very simple to determine the appropriate set of variables for conditioning, once we have identified our beliefs about the causal relationships at play.

According to the theory, a set of variables C is sufficient to achieve conditional exchangeability if conditioning on C blocks all '*backdoor paths*' between X and Y (i.e. spurious paths that transmit non-causal associations due to one or more common causes – referred to as **confounding**) and leaves all causal paths intact[3]. In practice, conditioning may be attempted by **restricting** the analysis to a single value of C, **stratifying** the analysis over fixed levels of C, or including C as a covariate in a multivariable regression*.* If your assumption you have chosen an appropriate approach (and modelled it appropriately in the regression model), this will provide the best possible estimate of the average causal effect of X on Y.

The identifiability conditions

The Potential Outcomes Framework imposes four 'conditions' on our ability to accurately identify and estimate a causal effect. The first (and most important) of these we have already discussed; i.e. that the units of analysis must be exchangeable. Clearly, we are will obtain a biased estimate of a true effect of a 'causal factor' if we try to compare two very different groups of participants with very difference probabilities of the outcome.

The three other identifiability conditions are **positivity**, **consistency**, and **no interference**, which will be briefly discussed.

Positivity

Exchangeability requires that the 'causal factor' of interest, known as the **exposure,** is assigned 'as random'; whether because it is entirely allocated at random or is allocated random conditional on the set of (**confounding**) variables C. This requires **positivity**, i.e. that all levels of exposure are possible for all participants. This is violated if, for one reason or another, some levels of exposure are not observed in some population subgroups.

This might occur if we don't have a large enough sample of participants within each population subgroup (known as **data sparsity**). Suppose we wanted to study the 'effect' of hours-worked on grade performance, conditional on gender. This would clearly not be possible in an all-girls school and is one reason why randomised controlled trials tend to recruit equal numbers of individuals to each level of the intervention. If the school included a few boys, we would still struggle to estimate this causal effect because of the lack of information among the boys, a problem described as a '**near violation**' of positivity.

Random positivity occurs when we do not have information on one or more levels of exposure in one or more subgroups, but these are still possible. In random positivity, it might be reasonable to estimate the potential outcomes for this (unobserved) level of exposure by 'borrowing' (i.e. interpolating) information from our other observations such as when the exposure is continuous. For example, if you are studying the 'effect' of age on diabetes risk in adults aged 40-70-years, but you don't have any data for adults aged 64. Though not observed, the consequences of age=64 can still be approximated from the information provided by those

<64 years and those >64 years. If interpolation is not reasonable, then we instead must restrict our inferences to only those subgroups where we have available information, e.g. in the example of studying an all-girls school, we could only estimate the relationship between hours-worked on grade performance in girls.

Structural positivity occurs when we do not observe one or more exposure levels because they are impossible. For example, though trivial, it is impossible to study the 'effect' of human age above 150 years on walking speed. More realistically, it is also impossible to study the effect of child welfare support on school performance in all children, because wealthier families do not receive child welfare support. Again, we would clearly have to restrict our inference to families eligible for child welfare support. Structural positivity is arguably more serious than random positivity because it is not simply a problem of insufficient data, but indicates you are trying to estimate an impossible process or an effect that may not exist.

Note: Although we have focussed on positivity in relation to the exposure, positivity also applies to *all* other variables in your model. I.e. all levels of the outcome should be *possible* for all levels of the exposure; and these both should be possible across *all* levels of all confounding variables within C. Therefore, problems of data sparsity can quickly arise even in very large datasets!

No interference between units

The **no interference between units** condition is part of the **stable unit treatment value assumption** (**SUTVA**). Put simply, it requires that there are no 'spillover' effects of the exposure between participants. This is violated if the exposure status of any one study participant someone effects one or more others.

Suppose, for example, you were interested in studying the effect of a new flu vaccine on risk of developing flu in an individual school, giving half the children the vaccine and half the children the placebo. This would likely create inference, because the probability of contracting flu is determined by the prevalence in the community in which you are part. If the vaccine was effective, the prevalence would be much lower and those who are not vaccinated would also benefit. Since the 'unvaccinated' group would thus have a lower prevalence of flu than if no vaccination had occurred, the apparent causal effect would be underestimated.

Consistency

The **consistency** condition is the other part of the SUTVA assumption. Put simply, it requires that the effect of an exposure is the same for all individuals who receive that exposure. This would be violated if the exposure had different effects in different subgroups; because the resulting 'average' effect would not relate to any specific subgroup.

Consistency violations most likely occur in **multidimensional latent variables** such as 'socio-economic position' (SEP) that include multiple domains or **composite variables** that have been algebraically constructed from more than one variable, such as 'adverse childhood experiences' (ACEs). Formally, neither SEP nor ACE have a consistent causal effect because the same level of either variable could reflect countless different variable combinations, each of which are likely to have different causal effects on the outcome. Similarly, a single unit change in the exposure (e.g. from two ACEs to three ACEs) could indicate various different exposures have changed. This is not to say that these variables are not 'causes', but that it is not possible to estimate a consistent causal effect, because you have instead produced a confused average of various different causal effects, the value of which is open for debate.

For further reading see[12].

1.4 Causal Directed Acyclic Graphs (DAGs) and Covariate Roles

Learning objectives

- Explain the concepts behind drawing a directed acyclic graph (DAG)
- Outline the definitions of different covariate roles in a multivariable model
- Distinguish between when to adjust / not to adjust for mediators in a multivariable regression model
- Understand and be able to explain the "Table 2 Fallacy"

Causal Graphical Models & Directed Acyclic Graphs (DAGs)

Causal graphical models serve as a visual summary of the hypothesised causal relationships amongst variables believed to be operating according to *a priori* knowledge, understanding, and – in the case of the relationships being tested in an analysis – conjecture. This visual summary of variable inter-relationships is used in causal analysis and in expert-systems research.[13] Such diagrams are also increasingly adopted in the epidemiological community,[14] yet they remain relatively novel and considerably underutilised.

Causal graphs may be used in a variety of ways: to think clearly about how the exposure, outcome, and potential confounding variables are causally related; to communicate these causal inter-relationships to the reader; to indicate therefore which variables are important to measure; and to inform the statistical modelling process, particularly in the identification of confounders, as well as competing exposures (see later for definitions). Causal graphs are the basis of a formal theoretical framework in which causal relationships can be identified and evaluated. The simplest kind of a causal graphical model is a **directed acyclic graph** (DAG).

A DAG consists of nodes that represent variables (e.g. X, Y) and arrows (or 'arcs') that depict direct causal effects (e.g. X→Y). To describe relationships between variables in a graph, we often read them like an ancestry tree and use kinship terminology. For example, in the diagram X→M→Y, M is a *child* of X and X is a *parent* of M; M and Y are *descendants* of X, and X and M are *ancestors* of Y. A graph is called a *directed acyclic graph* (DAG) if no variable is an ancestor of itself (i.e. no loops exist). DAGs are often depicted with the nodes arranged in temporal or causal order, with the earliest occurring variables on the left and the latest occurring on the right (though this is not mandatory). Arrows in a DAG reflect *a priori* assumptions about cause and effect in a specific context, some based on firm knowledge and understanding of actual or likely relationships between variables, others based on speculative hypotheses (including the specific relationships being examined in the analyses).

Note: These assumptions cannot be inferred empirically from the data on which the analyses are to be conducted but are required *a priori,* to select and interpret the correct statistical model.

Despite the potential visual complexity of some DAGs, they are still an oversimplification of the causal relationships amongst the constituent variables. A DAG does not indicate whether an effect is harmful or protective, whether **effect modification** (otherwise recognised as statistical interaction, which we cover in detail later) is occurring or not,[15] or whether a cause is sufficient or necessary.[16] DAGs correspond to a network of variables with probability distributions (realised as the covariance structure amongst all variables).

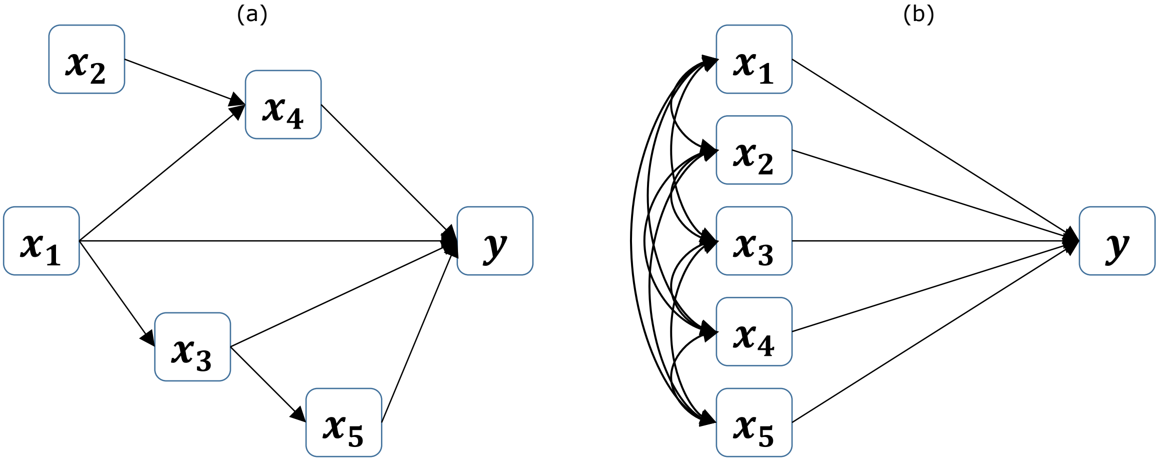
Note: This is not an exact 1:1 correspondence, as there are always multiple network probability distributions / covariance matrices that fit the DAG, and multiple DAGs that correspond to a network probability distribution / covariance matrix. If the DAG contains only observed (i.e. measured) variables, there is always at most one DAG (sometimes none) that fits a probability distribution; if we allow for unobserved or 'latent' variables, there can be several. We should always be mindful of the distinction between causality, as captured by a DAG, and parametric realisation of the variables and their relationships, as described by a covariance structure.

One key strength of a DAG is that it enables researchers to think clearly and logically about their research question(s), and to make explicit any assumptions being made about the relationships amongst the variables included. This visual summary is used to communicate these inter-relationships to other researchers, and hence it is easy to identify if important variables are missing from the DAG or whether any of the relationships are contentious as specified.

Limitations of the linear model

It is important to recognise that a DAG represents what is perceived to be happening causally (i.e. a hypothesised sequence of causal events that might fit the observed data), but when data are examined in a multivariable linear model the inter-variable *correlation structure* is not the same as the inter-variable *causal structure*. This is illustrated in Figure 1.4.1. Here the researcher's *a priori* perception of causal relationships is depicted in the DAG (a), which reflects causes, not correlations, and therefore may look very different to how the variables are 'perceived' statistically speaking (in terms of correlation) by a linear model, as shown in (b).

Figure 1.4.1: The relationship between an outcome y and five covariates $x_1 \dots x_5$: (a) their hypothesised causal relationship within a DAG; and (b) their correlation / covariance relationship in a linear model.



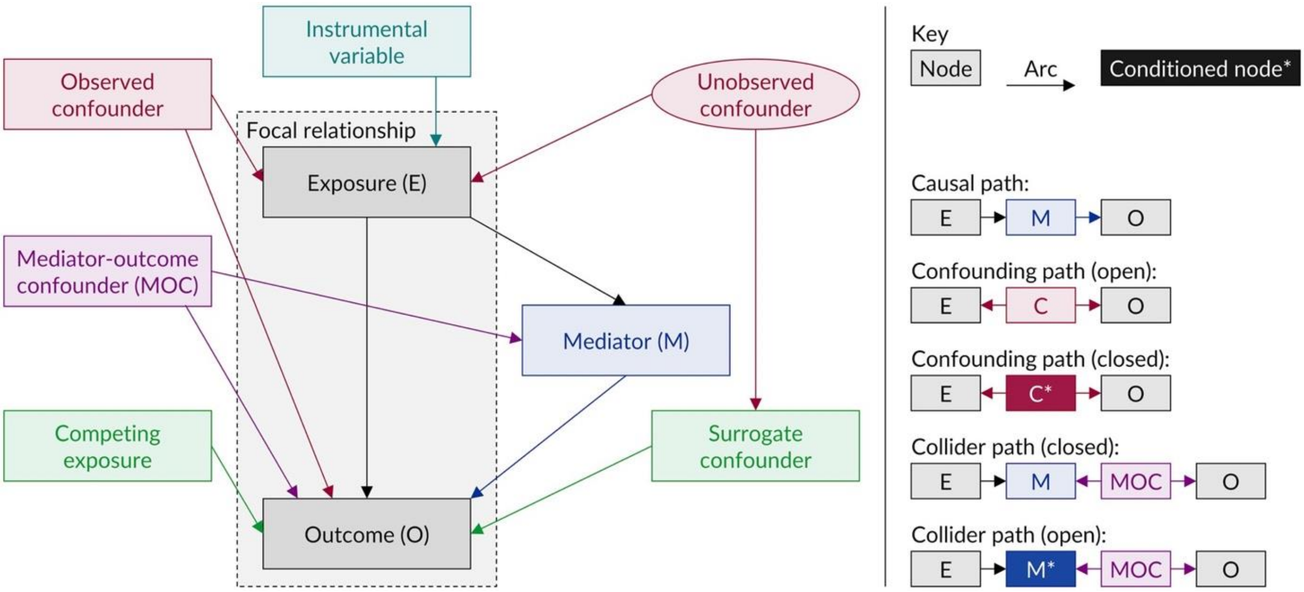
In Figure 1.4.1, the relationship between the outcome y and five covariates $x_1 \dots x_5$ are shown: (a) the presence of directed arrows signify the presumed causal links and the **absence of arrows** depict explicit assumptions of **no direct causal relationship** (in either direction); (b) directed arrows link all covariates to the outcome, suggesting a potential statistical association, and bi-directed arrows also link all five covariates to each other, depicting potential correlation (direct cause is not implied and, were there any causal link, the direction of causality is not specified). The DAG in (a) depicts variables x_1 and x_2 as causally unrelated, and as they share no common ancestor, they are also statistically independent, i.e. un-correlated. The model in (b) depicts all 5 covariates as potentially associated / correlated with the outcome and each other, though in practice, any pairwise correlation may be zero. The meaning and implications of arrows in a DAG are thus very different to the arrows in the linear model, which represents a one-to-many relationship without constraint; hence, the covariance matrix amongst the five covariates and the outcome (denoted $\boldsymbol{\Sigma}$) is freely estimated in the linear model.

When representing variables in a DAG, it is helpful to think about time in line with causality, but this must be derived using *a priori* knowledge *external* to the data. Using graphical model theory, if the data are consistent with the DAG, causality may be inferred, and the extent of causal effect estimated. Causality cannot unequivocally be proven in observational studies, even for longitudinal data, though it is convenient to assume potential causality and use graphical model theory to evaluate this where possible. A careful and principled approach to developing a DAG should always be adopted. We wish to avoid a situation where we have only a vague idea of the potential causal structure; we investigate the data via bivariate correlations or linear modelling; and then cavalierly use the discrepancies between what is observed for the data compared to that predicted from the DAG to revise the DAG accordingly.

Covariate roles

The key purpose of constructing a DAG is to be able to determine the 'role' that each variable has with respect to the exposure-outcome focal relationship (see Figure 1.4.2). This helps inform subsequent model building, i.e. which variables should be included as covariates in the model. The biggest focus has naturally been on adjusting for potential confounders in order to minimise confounding bias, although various other biases may also be avoided by appropriate adjustment (or not) for variables, depending on their covariate roles and study context.

Figure 1.4.2: Illustration of the main components of the DAG, the most common types of contextual variables and the most common types of paths. The DAG has been visually arranged so that all constituent arcs flow from top-to-bottom.[17]



Confounder

There have been various attempts at defining confounding, which broadly divide into two camps: ‘comparability-based’ and ‘collapsibility-based’.[18] In terms of the former, confounding is said to occur when there are differences in the risk of disease and/or healthcare practice in the unexposed and exposed populations that are not due to the exposure, but due to non-exposure variables that may be referred to as confounders. This results in bias in the estimate of the effect of a particular exposure on disease.[19] The second definition is founded on the premise that, in the analysis phase of a study, confounding may be reduced or eliminated by adjusting the analysis for, or stratifying the analysis by, potential confounders.[19] The latter definition is based solely on statistical considerations, and confounding is said to occur if there is a difference in the unadjusted and ‘collapsed’ estimates of the effect of the exposure on disease; estimates are said to have been adjusted for, or stratified by, the potential confounder.

Although both camps are sometimes considered indistinguishable, if confounding is correctly considered to be a causal rather than statistical concept, the comparability-based definition is to be adopted.[20] Based on graphical model theory, the accepted definition of a variable being a confounder within a causal inference framework is that it must be:[21]

- a cause of the outcome
- a cause of the exposure
- unaffected by the exposure (i.e. not on the causal path from the exposure to the outcome)

In DAGs, we can easily recognise confounders as those variables that are ancestors of both the exposure (X) and outcome (Y) via two independent paths; for instance, in X←C→Y, the variable C is a confounder but in C→X→Y, it is not. This is the strict definition of confounding, though careless use of the term ‘confounder’ is often used to describe what we now recognise as ‘**competing exposures**’ or ‘**mediators**’, which we cover next.

Note: In a linear model, confounders are correlated (i.e. *collinear*) with the exposure, which is why adjustment for confounders exploits collinearity to modify the estimated exposure-outcome association. This is an example of a situation in which collinearity is a good thing.

It is not always possible or necessary to measure and adjust for all known confounders. Graphical model theory can be used to search for covariate sets that qualify as 'adjustment sets' that remove all known confounding. The graphical rule used to find such sets is the 'back-door criterion',[22] implemented automatically in the online tool DAGitty (http://www.dagitty.net/) and the **R** package dagitty.[23]

Epidemiological criteria used to check if a variable is classified as a confounder should be based on the comparability definition, though this restricts how variables in a multivariable linear model might be viewed if causality is to be inferred. The previous, liberal use of 'confounder' is not permitted, and new terminology is needed for the more narrowly defined role of different variables in multivariable linear regression.

Proxy

We first introduce the concept of a 'proxy' variable – one that is measured and may sufficiently capture the variation of another variable, which itself may be measurable but not present in the study data (possibly overlooked during the initial study design) or not measurable and therefore latent. Proxies are important as they enable us to internalise a correspondence between what we have in our data and what we seek to frame in terms of 'real-world' concepts. For instance, 'education' may be measured differently – e.g. highest educational attainment or years of schooling, whichever is available in our dataset – yet such measurable variables are only proxies for 'education' that enable researchers to describe relationships between what 'education' encompasses conceptually in terms of cause-and-effect with respect to other variables in their data. There is unlikely to be a perfect correspondence (perfect correlation, statistically speaking) between highest level of education attained and years of schooling. Nevertheless, both measures, despite being imperfect (in effect suffering 'measurement error') allow us to capture the essence of concepts we wish to describe to investigate potential causal relationships in our data. This may seem obvious for variables such as education, but consider age: Do we mean 'chronological age' or 'biological age',[24] as there is a difference?

Less obvious, though equally important, is that some variables in a causal chain may be missing (i.e. not recorded in our dataset), yet their implicit presence is central to the correct drawing of arcs linking variables. Consider the three variables for an individual for whom we have information regarding their parents, their diet during childhood, and their BMI when entering adulthood: 'parental education' (PE), 'childhood diet' (CD), and 'adult obesity' (AO). It seems reasonable to draw a DAG as PE→CD→AO, where we surmise a causal chain from PE through CD to AO, since more educated parents are more likely to provide the kind of childhood environments, including dietary influences, that lead to a lower risk of obesity as individuals enter adulthood. Following this logic, if information regarding individuals' diets were absent from the study dataset, we may nevertheless surmise PE→AO; this does not mean that parental education *directly* causes obesity in offspring's adulthood, but that because 'childhood diet' is not present in the data, we drop CD from the DAG but retain the causal arc from PE to AO because CD is a descendent of PE and ancestor to AO, i.e. PE is a measured proxy of the unmeasured CD. Many hypothetical proxies may exist as descendants of one variable and ancestors for another variable, thereby linking the two by proxy.

It is important to recognise how variables in our data may subliminally supplant a more complex array of factors in which we are interested clinically, biologically, or from an ecological perspective – and necessarily so, as it facilitates the exposition of what are typically complex research questions. The implicit distillation processes we go through to arrive at the models we employ in addressing our research questions are as limited (i.e. 'approximate') as the models themselves in representing 'truth'. We are prone to overlooking such simplifications of what our data represent.

Competing exposure

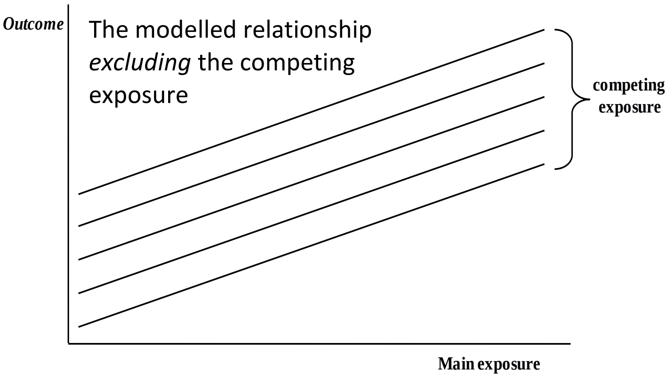
For a variable to be considered a **competing exposure**, it must be:

- a cause of the outcome, or a proxy of a cause of the outcome

- *not* a cause (or a proxy of a cause) of the main exposure
- unaffected by the main exposure (i.e. not a descendant of it)

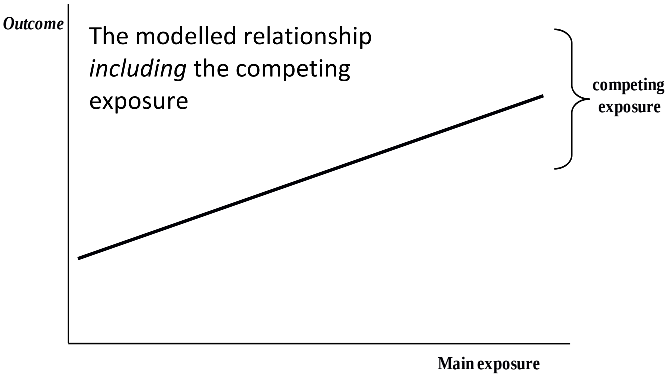
A competing exposure is not a confounder, but researchers often conflate the two.

The estimated association (slope) between a main exposure and outcome is unaffected if the model includes an uncorrelated competing exposure, as orthogonal covariates do not impact each other's estimated coefficients.


Outcome
The modelled relationship *excluding* the competing exposure
competing exposure
Main exposure

In practice, competing exposures are unlikely to be completely orthogonal to the main exposure, especially when these are continuous variables, so their inclusion often changes the estimates very slightly.

If an orthogonal competing exposure is included in the linear model, **precision is improved**, because some of the outcome uncertainty is effectively 'explained' by the competing exposure.


Outcome
The modelled relationship *including* the competing exposure
competing exposure
Main exposure

In the population, a competing exposure and main exposure are assumed to be **causally unrelated** but may be **correlated**.

A competing exposure may be correlated with the main exposure in the **study sample** for one of two reasons:

- Although the main and competing exposures are **causally unrelated** at the population level, in the study data they may exhibit a non-zero correlation due to *chance* sampling; were the study repeated several times, *on average* the estimated association between the main exposure and outcome is correct (hence there is no statistical bias), but for any one study sample with chance correlation, the estimate is modified away from true – i.e. it suffers from some 'error'.

or

- There is an ancestor (observed or unobserved) that causes both, creating a **correlation** at the population level and therefore a likely correlation within any subsample. Inclusion of the competing exposure will modify the main exposure-outcome relationship, which is desirable, as the competing exposure is then also a **proxy confounder** (see next section) – the estimate then suffers no 'error'.

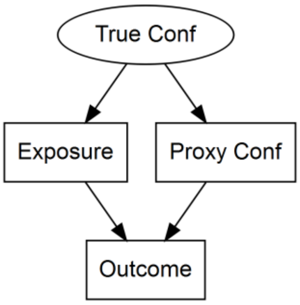
If a competing exposure is correlated with the main exposure in a study sample and this is due to chance (i.e. the first instance), the competing exposure should not be included in the model, as inappropriate modification of the main exposure-outcome estimate occurs, which trumps any advantage of improved precision – the estimate would be more precise but incorrect!

If, however, a competing exposure is correlated with the main exposure and this is due to a common ancestor variable causing both exposures (i.e. the second instance), inclusion of the competing exposure remains favourable (to adjust for proxy confounding), thereby removing bias and improving accuracy, whilst improved precision will also result (as a competing exposure) – a win-win!

The only way to be sure that inclusion is favourable from a confounder adjustment perspective is to use a DAG for all variables considered relevant, available or unobserved, and determine all possible adjustment sets (best done using either the online tool DAGitty, http://www.dagitty.net/, or the **R** package dagitty).[23]

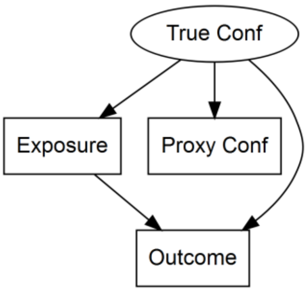
The decision whether to include the competing exposure in the linear model is determined by examination of the *a priori*, appropriately determined, DAG (which is non-parametric and can therefore only indicate which adjustment sets are appropriate rather than the exact nature of the variable relationships); in conjunction with knowledge of the data and its parametric sampling properties (which may indicate a non-zero correlation arising in the sample even though the DAG indicates the correlation should be zero in the population).

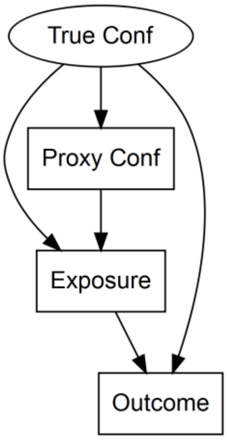
Proxy (surrogate) confounder



In the situation where the main and competing exposures are not directly causally related but are correlated due to a common ancestor variable that is causally related to both (and may even be unobserved; see right), the competing exposure is also a '*proxy* (of the ancestor that is a true) *confounder*', i.e. a **proxy confounder**.

Note: The true confounder need not be a direct parent of the outcome but were we to remove the proxy confounder from the DAG, the link between the true confounder and the outcome becomes direct; being an ancestor is sufficient to be a cause of the outcome.





Proxy confounders are not themselves confounders but lie on the causal path between confounders and either the exposure or outcome (but not both; else they would be confounders).

Proxy confounders are useful if true confounders are unobserved, since the closest thing we have to assess the impact of unobserved true confounders is through their influence via proxies.

Adjustment for confounding by proxy may be less effective than adjustment for confounding using genuine confounders, which may influence the choice of adjustment set (if there is more than one to choose from), but in the event that the true confounders are not measured or measurable, adjustment for proxies may be useful.

Mediator

A common challenge within epidemiology involves the appropriate engagement of mediators in linear regression models. Mediators are variables that lie on the causal path between the exposure and the outcome and transmit part of the total causal, i.e. they are caused by the exposure and in turn cause the outcome. It should therefore be obvious that if we seek to estimate the *total causal effect*, no mediators should be included in the adjustment set because this would adjust for (i.e. remove) part of the effect that we are interested in. However, adjustment for mediators may be justified on rare occasions. For example, if we were interested in estimating the *direct causal effect* of the exposure on the outcome, (i.e. the effect of the exposure that is not mediated by any other variables) we would need to adjust for all mediators. This approach, however, carries a high risk of inferential bias. First, if the causal relationships among all variables have not been sufficiently considered using expert *a priori* knowledge, it is possible that the effect being estimated is not the true direct effect if there are mediators on the causal path that we have failed to properly identify. Second, in some cases it might be that all the meaningful causal effect we are interested in is in fact the one transmitted through the mediators, and without them (i.e. the direct effect) we would not be able to estimate any *meaningful* effects. Third, and perhaps most importantly, adjustment for mediators without sufficient careful consideration and justification, risks invoking a phenomenon known as collider bias, which we will focus on later.

To illustrate some of the considerations that need to take place when deciding whether to adjust for a mediator, we consider two contexts in which the adjustment would have a different impact (by attenuating the estimated exposure-outcome relationship appropriately or inappropriately) and see how this relates to a hypothetical intervention. In our first context, we consider a variable that researchers often adjust for because they see it as a confounder, though in fact it is a mediator (i.e. a context in which mediator adjustment introduce inferential bias to the model estimates and is therefore inappropriate). In our second context, we consider a variable that is well-understood to be a mediator and adjusting for it is imperative to gain correct inference. We explore and explain this apparent contradiction, highlighting the key difference between the two scenarios in terms of hypothetical interventions and thereby indicating when mediator adjustment is appropriate or not.

Context 1: The relation between adult blood pressure and birthweight

In considering a potential relationship between adult blood pressure (BP) and birthweight (BW), researchers have questioned the validity of any association in part due to publication bias and/or inappropriate statistical adjustment for variables on the causal path (such as adult body size),[25] as the latter gives rise to the statistical artefact called 'reversal paradox'.[26] It has also been shown that simultaneous adjustment for two or more intermediate measures of body size exacerbates this artefact.[27] Nevertheless, it is suggested that some intermediate measures (e.g. adult weight, AW) are proxies for genuine confounders that are either unmeasured or, as yet, not identified (e.g. genes that simultaneously affect BW, adult body size, and adult BP).[28] Concern with this argument is that even if intermediate body size measures are a proxy for unmeasured or unknown genuine confounding, the reversal paradox does not go away; there are adverse effects of the artefact induced by the reversal paradox and genuine bias due to adjustment for proxy confounders.[29] In many situations it may be unclear, and therefore remain unresolvable, as to which direction, and of what magnitude, these effects alter the estimated model coefficient for the main exposure; they may be synergistic (add to) or antagonistic (oppose and partly cancel out). In any event, the inferential bias from the reversal paradox never goes away.

It helps to resolve this dilemma by asking: what is the *research question*; what *consequence* are we interested in; and how might we assess this via a (hypothetical) *intervention*?

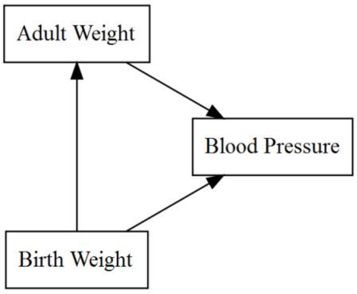
Context

These issues are **context specific**. For instance, do we wish to understand the impact of BW *per se* or, more likely, are we interested in what BW is a proxy for? Biologically, it is unlikely that body mass at birth in a **physical sense** is at all important in relation to adult BP; rather, it is what body mass at birth represents (i.e. what it reflects of **foetal development**), and whether this has some bearing on physiological status in later life. It is widely accepted that BW is a proxy for many things, not least **in-utero nutrition** (both quality and quantity); and the health of the foetus as affected by the health of the mother (both before and during pregnancy). To affect adult BP via intervention at the earliest stages of life (assuming BW were associated with adverse health outcomes in later life), one might seek to affect all factors reflected in the proximal value of BW. We might therefore seek to ensure mothers are fit and healthy before conception, as well as during pregnancy; we might seek to ensure mothers' diets are balanced, containing sufficient nutrients and calories for optimal foetal development; and we might seek to secure a more holistic positive environment to minimise physical and mental stress, avoid adverse lifestyle choices (e.g. alcohol, tobacco), and minimise disease exposures (e.g. measles, tuberculosis).

The complexity of BW as an exposure brings into question what it is that any unmeasured or unknown confounders confound: do they causally influence all, or just some, of the factors encapsulated in the proxy measure of BW? If some unmeasured or unknown confounders were genetic, for instance, how do genes determine the environmental factors that influence BW? Apart from operating via biological mechanisms that drive dietary habits and/or general health-related behaviours, many environmental influences of maternal and foetal wellbeing are determined by geographical, communal, and cultural circumstances, such as the availability of food and medicines (even in developed countries), the risk of exposure to disease or disaster – whether natural (earthquake, floods) or man-made (war) – and parochial dietary and lifestyle norms. This perhaps makes for an argument that any confounding, for which adult weight purportedly acts as a proxy, is tenuous and dilute for each potential confounder. One might argue that many other factors that may seem arbitrary, yet conveniently recordable, could similarly be considered proxy confounders and we soon become awash with possible proxies.

We might seem to overanalyse BW as an ‘exposure’, but this discussion serves to illustrate that the variables we use in a linear model are an abstraction of what we hope they reflect. When seeking causal inference, and thus when considering the role of various measures as confounders or proxy confounders, the **perspective** adopted is **subjective**. Most clinical variables have utility, though often only approximately encapsulating the essence of our research focus. We should remain mindful of this when undertaking linear modelling for causal inference.

Context aside, back to the specifics of the example



Stepping back from these important yet philosophical issues of the context and utility / meaning of the variables in our DAG, we now examine what is meant by adjusting for mediators. We form a theoretically sound perspective by constructing a DAG (see right), putting aside whether or not intermediate body size captures confounding by proxy, and examine the BP-BW relationship as though causal, with focus on a potential intervention just before BW is measured.

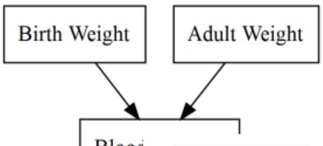
We ask the question: *What is the effect of one unit change in BW on change in adult BP?* We consider this with two model scenarios: one where we have BP as the outcome and BW as the exposure variable and no other covariates (i.e. BP~BW); the other where BP is the outcome, BW is the exposure variable and we include current adult weight (AW) as a ‘confounder’ (i.e. BP~BW+AW, ignoring the fact that AW is not a ‘true’ confounder). Critically, we assume a causal BW-AW relationship, or else AW would not be a mediator either but rather a competing exposure.

In using the model that includes AW to estimate the impact of change in BW on BP, we must evaluate the impact of change in BW on both AW and BP, along with the impact of an altered AW on BP. In using the model that includes only BW, we must evaluate the impact of BW on BP only.

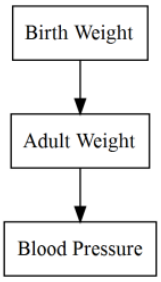
In the causal framework (BW→AW→BP and BW→BP), it is shown mathematically that the evaluated impact of one unit change in BW on BP is identical for both models yet more succinctly captured by the BW coefficient in the BP~BW model. The BW coefficient in the BP~BW+AW model does not reflect the *total* effect of BW on BP, as it must be modified by the effect of AW on BP.

From a causal inference perspective, asking: *What is the (hypothetical) intervention-generated effect of one unit change in BW on the change in BP?* – with the model to yield the answer is the model with only BW included; inclusion of the intermediate AW modifies the coefficient of effect for BW away from the true *intervention-generated* effect. Since BW→BP *and* BW→AW→BP (i.e. AW is a mediator), ‘adjustment’ for AW in the BP~BW model alters the inference sought of the BW coefficient (which is interpreted around the idea of an intervention on BW).

Note 1: If BW is not causally related to AW, it is a competing exposure (see right) and there would be no difference between the two models in the coefficient estimated for BW, and both models would capture the *total* causal effect of BW on BP correctly in the BW coefficient.



Note 2: If BW is not directly causally related to BP (see right), then adjustment for AW should completely remove the effect of BW.



Note 3: Both models (BP~BW and BP~BW+AW) are **statistically unbiased**, as they are correctly estimated; the second model suffers **inferential bias** (i.e. the estimated impact on BP of a hypothetical intervention on BW is biased).

Hypothetical intervention at the time of exposure assessment

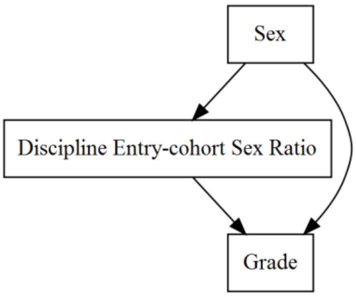
We thus conclude that when seeking to interpret an outcome-exposure relationship causally within a multivariable linear model, where interpretation of the exposure coefficient is predicated on an intervention at the time of (or just before) the exposure assessment, then inclusion in the linear model of mediators biases the model inference and hence its interpretation; the exposure model coefficient does not reflect the total causal impact of any hypothetical intervention on the outcome.

Context 2: Relation between sex and academic career progression

It is generally acknowledged that there are differences between the sexes, though what is due to nature or nurture is still debated.[30] It is nevertheless widely accepted in science (and increasingly accepted culturally, as reflected in legislation) that, notwithstanding variation within each sex, men and women are on average no different in their potential intellectual acuity.[31] It is thus reasonable to presume that it is entirely down to cultural differences experienced throughout life that leads to sex imbalances in the pursuit of different careers. Therefore, within professions for which there is no reliance on physique, uptake of jobs and progression through the ranks should proportionally be very similar. In academia, for instance, the proportion of men and women in each discipline at each grade should be roughly equal. This is, however, far from true (across the globe in fact). In the UK, this led to the formation of an equality charter, **Athena SWAN**: committed to advancing women's careers in science, technology, engineering, maths and medicine (STEMM) in higher education (see: http://www.ecu.ac.uk/equality-charter-marks/athena-swan/).

One metric used in raising awareness, and in monitoring the success of the Athena SWAN charter, is the proportion of women at each grade, e.g. the proportion of women professors per discipline. An implication is that we can assess the 'performance' of academic institutions to 'promote sex equality' through such a metric. The academic workforce today, however, is the product of each individual's experience over the years prior to their first appointment, including their journey through postgraduate and undergraduate training, and before that through secondary education, primary education, nursery, and home-life, along with the wider societal and cultural environments throughout their lives. When examining institutions for potential sex discrimination, we must take account of this.

In contemplating how to investigate academic institutions in their 'fairness' to promote men and women equally, we can look at the proportion of **successful appointments** by sex at each grade, and then *take into account* [i.e. adjust for] *the proportion of men and women applying each time*, though this information is unlikely to be available. Instead, we might take into account [i.e. adjust for] the **proportion of men and women eligible** for each appointment by considering **discipline-specific entry-cohort sex ratios**.



One problem is that entry-cohort sex ratios may vary over time, and the lag between entry and each appointment widens with grade seniority. For simplicity, we assume no change in discipline-specific entry-cohort sex ratios over time and consider hypothetical data for all academics in STEMM subjects comprising: academic **grade** (outcome), their **sex** (exposure), and each **discipline entry-cohort sex ratio** (mediator); see the DAG on the right.

The exposure (sex of the individual) precedes entry to any academic discipline and subsequent grades attained; each discipline-specific entry-cohort sex ratio precedes any subsequent grade attained and lies on the exposure-outcome path. Whether or not these relationships are causal must be determined. As cumulative lifecourse experiences differ by sex prior to entry into an academic career, the discipline-specific entry-cohort sex ratio is a proxy for these experiences in the same way as birthweight was for early-life exposures. A causal link between sex and discipline, hence entry-cohort sex ratio, is therefore implicit. In the absence of any sex discrimination, discipline-specific entry-cohort sex ratios should yield similar sex ratios in grade attainment, with proportions of each grade by each sex determined by the discipline; causality is again implicit.

In a linear model, discipline-specific entry-cohort sex ratio is a mediator. According to the BW~BP example, adjusting for discipline-specific entry-cohort sex ratios whilst examining the grade-sex relationship might be suspect. On the other hand, it is compelling to 'adjust' for discipline differences in the workforce sex ratios, as alluded to. To resolve this, we ask: what is the *consequence* we are interested in; and how might we assess it via (hypothetical) *intervention*? The answer to these questions helps frame the research question: *Are appointments to grade subject to sex discrimination?* The process that then takes place to address the research question occurs when the grade is attained, which is after the entry-cohort sex ratio was established. The consequence of interest is ensuring fairness in the *appointment* process. Hence, we need to adjust for the entry-cohort sex ratios because they differ at the time each appointment is made, thereby affecting the denominator of men and women entering the selection process.

The fairness being assessed (upon which one might hypothetically intervene) occurs at the time the outcome (grade) is measured, not when the exposure (sex) is measured, and importantly *after* the mediator (entry-cohort sex ratio) is measured. If we intervene to change establishments prone to sex discrimination, this would be to alter the appointment process, e.g. by ensuring that appointment committees are gender balanced, involving independent observers intervening if any part of the appointment process fails to give equality to all candidates, or other such actions **at the time of appointment**. The critical point is that any intervention necessarily takes place at or just before grades are attained (or not), and therefore *after* the time when discipline-specific entry-cohort sex ratios are established.

According to our DAG, the total causal effect of sex on appointment status in higher education comprises an indirect effect mediated by societal factors that lead to a certain entry-cohort sex ratio, and a direct effect not mediated by any such factors that preceded the application. Any policy change by academic institutions in the hiring process cannot hope to change the indirect effect (e.g. gender balanced committees cannot influence the choice of toys in nursery school). What is targeted is the direct effect: by adjusting for the mediator, we 'block' the indirect effect so only the direct effect remains, which is the relevant effect for our intervention question.

The key to understanding when to adjust for a mediator in a regression model is to ask: *when might an intervention be required that best informs our research question*? If the intervention occurs after the mediator, it is appropriate to adjust. Conversely, for mediators occurring after the intervention it is inappropriate to adjust. By framing research questions in terms of an intervention, it highlights which factors confound the **intervention-outcome relationship** as opposed to the **exposure-outcome relationship**. It is less important that factors considered for being conditioned on are genuine confounders or mediators, as long as all such factors precede the intervention point. This keeps the application and interpretation of conditional linear modelling firmly rooted in a causal framework. It is the need to arrive at causal inference that leads to such rigidity in the ways we think about, and employ, multivariable linear models (DAGs aid this).

We have considered two contexts: one in which adjustment for the mediator was inappropriate because what was to be estimated was the **total effect**; in the other context the desired effect was the **direct effect**, and so adjustment for the mediator was appropriate. The decision as to which effect is to be estimated determines whether to adjust for the mediator. As a rule of thumb, if the exposure variable is also a putative intervention target, it is the total effect that needs to be estimated. This explains why, in biomedical research, adjustment for mediators is uncommon, as the exposure is often a drug or a modifiable risk factor and is thus the target of the intervention.

There are, however, situations where the 'correct' statistical adjustment in a linear model is not obvious, or indeed even tractable, and we look at some of these instances. The first is a problem that plagued the literature with confusion for decades and is an illustration of Simpson's paradox; and the second is an illustration of the challenges with compositional data. We examine both these issues later.

Collider

A collider is a variable that is simultaneously caused by two or more separate causal paths (i.e. they 'collide' into it, as the name suggests). Variables that have causal arcs 'colliding' into another variable may or may not share a causal relationship. If two *independent* variables cause a collider, we do not expect them to carry any association and the presence of a collider does not transmit or give rise to additional associations between them. However, if the collider is inappropriately and/or inadvertently conditioned upon, a spurious **conditional dependency** is created between all variables on the previously unrelated paths. This phenomenon is known as *collider bias* (this is covered in a lot more detail later).

Instrumental variable

An instrumental variable is defined as a variable that is strongly associated with the exposure but has no residual association with the outcome and no association with the confounders of the exposure-outcome relationship.

An example of a variable acting as an 'instrument' can be demonstrated in the Dutch Hunger Winter Families Study – a unique cohort of women born in the Western Netherlands around the period of Nazi occupation. During the occupation, the Dutch people suffered extreme hardship including starvation and malnutrition. Women born during this period thus provided a 'natural experiment' of exposure to extreme calorie restriction and nutrient deficiency at different stages of pregnancy. But the interest was less in exposure to starvation directly, and more in the effect of maternal calorie and nutrient intake during pregnancy. Exposure to the Dutch famine was therefore an '**instrumental variable**' for the true exposure of interest, i.e. maternal calorie/nutrient intake during pregnancy.

Instrumental variables provide a useful opportunity to study an exposure-outcome relationship that is biased by confounding, because of their unique relationship with the exposure and outcome, and exposure-outcome confounders. An instrumental variable may cause the exposure directly (e.g. **famine** causing **low calorie intake**), where it is known as a **causal instrument**, or it may be correlated with a causal instrument (e.g. **tulip consumption** would be correlated with the degree of **famine**), where it is known as a **correlational instrument**. In theory, instrumental variables provide unique insights into the causal effect of an exposure on an outcome because any covariation between the instrument and the exposure is *not* due to confounding. By implication, any covariation between the instrument and the outcome must therefore be due to the causal effect of the exposure on the outcome directly. The causal effect of the exposure on the outcome can thus be estimated by dividing the apparent effect of the **instrumental variable** (**Z**) on the **outcome** (**Y**) by the apparent effect of the instrument on the **exposure** (**X**).

If: $\mathbf{Y} \sim \beta_1\mathbf{Z}$ and $\mathbf{X} \sim \beta_2\mathbf{Z}$
Then: $\mathbf{Y} \sim (\beta_1/\beta_2)*\mathbf{X}$

In **Mendelian randomisation**, a '**polygenic score**' instrumental variable is constructed from multiple genetic variants that appear to satisfy the assumptions of an instrumental variable (i.e. being correlated with the exposure but having no residual correlation with the outcome and unrelated to all confounders).

Assumptions and caveats:

- The main limitation of an instrumental variable analysis is that they are very sensitive to their assumptions, particularly that the instrument has no residual relationship with the outcome and no relationship with any confounders.
- This is especially important where the instrumental variable has a weak relationship with the exposure, such as with many polygenic scores. With weak instruments, modest assumption violations (such as a small residual correlation with the outcome) may be enough to overwhelm the true causal effect. It is therefore recommended that potential instrumental variables be tested and validated in at least two distinct populations before use.

Another limitation of an instrumental variable approach is that they do not guarantee the direction of causality (i.e. that Y precedes X), although this can be explored by studying a separate instrumental variable for the outcome. Unfortunately, even this approach will fail in the presence of time-varying relationships, i.e. where the exposure and outcome cause each other across time.

Statistical adjustment & causal interpretation of multivariable linear models

In causal inference, two variables are special:

- **exposure** (or treatment); and
- **outcome** (or endpoint).

All other variables are **covariates**. As we have seen, covariates have a variety of different roles from a causal inference perspective: they can be **mediators**, **confounders**, **surrogate confounders**, or **competing exposures**. If a suitable subset of covariates can be identified that removes confounding, we may proceed to estimate our causal effect using a multivariable linear model.

In regression models, there are only two types of variables:

- dependent variable (DV) and
- **independent variables** (IVs, or covariates).

No further distinction is made between the IVs – specifically, the exposure is by no means a "special" IV and is treated just like any other covariate. Thus, there is a conceptual mismatch between causal graphical model theory (as depicted by DAGs, which lead us to formulate a multivariable linear model that highlights the exposure-outcome relationship adjusted for confounding) and the standard perception of a regression model. This conceptual mismatch often leads to misinterpretation of the results from a multivariable linear model.

Within observational research, it is important to adjust for confounding to reduce potential biases. Other forms of adjustment may be undertaken, e.g. for competing exposures, which are not true confounders but can improve model precision (though recall: some competing exposures might also double as surrogate confounders). It is clear that adjusting for mediators (variables that lie on the causal path from exposure to outcome) presents a challenge, as this may bias the intended model *inference*.

DAGs can help us examine carefully when and how to make 'appropriate' statistical adjustment for mediators in a linear regression model. To do this, we must recognise three key ingredients to the application and interpretation of multivariable regression models:

- ***causality*** – the framework in which confounding is defined
- ***intervention*** – whether real or hypothetical, as a basis of thinking about what has meaning in relation to the research question that drives interpretation of the model coefficients
- ***context*** – a 'catch-all' for remaining issues, but important for the recognition of extraneous factors that validate or challenge the appropriateness of the methodologies adopted; an example is how we understand the abstract meaning of variables in our DAG (discussed at length earlier for birthweight).

Judea Pearl formulated in 1995 a new calculus for application to causal graph theory coined *do*-calculus.[32] Pearl's calculus facilitates identification of causal effects in non-parametric models as well as proving useful in mediation analysis,[33] transportability,[34] and the recently emergent domain of meta-synthesis (the fusing of empirical results from diverse studies conducted on heterogeneous populations, under different conditions, to synthesize an estimate of a causal relationship in some target environment). We do not consider this calculus in detail but borrow the '*do*' component, i.e. the concept of intervention. When considering the implications of causality in model selection and model interpretation, it helps to think about the role of intervention, either real or hypothetical. Drawing meaningful inference in observational research from a linear model then boils down to identifying the context in which inference has utility. This is best realised by asking: *What is the causal consequence I am interested in?* This helps target an intervention that corresponds to the research question.

Estimating total causal effects

When estimating the total causal effect of an exposure on an outcome, adjusting for all **confounders** is essential to block all confounding paths. When a known or hypothesised confounder is unmeasured or not possible to measure, it is possible to adjust for another variable (i.e. **surrogate confounder**) that sufficiently captures variation in the original confounder, although the adjustment would not be perfect. Adjustment for **mediators** should be avoided because this would block some (or most) of the true causal path and a total causal effect cannot be estimated. Inappropriate adjustment for mediators also risks invoking even more serious problems, such as collider bias, an example of inferential bias phenomenon (more on this topic later). Adjustment for **competing exposures** is optional and may improve precision in the estimates, though extra care should be taken to ensure that such variables are indeed competing exposures, and not, for example, mediators. Similar considerations apply to adjustment for surrogate confounders; while useful for capturing the effects of unmeasured known confounders, they might also act as mediators, in which case adjustment should be avoided.

A summary of covariate roles and general statistical adjustment considerations are available in Table 1.4.1.

Table 1.4.1. A summary of the most common covariate roles, their relationships with the exposure and the outcome, general adjustment considerations, and additional considerations.

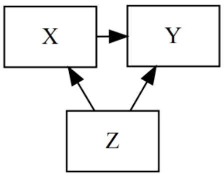
Variable role	Relationship with the exposure	Relationship with the outcome	Adjustment	Additional considerations
Confounder	Causes the exposure	Causes the outcome	Required to block all confounding paths	
Mediator / Collider	Caused by the exposure	Causes the outcome	Should be avoided when estimating the total causal effect and to minimise the risk of collider bias May be justified when the direct causal effect is sought if risk of invoking inferential biases is minimised, or for sensitivity analyses	
Competing exposure	No causal relationship	Causes the outcome	Optional to improve precision in the estimates	Important to ensure that it is not a mediator (i.e. relationship with the exposure overlooked) May also act as a proxy/surrogate confounder
Proxy/surrogate confounder	No causal relationship	Causes the outcome	Possible if the original confounder is not available and is captured sufficiently by the proxy	Important to ensure that it is not a mediator (i.e. relationship with the exposure overlooked)
Instrumental variable	Causes (or highly correlated with) the exposure	No causal relationship or correlation	May be used instead of the exposure in Mendelian randomisation approaches	

Table 2 Fallacy

One particularly widespread misconception is known as **mutual adjustment**, recently called the 'Table 2 fallacy',[35] since the first table in most epidemiological articles usually describes the study data and the second table reports the results of a multivariable regression model in which the erroneous efforts to illustrate mutual adjustment often appear.

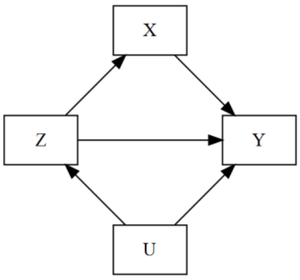
To illustrate the fallacy, let us assume that we wish to estimate the effect of X on Y. We know (e.g. from a DAG) that there is only one confounder, Z, so we run the regression Y~X+Z. If our background knowledge and the statistical assumptions of the regression (e.g. normality) hold, then the coefficient of X estimates the **total causal effect** of X on Y. The 'Table 2 fallacy' is the belief that we can also interpret the coefficient of Z as the effect of Z on Y; indeed, in larger models, the fallacy is the belief that all coefficients have a similar interpretation with respect to Y.

To see why this is not true, look at the DAG that matches our scenario: Z→X→Y & Z→Y (see right). With respect to the X→Y effect, adjustment for Z removes all confounding, but what does including X in the model mean for the effect of Z on Y?



As we can see, X is a mediator of the Z→Y effect, but adjustment for a mediator is erroneous when estimating the total causal effect; the Z coefficient in our model cannot be interpreted as such. Instead, we could interpret it as the 'direct effect' of Z on Y when X is held constant, and this may be stronger than, weaker than, or opposite to the total effect. It would seem, from this example, that we can at least interpret every coefficient as a causal effect: some total and some direct.

To see that this can also fail, let us add another variable to our DAG. We include U, which affects both Z and Y (see right). Despite the addition of this new variable, it is still sufficient to adjust for Z to unconfound the X→Y effect, so the validity of the X coefficient is unchanged – can you see why? Upon examining Z in this situation, however, we encounter difficulties.



The new variable U acts as a confounder of the Z→Y relationship, which means that we would have to interpret the Z coefficient as a 'direct effect that is *confounded* by U' – not exactly a helpful interpretation. Indeed, no single multivariable linear model could ever estimate the causal effects of X and Z at the same time: estimating the effect of X means we *must* include X in the model, but to estimate the effect of Z we *must not* include X.

In general, it is impossible to identify multiple causal effects using a single linear model, and we can usually interpret, at most, just one coefficient in such a model as a **total causal effect**. If we are interested in multiple causal effects, we need multiple (**separate**) regression models.

In the 2nd DAG, we can obtain the effect of X from the model Y~X+Z because adjustment for Z unconfounds the X→Y effect, and we can obtain the effect of Z from the model Y~Z+U because adjustment for U unconfounds the Z→Y effect. The concept of 'mutual adjustment', as often encountered in the literature, is seriously misleading and erroneous.

DAY2

2.1 The Target Trial Framework

Learning outcomes

- Describe the broad aims of the **target trial framework**
- Identify the seven components of the target trial framework and discuss the main features of each
- Describe **prevalent user bias** and **healthy user bias**
- Describe **immortal time bias** and discuss strategies to avoid it

Introduction

Where possible, the most reliable way to estimate a causal effect of an exposure on an outcome is to conduct a randomised controlled trial (RCT). Relatively few exposures in the health and social sciences are however suitable for experimentation, due to practical and ethical limitations. Obesity, for example, cannot practically be assigned to people 'at random'; similarly, it would not be ethical to willingly expose people to harmful substances such as smoking. In these situations, we must attempt to estimate causal effects from observational data, a notoriously difficult task that requires specialist methods such as the potential outcomes framework (covered in Section 1.3) and directed acyclic graphs (covered in Section 1.5). Unfortunately, these methods can be difficult to translate from theory into practice without training. The estimand, estimator, estimate idea – for example – is relatively simple to learn in theory, but it is not always so obvious how to use in a real-world context. These problems – and the use of jargon – can make causal methods off-putting for some applied scientists.

Hernan and Robins attempted to narrow the 'gap' between conventional quantitative training and the use of formal causal framework by introducing the '**target trial framework**'.[36] Built on the language of experimental studies, the approach – which they term an 'organising principle' – encourages researchers to formally describe the '**target trial**' that they would (in theory) use to study their effect of interest, and then explain how they will emulate this trial in observational data. For this reason, the approach is sometimes also known as '**target trial emulation**'.

The Target Trial

The target trial is the hypothetical trial that you would conduct to estimate your causal estimand of interest. The target trial is not a real trial, but a thought experiment. Hernan and Robins however encourage you to make the target trial as close as possible to a plausible trial and even argue: '*if we can't think of a target trial, we probably do not have a causal question*'[37]. The target trial framework therefore encourages you to think of and study a '**well defined intervention**', i.e. something that you could conceivably introduce in practice. This suggests drugs (e.g. metformin), therapies (e.g. cognitive behavioural therapy), protocols (e.g. routine testing), and policies (e.g. stay-at-home orders), rather than broader concepts like 'socio-economic position', 'ethnicity', or 'obesity'.

Two protocols

The target trial approach encourages you to design two protocols, one for the hypothetical target trial, and one for the observational study that aims to emulate that trial. This allows you to explicitly say what you would do 'in the idealised world' where a trial was possible and highlight the areas of compromise in the observational study. The target trial will typically be a pragmatic trial, since many features of a strict randomised trial will not be possible for the causal effect of interest (e.g. blinding of the treatment allocation). The protocols for the target trial and emulation study contain seven components, the **eligibility criteria**, **treatment strategies**, **assignment procedures**, **follow-up period**, **outcome**, **causal contrasts of interest** (estimand), and **analysis plan** (estimator).

Eligibility criteria

The eligibility criteria outlines who would be eligible for recruitment in the trial, and therefore what sample you need to emulate in your observational study. It should be unambiguous, clearly stating the inclusion and exclusion criteria. This step may seem extremely obvious but it can prevent many problems, because you can only recruit based on information known before recruitment! This prevents **'conditioning on the future'** (see Sections 3.1 and 4.3) or other common sampling errors, where we create a misleading sample based on future information. For example, it would not be possible to recruit pregnant women who 'end up having two or more births', yet this is a common sample in observational studies looking at repeat adverse outcomes in pregnancy. Similarly, in longitudinal studies, it is common for analysts to create a sample of people 'with complete information at all time points', which again would not be known at baseline.

In Hernan and Robin's example, the eligibility is described as: *"Postmenopausal women within 5 years of menopause between the years 2005 and 2010 and with no history of cancer and no use of hormone therapy in the past 2 years."*

Treatment strategies

The treatment strategies outline all possible exposure regimes. In an observational study, it is inconceivable that you will ever be comparing a treatment to a placebo; and you won't have trial managers monitoring and enforcing people's adherence. This shows how the target trial will generally have to be a pragmatic trial, comparing treatment to usual care, different levels of treatment, or similar. It is not sufficient to simply compare the 'exposed' vs the 'unexposed' because: a) this is not philosophically equivalent to a trial (in which treatment is *initiated*); and b) it risks '**prevalent user bias'** (or '**healthy user bias**'), a form of selection bias that arises because units that are already exposed are likely to be structurally different from those that are not exposed.

In Hernan and Robin's example, the two treatment strategies are: 1) *"Refrain from taking hormone therapy during the follow-up"* (reference group); and 2) *"Initiate estrogen plus progestin hormone therapy at baseline and remain on it during the follow-up, unless diagnosed with deep vein thrombosis, pulmonary embolism, myocardial infarction, or cancer"* (treatment group).

Assignment procedures

The assignment procedure describe our efforts to produce exchangeable units of analysis; in a simple target trial this will almost always be random assignment at baseline. It is very rare – though not inconceivable – that a person will be unaware of their exposure status in an observational setting, therefore the assignment procedure for the target trial will usually not be blind. Since the observational study is likely interested in the real world effect, it is arguably not a weakness to study the unblinded effect as this will likely be most similar to what you will observe in practice.

Emulating random assignment in the observational study requires an appropriate study design or analysis technique, and will usually involve measuring and controlling for relevant confounders. Hernan and Robin's recommend several sensitivity analyses to check for unobserved and residual confounding, including reversing the treatment strategies; so instead of looking at the effect of treatment initiation in a treatment-naïve sample, look at the effect of discontinuing treatment in a treatment-exposed sample. Alternatively, consider studying the effect of the exposure on one or more control outcomes, i.e. outcomes that you would NOT expect there to be an effect. Similarly, consider studying additional treatment controls, i.e. treatment strategies with similar indications to the primary exposure, but which are not expected to have a (large) causal effect.

In Hernan and Robin's example, the assignment procedures are: *"Participants will be randomly assigned to either strategy at baseline and will be aware of the strategy to which they have been assigned".*

Follow-up period

The follow-up period section describes the start and end of the time that all participants will be followed. Care should be taken to consider the time that follow-up starts to minimise **immortal time bias**, discussed below, which occurs when the start of follow-up is aligned with either the eligibility and/or treatment assignment.

In Hernan and Robin's example, the follow-up period is: *"Starts at randomization and ends at diagnosis of breast cancer, death, loss to follow-up, or 5 years after baseline, whichever occurs first".*

Outcome

The outcome section describes our effort to monitor and identify the outcome. Where possible, in a randomised controlled trial, the outcome is usually ascertained systematically without knowledge of the treatment status (to avoid '**diagnostic bias**'). In most observational studies, this will not be possible, since the participant and diagnostician will usually be very aware of the participants exposure status. The most important consideration here is therefore about the completeness of ascertainment; so with a preference towards objective ascertainment, multiple ascertainment, and/or validation. The biggest problems will occur wherever certain values of the exposure may differentially encourage additional screening or testing (for example, women receiving hormone replacement therapy are more likely to have blood spotting, which is a symptom of endometrial cancer, leading to a higher chance of referral and diagnosis).

In Hernan and Robin's example, the outcome is: *"Breast cancer diagnosed by an oncologist within 5 years of baseline".*

Causal contrasts of interest

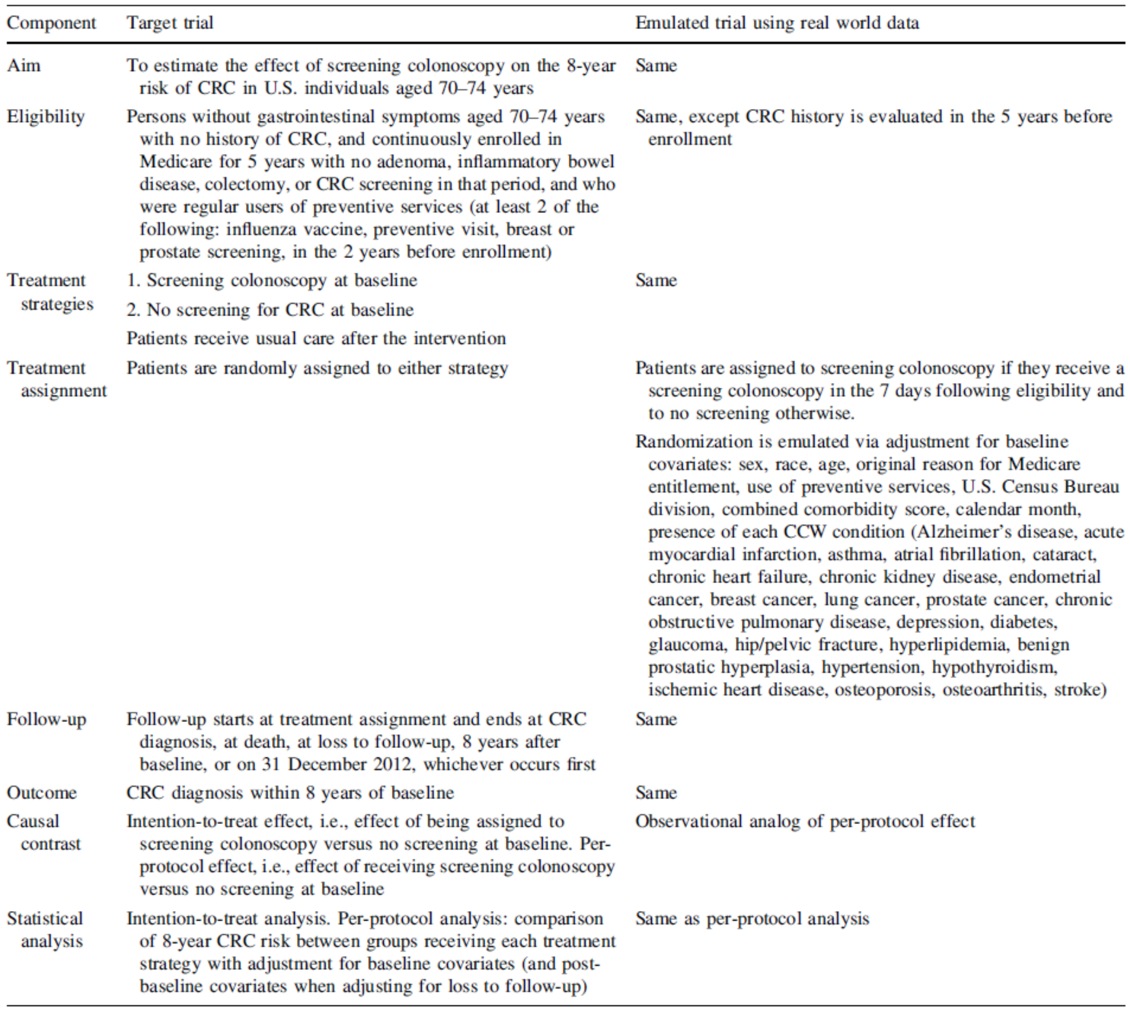
This section describes the estimands. In the target trial, these are likely to be the **intention-to-treat effect**, the **per-protocol effect**, or both. The intention-to-treat effect is the total causal effect of being allocated to a specific treatment strategy, regardless of whether it is then followed. The per-protocol effect is the total causal effect of being allocated to a specific treatment strategy *and* adhering to that strategy for the duration of the treatment period. It is probably preferable to attempt to estimate both, but in observational data it may not be possible to estimate the intention-to-treat effect, because it may be impossible to distinguish between someone who goes against one treatment protocol, and someone who was 'assigned' a different treatment strategy.

Analysis plan

This final section describes the planned analytical approach, although only in summary detail. Where an intention-to-treat analysis is possible, adjustment will be necessary for baseline confounders. Where a per-protocol analysis is attempted, adjustment will be necessary for baseline confounders and time-varying confounders (using **g-methods** – see Section 3.3). Inverse probability weighting may also be necessary in the presence of differential attrition.

Example

Most implementations of the target trial framework present the protocol for both the target trial and emulation study in a table like the following (from Garcıa-Albeniz et al 2017[38]):



Component	Target trial	Emulated trial using real world data
Aim	To estimate the effect of screening colonoscopy on the 8-year risk of CRC in U.S. individuals aged 70–74 years	Same
Eligibility	Persons without gastrointestinal symptoms aged 70–74 years with no history of CRC, and continuously enrolled in Medicare for 5 years with no adenoma, inflammatory bowel disease, colectomy, or CRC screening in that period, and who were regular users of preventive services (at least 2 of the following: influenza vaccine, preventive visit, breast or prostate screening, in the 2 years before enrollment)	Same, except CRC history is evaluated in the 5 years before enrollment
Treatment strategies	1. Screening colonoscopy at baseline 2. No screening for CRC at baseline Patients receive usual care after the intervention	Same
Treatment assignment	Patients are randomly assigned to either strategy	Patients are assigned to screening colonoscopy if they receive a screening colonoscopy in the 7 days following eligibility and to no screening otherwise. Randomization is emulated via adjustment for baseline covariates: sex, race, age, original reason for Medicare entitlement, use of preventive services, U.S. Census Bureau division, combined comorbidity score, calendar month, presence of each CCW condition (Alzheimer's disease, acute myocardial infarction, asthma, atrial fibrillation, cataract, chronic heart failure, chronic kidney disease, endometrial cancer, breast cancer, lung cancer, prostate cancer, chronic obstructive pulmonary disease, depression, diabetes, glaucoma, hip/pelvic fracture, hyperlipidemia, benign prostatic hyperplasia, hypertension, hypothyroidism, ischemic heart disease, osteoporosis, osteoarthritis, stroke)
Follow-up	Follow-up starts at treatment assignment and ends at CRC diagnosis, at death, at loss to follow-up, 8 years after baseline, or on 31 December 2012, whichever occurs first	Same
Outcome	CRC diagnosis within 8 years of baseline	Same
Causal contrast	Intention-to-treat effect, i.e., effect of being assigned to screening colonoscopy versus no screening at baseline. Per-protocol effect, i.e., effect of receiving screening colonoscopy versus no screening at baseline	Observational analog of per-protocol effect
Statistical analysis	Intention-to-treat analysis. Per-protocol analysis: comparison of 8-year CRC risk between groups receiving each treatment strategy with adjustment for baseline covariates (and post-baseline covariates when adjusting for loss to follow-up)	Same as per-protocol analysis

Time zero and immortal time bias

One of the most important benefits of a target trial approach is avoiding **immortal time bias**. Immortal time bias occurs in observational studies when there is a time delay between eligibility, treatment assignment, and the analytical time zero. This occurs most commonly when the exposure either implicitly or explicitly contains time (e.g. '7-days of steroid therapy' or 'attended at least three sessions of chemotherapy'). Here, anyone who dies between the initiation and completion of treatment cannot be included in the sample, making this person 'immortal'. In comparison to any exposure with a different time requirement, the survival will therefore appear higher. The best way to avoid immortal time bias is to avoid time gaps between eligibility, treatment assignment, and analytical time zero.

Example 1: Bloodletting for Yellow Fever

In 1793, Benajamin Rush claimed to have discovered that being bled 7-times was a cure to Yellow Fever because none of the patients that he had bled more than 7-times had ever died. Any patient that didn't survive long enough to be bled seven times, of course, was considered further proof that 7-bleeds were required to keep the patient alive!

Example 2: Statins and risk of CVD in people with diabetes

Yee *et al.* 2004 reported that 'statin use' among adults with type 2 diabetes was associated with a 10-month delay in needing to start insulin treatment.[39] 'Statin use' was however defined as 'at least 1 year of use'. Any statin user with <1 year treatment who required insulin treatment was therefore misclassified as being a non-user. When the same dataset was correctly analysed, the effect reversed.[40]

Example 3: Time between the first and second dose and COVID vaccine efficacy

The longer the time between the first and second dose of your vaccine, the higher the chance of contracting COVID before you complete the vaccination regimen, and the higher the chance you fail to reach eligibility.

For the same length of follow-up, people in the sample with a longer gap therefore have a longer period of 'invulnerability' to COVID; and a shorter length of vulnerability. If analysing only the vaccination group, this would cause a huge bias. However, since vaccine efficacy explicitly involves comparing treated to placebo groups, the invulnerability should scale equally in the treatment and placebo groups, resulting in no overall bias.[41]

Analytical solutions to immortal time bias

When it is not possible to avoid some time gap between the eligibility, treatment assignment, and/or allocation of time zero, immortal time bias can also be avoided with analytical solutions. Several different strategies are suggested by Zhou *et al.*[42] The simplest is to discard all participants in the untreated group who fail to reach the same duration of time as the treated group; e.g. if statin-users are those requiring 1-year of statin treatment, then all non-users with less than 1-year of follow-up are also discarded. This approach avoids the problem of immortal time bias but is statistically inefficient because you are discarding important observations. Alternatively, a time-dependent analysis can be used, in which participants switch from untreated to treated groups as their eligibility changes.

Finally, a cloning approach may be used, where all participants are duplicated at baseline and entered into both the treated and untreated group (since their status is 'unknown' until the minimum time until treatment assignment has passed). Once the pair becomes compatible with a particular treatment strategy, the mismatching clone is censored. Further details of this approach are provided in a tutorial by Miguel Hernan[43].

2.2 Drawing and Evaluating DAGs for Applied Research

Learning objectives

- Identify and describe each component of a DAG
- Be able to systematically approach the drawing/specification of a DAG

Definition and terminology

Within causal inference, DAGs are graphical, nonparametric representations of hypothesised causal relationships between measured ('observed') and unmeasured ('unobserved' or 'latent') variables. Current convention is to represent measured variables as squares or rectangles and unmeasured variables as circles or ellipses, although this is not universally applied. These representations of variables are termed 'nodes' (or 'vertices'), and the causal paths between variables are represented by unidirectional arrows termed 'directed arcs' (or 'directed edges').

Three key characteristics of DAGs are that:

- Causal paths between variables **must be** unidirectional (i.e. each of the variables connected by a causal path can only operate as either cause or effect, and not both);
- A variable **must not** cause either itself or one of its own causes (i.e. there should be no cyclical paths, hence the name 'directed **acyclic** graph'); and
- While a direct path between two variables only indicates the **possibility** that these variables are causally related (even if only to a modest extent), the absence of a direct path between two variables reflects the absence of any such causal relationship (i.e. greater **certainty** and importance is afforded the absence of a causal path than the presence of one).

Epidemiological utility – past, present and future

We have seen how DAGs have substantial utility for displaying – and supporting robust analyses of – hypothesised causal relationships. DAGs facilitate what might be termed a 'causal gaze' – a perspective from which complex (causal) processes can be simplified, characterised in graphical form and then examined, disentangled, debated and resolved using an established framework of rules (including the three key characteristics listed above).

DAGs also facilitate the identification of variables operating in very specific ways within any hypothesised causal system, each of which requires careful attention when designing statistical models to generate causal inference. As described in the preceding session, these include:

- The specified (or 'main') '**exposure**' (the putative cause within the '**focal relationship**' under examination);
- The specified '**outcome**' (the putative effect / consequence within the '**focal relationship**' under examination)
- '**Confounders**' (covariates relating to events, processes, items or characteristics which, as specified, come into being/occur **before** *both* the exposure *and* the outcome, and are therefore potential causes of both);
- '**Mediators**' (covariates relating to events, processes, items or characteristics which, as specified, come into being/occur **after** the specified exposure but **before** the specified outcome, and are therefore potential consequences of the specified exposure and potential causes of the specified outcome); and
- '**Competing exposures**' (covariates that are causally **unrelated** to the specified exposure, but which precede, and are therefore **potential** causes of, the specified outcome).

By identifying variables operating in these ways within the hypothesised causal system, DAGs have extensive utility in statistical modelling for causal inference by ensuring that models:

- Identify, and adjust for, those covariates specified as potential confounders;
- Do **not** (ordinarily – see earlier) adjust for covariates specified as likely mediators (since the adjustment for such variables can create bias due to the 'reversal paradox'[26]); and
- Can identify and adjust for covariates specified as competing exposures, wherever such adjustment strengthens the value or interpretation of the model produced.

The application of '**graphical model theory**' to DAGs[2] can further enhance adjustment for confounding by identifying any alternative '**minimally sufficient adjustment sets**' of covariates specified as potential confounders.[3] This can be of great practical value in those circumstances where: not all the specified potential confounders have been measured; or not all of the specified potential confounders can be measured with reasonable accuracy and precision (or within the resources available).

Beyond these 'early benefits' of DAGs (i.e. improving the transparency of *a priori* hypotheses; reducing inappropriate adjustment for mediators; and enhancing the selection of confounders for adjustment), DAGs also have substantial potential utility for: identifying and estimating the extent of unobserved confounding (where the DAGs involved permit this); evaluating whether any given DAG (as specified) is consistent with the observed dataset(s) it was intended to represent;[4] and elucidating invalid or inappropriate analyses.

Conceptualising variables and contextualising cause

Although DAGs can sometimes offer simple representations of what might otherwise be complex causal processes, many can be challenging to draw (or, rather, to 'specify'), not least when:

- The variables involved represent poorly defined and/or understood concepts/constructs;
- The variables, though measured at one point in time, reflect events, items, processes or characteristics that came into being/occurred at previous points in time; and
- The causal processes the DAG is intended to reflect are influenced by the context(s) in which these occur.

Hypothesising the potential causal relationships between each of the constituent variables (be they manifest or latent) requires that we not only recognise precisely what each variable represents (be that an event, an item, a process, or a characteristic), but also that we have substantial understanding of each potential causal relationship based upon clear theoretical principles and/or robust, *external* empirical evidence (where, by *external* we mean from data external to the study in question). This can be extremely challenging, especially in hypothesised causal systems where there is incomplete understanding, limited robust *external* empirical evidence, or where the theoretical principles involved are unclear, uncertain or contested. Nonetheless, even under these circumstances, 'temporality' (i.e. the simple rule that the past precedes the present) can often provide a sufficient theoretical basis upon which DAG specification can proceed, providing it is possible to identify the temporal sequence of the variables involved. Thereafter, there is no reason why alternative DAGs (particularly, and preferably, when specified *a priori*) might be specified that reflect specific uncertainties and therefore guide causal inference analyses using DAG-informed sensitivity analyses.

Determining the temporal sequence of variables within a DAG requires establishing the temporal relationship of measurements operationalised as nodes that are fixed in time either: (a) by nature of the variable concerned (i.e. where the variable is '**time-invariant**' and varies only across subjects/participants and not over time; e.g. sex or place of birth); or (b) by the specific point in time at which the variable concerned was measured (i.e. where the variable is '**time-variant**' and varies not only across subjects/participants but also over time; e.g. body mass or food intake). Importantly, every measurement of a time-variant variable captures not only the value prevailing at the point of measurement, but also the cumulative 'experience' of that variable over the time preceding measurement (such that the measurement made might be considered to represent a value that has 'crystallised' at, or up until, that point in time).

The precise time at which a time-variant variable (and the concept/construct this represents) is 'crystallised' is crucial for considering where it should be placed in the temporal sequence of nodes that form a causal

DAG. This is because temporality is key to establishing which variables (as manifestations in time of the 'crystallised' properties they reflect) can plausibly act as **potential** causes of other variables (given that only past nodes can cause subsequent nodes). Indeed, the very notion of time-variant variables – which may reflect properties from **either the present or the past (or both)**, that have accumulated over time – can make them especially difficult to position within a DAG (both conceptually and functionally). A simple example of such a variable might be body height which might be considered a time-variant variable when measured during childhood, but which might appear time-invariant when measured in adulthood (having crystallised at the end of adolescence, thereafter remaining the same until the decline in height commonly accompanying senescence later in life).

The causal relationships between variables (whether time-variant or time-invariant) may also change between contexts, such that a valid causal relationship in one context may be reversed in a second, or entirely impossible/implausible (and therefore absent) in a third. Drawing DAGs therefore requires not only careful thinking about the meaning of all of the constituent variables, but also how these are likely to be ordered, in time, within the specific context being modelled – a context that extends not only to the specific historical, social and physical environment concerned, but also to the very different 'analytical contexts' that exist for different study designs, sampling strategies, and data acquisition processes.

Understanding any given variable, what this purportedly measures, and what this means in any given context, is therefore both challenging *and* critical to correctly specifying DAGs that are capable of informing robust statistical models of hypothesised causal relationships. There may be instances in which the level of ambiguity or a lack of knowledge and understanding means there is little confidence to support accurate specification of even the most hypothetical DAG. Yet the impossibility of knowing, *a priori*, everything necessary about the processes involved in any causal system does not mean that the resultant DAG (specified *in the absence of definitive evidence*) has little to offer to strengthen our confidence in causal inference modelling. This is because, while challenging to specify and impossible to perfect, DAGs nonetheless make the process of causal estimation far more transparent to both the analysts concerned and to others. By helping analysts to identify (and hence avoid) some of the more obvious (and sometimes less obvious) errors that influence the analysis of observational data for causal inference, even 'uncertain' DAGs can help improve the analysis of causal inference.

Drawing DAGs in four simple steps using temporal logic

Notwithstanding the conceptual and operational issues considered above, there are four simple rules (based on the unassailable 'temporal logic' that the past precedes the present) that can help to improve the drawing/specification of DAGs to represent hypothesised causal processes.

- *All nodes should initially be considered as potentially 'time-variant' measures of the variable they represent:* this ensures that the properties attributed to **measured** variables include those that may have crystallised prior to the time at which the variable was measured.
- *Simultaneously crystallising variables are likely to share common causes (whether latent or manifest):* this allows for any such 'contemporaneously crystallising' nodes to be **correlated** without being specific about any direct causal links, nor having to specify the direction of any such cause (were this to be present).
- *Only preceding nodes act as causes of subsequent nodes:* this requires nodes acting as causes to have properties that crystallised **before** those of any nodes they affect.
- *Temporality confers the* ***potential*** *for causality:* this means that causal paths (i.e. arcs or edges) should only be missing within a DAG where these: do not follow temporal logic; or where there is robust, *external* empirical evidence that the given causal path does not exist.

These four rules can be translated into a series of tasks that greatly facilitate the specification of DAGs based on all constituent variables (whether observed or unobserved) that are thought to be relevant to the focal relationship under examination:

First, determine **when** each observed variable (regardless of when measured) was likely to have 'crystallised'; then specify **when** each unobserved variable is considered (theoretically) to have crystallised; and arrange both sets of (observed and unobserved) variables in a temporal sequence, allowing for groups of variables that crystallised at the **same** point in time to be situated contemporaneously;

Second, for each group of contemporaneously crystallised/situated variables, add a new latent (i.e. unobserved) variable operating as a common cause temporally situated (for convenience) **immediately** preceding the contemporaneously crystallised / situated group of variables.

Third, add directed arrows from **all** preceding variables to **any** subsequent variable(s), ensuring there are no missing arrows from any preceding variable to any subsequent variable.

The first three steps generate what is termed a '**forward saturated DAG**' (meaning that it includes all possible causal paths between preceding and subsequent variables). When drawn in a straight line (e.g. from left to right, from past to present), with variables arranged in the order in which these crystallised, and with causal paths delineated using curved lines, such DAGs often take on the appearance of an 'onion' (hence the colloquial term '**onion DAG**').[5]

Importantly, a **fourth step** may be required when there is sufficient evidence to warrant excluding a directed arrow between a preceding and subsequent variable, thus:

Fourth, remove **only** those directed arrows between variables where these do not follow temporal logic (this should not occur if the third step, above, has been correctly implemented) *or* where there is sound knowledge or robust, *external* empirical evidence that the given causal path does **not** exist.

Finally, an optional **fifth step** might involve introducing additional, unlabelled latent variables in between each of the variables as arranged (i.e. such that each of the measured/acknowledged manifest and latent variables originally included are now preceded by an additional unlabelled latent variable). This step ensures the DAG acknowledges the likelihood of latent confounders and mediators influencing the causal pathways specified and can prompt analysts to probe what these latent confounders might be and what importance they might play in possible under-adjustment for confounding.

Summary

After highlighting the early benefits of DAGs (in facilitating the conceptualisation and dissemination of causal systems and processes, and helping to reduce a range of common flaws and errors in the modelling of causal systems), we also examine several implicit and explicit conceptual and contextual challenges to drawing (or 'specifying') DAGs. These challenges relate to both: the causal meaning of what constitutes a 'variable' (and the 'nodes' used to represent these as markers of past or present events, items, processes, or characteristics); and the important role that context plays in determining what such variables mean, and how they are conceptualised and operationally specified. The four key rules outlined, based on temporal logic, can be applied using four-five simple steps to draw/specify DAGs consistently, thereby improving intra- and inter-analyst reliability and reducing the potential for error.

2.3 Wright's Path Rules & Parametric Considerations

Learning objectives

- Learn about Wright's path rules and understand how these underpin Structural Equation Models (SEMs), Structural Causal Models (SCMs), and parametrised DAGs
- Appreciate how to evaluate DAG-data consistency and avoid data-driven model building
- Understand how to simulate data that respects the underlying data generating mechanism (DMG) using a simplified parameterised DAG
- Understand key differences between DAG-informed simulation of the DGM directly and covariance-matrix simulation of the consequences of the DMG

Informing simulations from a causal perspective

Simulation studies are useful when evaluating the performance, adequacy, and properties of both current and novel statistical methods under a wide variety of settings.[44,45] A poorly designed simulation study may lead to poorly simulated data, which may then potentially affect the conclusions drawn from the statistical models being investigated. The validity of a simulation depends upon how well it reflects the underlying data generation mechanism (DGM), not merely the consequences of it. A suitable DGM that accurately describes the causal relationships amongst all variables will improve the generalisation of the simulation and hence the generalisation of results obtained from the many evaluations undertaken on the simulated datasets. A well-structured and carefully considered DGM guides a more robust interpretation of the results.

Data simulation informed by a directed acyclic graph (DAG) is fundamentally different from simulation using a covariance (correlation) structure. A key principle is the distinction of respecting the DGM directly, as it unfolds longitudinally, opposed to the more common yet naïve approach of capturing only its *consequences* as seen in the observed covariance structure at specific (i.e. cross-sectional) time points.

This is an important distinction because the resultant (cross-sectional) covariance structure cannot yield any intrinsic latent features that occurred during the temporal process. By aligning simulations to the temporal causal processes inferred, the potential data structures generated are mapped reliably onto the potential solution space that reflects the mechanism(s) behind observed data. To illustrate, we revisit briefly the origins of Structural Equation Models (SEMs)[46], the thinking that came later to Structural Causal Models (SCMs)[47] and DAGs[13], and introduce Wright's path rules.

Wright's path diagrams and his 'path tracing' rules

Wright developed the very first **path models** (precursors to SEMs, SCMs, and DAGs). His path models had path coefficients that were estimated on the basis of the correlation of both observed and latent variables – a pioneering idea at the time.[48,49] Wright used path coefficients to show how bivariate correlations among variables can causally quantify relationships for these variables using a system of linear equations, unpicking the causal structure.

<u>Note</u>: Deriving causal structure from data can be achieved in only limited circumstances and involves *a priori* knowledge (i.e. external knowledge of the underlying causal structure), as was the case for Wright's problem.

<u>Note</u>: Wright's path models (initially) made very simplistic parametric assumptions of **multivariate normality** and **no interactions** amongst the variables considered – whilst extension beyond this now exist, these are not discussed here, since it is the foundations of Wright's thinking that is of intrinsic value.

The oversimplifications made by Wright's approach to the development, modelling, and interpretation of his path diagrams may seem to be of limited use for real-world complex scenarios, but the insights they provide from 'toy' perspectives are often astounding and extremely valuable. Wright's achievement is therefore important and should be understood as useful for investigating simplifications of the complex real world because, through simulation, important insights may be gained – as will be shown in this course.

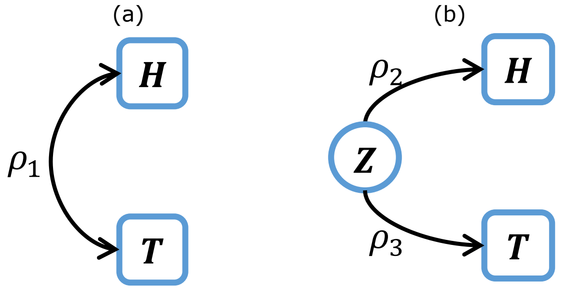
We use the ‘toy’ simplicity (though very real problem) from Wright’s original work to show how links between bivariate correlations of pairs of variables and model parameters in a system of linear equations might estimate *causal effects* of one variable on another.[46]

Wright proposed a set of rules for examining a linear, multivariate normal path diagram that generates a system of equations to describe relationships among all variables,[48,49] where the Pearson correlation between any two variables in the path diagram is expressed as a contribution of all possible paths between them, both direct and indirect. The numerical contribution of an indirect path is the product of the path coefficients for each constituent arrow along the route (**note**: this does not hold with interactions or nonlinearity).

For any compound path:

- **Loops are not allowed**. This means that one cannot pass through the same variable twice when following a particular route.
- **No going forward and then backward**. This means that if one goes forward on a particular route or path, one cannot go backward to the variable(s) along the same or alternative backward route.
- **A maximum of one bidirectional arrow is allowed for each path**. This allows for correlations, i.e. causal flow that is not explicitly specified; but curved arrows are allowed only once for each path.

Figure 2.3.1: (a) A path diagram depicting correlation between variable $\boldsymbol{H}$ and variable $\boldsymbol{T}$; and (b) a path diagram depicting a causal path from variable $\boldsymbol{Z}$ to variable $\boldsymbol{H}$ and a causal path from variable $\boldsymbol{Z}$ to variable $\boldsymbol{T}$; the bivariate correlation between $\boldsymbol{H}$ and $\boldsymbol{T}$ is $\rho_1 = \rho_2\rho_3$.



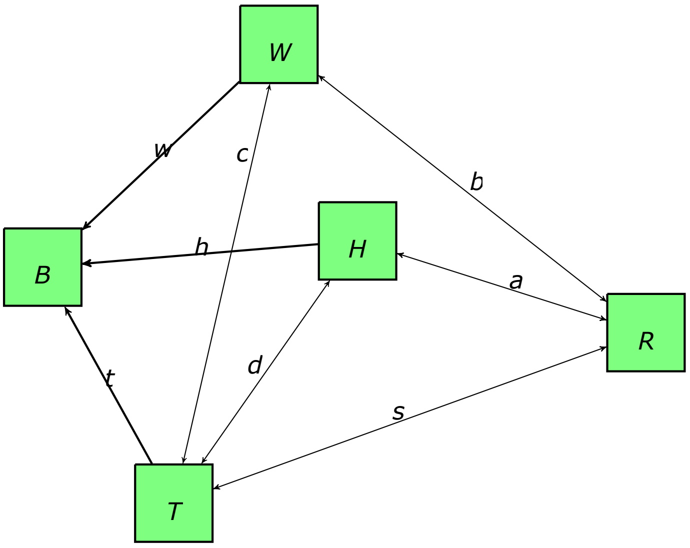
The last rule violates the construction of a DAG but is overcome by introducing a latent variable as a common ancestor of the two variables. The path coefficient of the bidirectional arc is the product of the two unidirectional arcs (Figure 2.3.1).

To illustrate, we go through Wright’s example when investigating causal factors that determine wet bulb depression ($\boldsymbol{B}$), i.e. the difference between dry-bulb and wet-bulb temperature of a thermometer (Figure 2.3.2).

If there is 100% humidity, dry-bulb and wet-bulb temperatures are identical.

Factors Wright considered were: temperature $\boldsymbol{T}$, absolute humidity $\boldsymbol{H}$, and wind velocity $\boldsymbol{W}$. He introduced radiation $\boldsymbol{R}$ as another factor correlated with all causal factors. Wind velocity was assumed to be correlated with temperature and radiation. Let $\beta_{\boldsymbol{BT}} = t$ be the path coefficient measuring the relative influence of temperature on wet-bulb depression, $\beta_{\boldsymbol{BH}} = h$ be the path coefficient measuring the relative influence of humidity on wet-bulb depression, and $\beta_{\boldsymbol{BW}} = w$ be the be the path coefficient measuring the relative influence of wind velocity on wet-bulb depression. Assuming that $\{c, d, a, b, s\}$ are bivariate correlations between $\boldsymbol{W}$ and $\boldsymbol{T}$, $\boldsymbol{T}$ and $\boldsymbol{H}$, $\boldsymbol{H}$ and $\boldsymbol{R}$, $\boldsymbol{R}$ and $\boldsymbol{W}$, $\boldsymbol{R}$ and $\boldsymbol{T}$, respectively, we use Wright’s rules to find the bivariate correlations: $\rho_{\boldsymbol{BW}}$, $\rho_{\boldsymbol{BR}}$, $\rho_{\boldsymbol{BT}}$, $\rho_{\boldsymbol{WR}}$, $\rho_{\boldsymbol{WT}}$, $\rho_{\boldsymbol{RT}}$.

Figure 2.3.2: A path diagram depicting causal relations between wet-bulb depression ($\boldsymbol{B}$), wind velocity ($\boldsymbol{W}$), radiation ($\boldsymbol{R}$), and temperature ($\boldsymbol{T}$) taken from (Wright, 1921).



Links between (causal) path coefficients and bivariate correlations then follow.

- The correlation between $\boldsymbol{B}$ and $\boldsymbol{W}$: There is a direct effect from $\boldsymbol{W}$ to $\boldsymbol{B}$ represented by w. There is an indirect path from $\boldsymbol{W}$ to $\boldsymbol{B}$ through $\boldsymbol{T}$, which can be represented by tc. Hence, the total correlation between $\boldsymbol{B}$ and $\boldsymbol{W}$ is given by $w + tc$.
- The correlation between $\boldsymbol{B}$ and $\boldsymbol{R}$: There is no direct path (effect) from $\boldsymbol{R}$ to $\boldsymbol{B}$. There are three indirect paths from $\boldsymbol{R}$ to $\boldsymbol{B}$. The first indirect path goes through $\boldsymbol{T}$, which can be represented by ts. The second indirect path goers through $\boldsymbol{W}$, which can be represented by bw. The third indirect path goes through $\boldsymbol{H}$, which can be represented by ah. Hence, the total correlation between $\boldsymbol{B}$ and $\boldsymbol{R}$ is given by $ts + bw + ah$.
- The correlation between $\boldsymbol{B}$ and $\boldsymbol{T}$: There is a direct effect from $\boldsymbol{T}$ to $\boldsymbol{B}$ represented by t. There are two indirect paths from $\boldsymbol{T}$ to $\boldsymbol{B}$. The first indirect path goes through $\boldsymbol{H}$, which can be represented by dh. The second indirect goes through $\boldsymbol{W}$. Hence, the total correlation between $\boldsymbol{B}$ and $\boldsymbol{T}$ is given by $t + dh + wc$.
- The correlation between $\boldsymbol{B}$ and $\boldsymbol{H}$: There is a direct effect from $\boldsymbol{H}$ to $\boldsymbol{B}$ represented by h. There is one indirect path from $\boldsymbol{H}$ to $\boldsymbol{B}$, through $\boldsymbol{T}$, which can be represented by dt. Hence, the total correlation between $\boldsymbol{B}$ and $\boldsymbol{H}$ is given by $h + dt$.
- The correlation between $\boldsymbol{W}$ and $\boldsymbol{H}$: There is no direct effect from $\boldsymbol{H}$ to $\boldsymbol{W}$ and there is indirect path from $\boldsymbol{H}$ to $\boldsymbol{W}$. Hence, the total correlation between $\boldsymbol{W}$ and $\boldsymbol{H}$ is zero.
- The correlation between $\boldsymbol{W}$ and $\boldsymbol{T}$: There is a direct effect from $\boldsymbol{T}$ to $\boldsymbol{W}$ represented by c and no indirect routes. Hence, the total correlation between $\boldsymbol{W}$ and $\boldsymbol{T}$ is given by c.
- The correlation between $\boldsymbol{W}$ and $\boldsymbol{R}$: There is a direct effect from $\boldsymbol{R}$ to $\boldsymbol{W}$ represented by b and no indirect routes. Hence, the total correlation between $\boldsymbol{W}$ and $\boldsymbol{R}$ is given by b.
- The correlation between $\boldsymbol{R}$ and $\boldsymbol{T}$: There is a direct effect from $\boldsymbol{R}$ to $\boldsymbol{T}$ represented by s and no indirect routes. Hence, the total correlation between $\boldsymbol{R}$ and $\boldsymbol{T}$ is given by s.

The result is a map of the (causal) path coefficients in the path diagram onto the bivariate correlation matrix derived for all variables. We list directly assigned and indirectly calculated causal paths and corresponding bivariate (Pearson) correlations in Table 2.3.1.

Table 2.3.1: Comparisons of the direct or indirect total path coefficients between pairs of variables in Figure 2.3.1 and their corresponding bivariate (Pearson) correlation.

Causal path coefficient	Bivariate (Pearson) correlation
$\boldsymbol{W} \rightarrow \boldsymbol{B} = w$	$w + tc$
$\boldsymbol{H} \rightarrow \boldsymbol{B} = h$	$h + dt$
$\boldsymbol{T} \rightarrow \boldsymbol{B} = t$	$t + dc + wc$
$\boldsymbol{R} \rightarrow \boldsymbol{B} = ah$	$ts + bw + ah$
$\boldsymbol{H} \rightarrow \boldsymbol{W} = 0$	0
$\boldsymbol{H} \rightarrow \boldsymbol{T} = unspecified$	d
$\boldsymbol{W} \rightarrow \boldsymbol{T} = unspecified$	c
$\boldsymbol{R} \rightarrow \boldsymbol{T} = unspecified$	s
$\boldsymbol{R} \rightarrow \boldsymbol{W} = unspecified$	b
$\boldsymbol{R} \rightarrow \boldsymbol{H} = unspecified$	a

It can be seen that the causal relationship between two variables is not always the same as their bivariate correlation, confirming what is widely known, i.e. $correlation \neq causation$. However, perhaps less obvious and less widely appreciated, this also demonstrates how $causation \neq correlation$, i.e. it is feasible to have a non-zero causal link between two variables that exhibit a zero (or near zero) correlation, due to underlying causal structure among a set of variables.

It is interesting to note that, as Wright depicted his assumed causal structure, his path diagram implied the very strong assumption that $\boldsymbol{W}$ and $\boldsymbol{H}$ have no causal link (within the population), which means that, on average across multiple samples, they should exhibit a bivariate correlation of zero. This also implies that $\boldsymbol{W}$ and $\boldsymbol{H}$ have no common unobserved causal ancestors – given that one is wind velocity and the other is humidity, this might seem untenable. This could be remedied by adding a bidirectional arc between $\boldsymbol{W}$ and $\boldsymbol{H}$, but we do not consider this. It is just important to recognise how Wright's path diagram explicitly states his assumptions, even if we might not agree with them. The true underlying DGM is not known, but we can now evaluate consistency between the observed data and our presumed path diagram/DAG (see later).

Structural Equation Modelling (SEM)

Although it is possible to extend Wright's path diagrams to multivariate non-normal distributions (in any combination), interactions remain challenging and require more sophisticated approaches, such as causal mediation analysis (CMA)[50] and four-way decomposition[51]. Path diagrams may be viewed as the first step in building an SEM. However, going straight to an SEM makes several parametric assumptions that probably warrants careful scrutiny, i.e. assumptions of linearity, multivariate normality (or related distributional properties), and no interactions all need verification in practice – this, sadly, is rarely undertaken. In addition, most SEMs are devised from the premise of *adding* arcs between pairs of variables, opposed to starting with a 'saturated' path diagram/DAG and justifying the *removal* of arcs.

Structural Causal Models (SCMs)

The ideas relating to structural causal models (SCMs) were proposed by Pearl in the mid-1990's,[13,52,53] when he introduced nonparametric causal diagrams for identifying causal effects from observational data. SCMs are defined as mathematical models that are used to represent causal relationships between variables, represented in the form of a Directed Acyclic Graph (DAG). DAGs are simply nonparametric and 'acyclic' path diagrams, where bidirectional relationships are replaced with a latent and two directional arcs (as in Figure 2.3.1).

Evaluating DAG-data consistency

As explained in the article that introduces *dagitty* the *R* package,[54] it is now feasible to evaluate consistency between one's DAG and an observed dataset. This effectively allows evaluation of every implied constraint within a DAG, i.e. the absence of arcs can be assessed for consistency against what this implies with respect to bivariate correlations within the observed data. These evaluations may be viewed as 'local' tests – in that each implied constraint involves only two or a few variables and the assessment does not need to involve all variables in the DAG. Most SEM model-fit assessments, in contrast, are 'global', though there 'modification indices' within SEM software packages attempt something similar to the local tests within *dagitty*. In practice, however, overzealous use of local 'tests' in *dagitty* (or modification indices in SEMs) risks the development of a DAG (or SEM) to be primarily data-driven and not theory-driven, which it should always be.

Clearly, if a dataset is not remotely consistent with the assumed DAG, things are wrong, but this should only encourage taking a step back and having some deep reflection, opposed to assuming the DAG is wrong and begin altering it. The DAG may be incomplete; it is possible you are missing some key variables. If it is unclear how this might be, as you have accounted for all possible confounding, for instance, the missing variable(s) could well be related to data provenance, i.e., what is commonly termed **selection bias** (aka **collider bias**, as we discuss later).

Data simulation using a DAG

Simulation studies should be kept as simple as possible (but realistic). It is okay to discuss the assumption of 'no unmeasured confounding' for toy examples where a method is being evaluated for its ability to recover the simulated 'truth'. All scenarios simulated must be approximate representations of reality, else there is little gained by the exercise. Bad simulations risk yielding false conclusions that could mislead regarding the appropriateness or robustness of the methods under evaluation. While important to simulate data adopting a DGM, the hypothesised DMG (and its associated DAG) must be accurate, which is the hard part of this process and why many simulations are so challenging.

When simulating data using a causal graph, there are several stages:

- **Theory**: In the first step, we hypothesize possible causal relationships between variables within an observational study setting.
- **Developing a causal model**: The second step involves translating hypotheses generated in step 1 into a possible causal graph. A sense of temporality is important. This involves showing how the variables are structurally related under naïve assumptions of linearity and some (possibly mixed) multivariate distribution (by assigning the standardised *beta* coefficients to the causal graph as path coefficients). It is important to note that these are NOT the same as bivariate correlations.
- **Derive the covariance matrix**: Once a causal model has been specified within a parametrised DAG, assess it to ensure that that the covariance matrix is positive definite. The causal structure imposed will limit the range of path coefficients; sometimes considerably so.
- **Simulate data**: The next step is to simulate the data. The distributional form of each variable can be specified (all multivariate normal variable or all binary variable can be simulated with the *R* package *dagitty* – other (mixed) distributions need a different *R* routine called GenData, see[55]). Standardised data are simulated since only the covariance *structure* (i.e. correlation matrix) is important for causal relationships, which means that initially we have covariance ≡ correlation. The standardised data may be transformed (i.e. rescaled) to represent real data, which have the correlation structure but the covariance matrix alters.
- **Check model statistics**: Compute basic statistics to assess the simulated data relationships and contrast to real-world scenarios – where there are discrepancies, tweaks to the path coefficients in the DAG may be explored and the process repeated until a satisfactory simulated data is achieved.

Assumptions

It is important to realise that initially naive assumptions are necessarily made about the causal graph to arrive at the desired data structure – this will a be a 'toy' study, not a real-world full-blown exploration of reality.

Assumptions around linear relationships between variables indicated by unitary arrows in a causal diagram are readily made without careful questioning. If there are nonlinear relationships between two variables in nature, this might be rectified by adopting a suitable variable transformation. Similarly, the assumption of all data being multivariate normal may be untenable. However, the same solution can apply (and solve both problems simultaneously) – for instance, when studying blood insulin levels, these often exhibit an underlying log-normal (i.e. skewed) distribution within the population and the relationship with other variables may be nonlinear. Taking the log of these measures should yield an underlying normal and may be sufficient to obtain approximately linear relationships with all other variables. Simulating multivariate normal data is therefore satisfactory for illustrating the consequences of different methods for 'toy' scenarios.

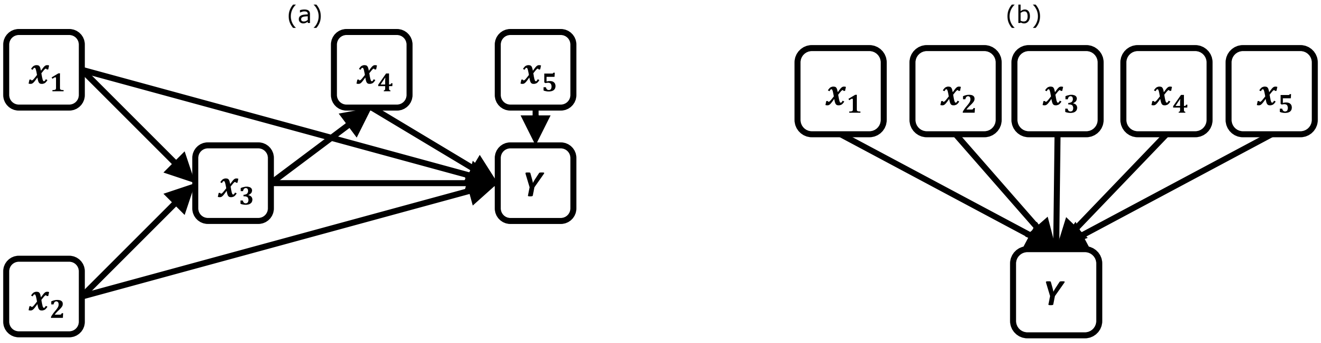
The assumption made that the simulated dataset is not influenced by factors not included in the dataset is sometimes judged naïve, but it can suffice to illustrate how methods work or to investigate hypothetical causal relationships in the absence of unmeasured confounding. One can always proceed to consider confounding. The view that 'toy' datasets never perfectly emulate real-world scenarios does not prohibit astoundingly insightful enquiries, as we show in this course.

Finally, path coefficients in the causal DAG will never represent the true causal relationship for real-world data (especially if some variables were transformed to overcome non-normally distributed measures and/or nonlinear relationships) – this is fine, as we only seek to explore the nature of the mechanisms or methods under investigation. The path coefficients deployed and results observed should be treated as approximate and subjected to sensitivity analyses to create confidence in the overall picture derived from the simulation.

Important implications surrounding DAGs and complex DMGs

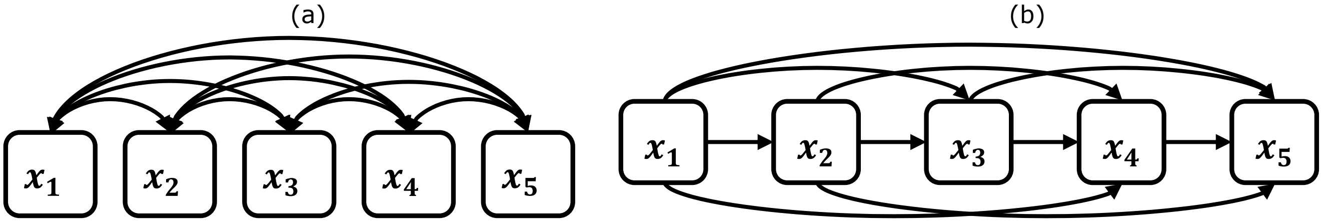
It is very important to recognize that the DMG (and resultant DAG that seeks to represent it) is a construct of theory in the researcher's mind. When data are subject to various statistical methods, such as regression analysis, the perception of variable interrelationships within the human mind and within software deploying those methods rarely have a 1-to-1 map. For instance, in Figure 2.3.3, we depict how the data relationships may be thought of by the researcher compared to how regression software 'sees' the data.

Figure 2.3.3: Path diagrams depicting: (a) the researcher's view of the data generating mechanism of the variables within the dataset; and (b) the statistical software's view of the data generating mechanism for a multivariable regression model.



From the perspective of statistical analyses, the data might be viewed as representing a snapshot in time, and therefore be thought of as **cross-sectional**; especially if all variables were collected contemporaneously. For many datasets, however, different variables – irrespective of when their values were elicited – represent different snapshots in time and might be considered **longitudinal**. This distinction is specific to each dataset and context, but matters as to how we conceive the data generating mechanism (DMG). The implications of this in terms of depicting the temporal relationship of variables within a DAG is seen in Figure 2.3.4.

Figure 2.3.4: (a) Path diagram depicting conditional independence among five cross-sectional variables that crystalise contemporaneously; and (b) a DAG depicting temporal dependence among five longitudinal variables that crystalise sequentially.



Parameterised versions of the two diagrams in Figure 2.3.4 have vastly different implications regarding the limitations in the ranges of potential path coefficients. This is readily understood by using Wright's path rules. In Figure 2.3.4 (a), any one standardised path coefficient may vary from -1 to $+1$, i.e. for any two variables x_i and x_j, their standardised path coefficient is: $\rho_{ij} \leq \pm 1$. This does not mean that all path coefficients may take such a full range of values. If $\rho_{12} = -1$, (i.e. x_1 and x_2 are perfectly negatively correlated), we recall what in terms of a DAG (see Figure 2.3.1) the bidirectional arc between x_1 and x_2 implies a common ancestor. In this instance, the common ancestor is fully deterministic of both x_1 and x_2. As a consequence, there can be no other arcs into either x_1 or x_2 (as they are fully determined by their common ancestor), which implies that all other path coefficients involving these two variables must be zero, i.e. $\rho_{1k} \equiv \rho_{1k} \equiv 0$ for $k = 3, 4, 5$. There are thus constraints on all path coefficients in Figure 2.3.4 (a) implied by the values of the other path coefficients, but in principle, any one path coefficient could be as large as ± 1. In practice, in most datasets, most path coefficients in Figure 2.3.4 (a) would be far from -1 or $+1$, and while other paths are constrained, not extensively so.

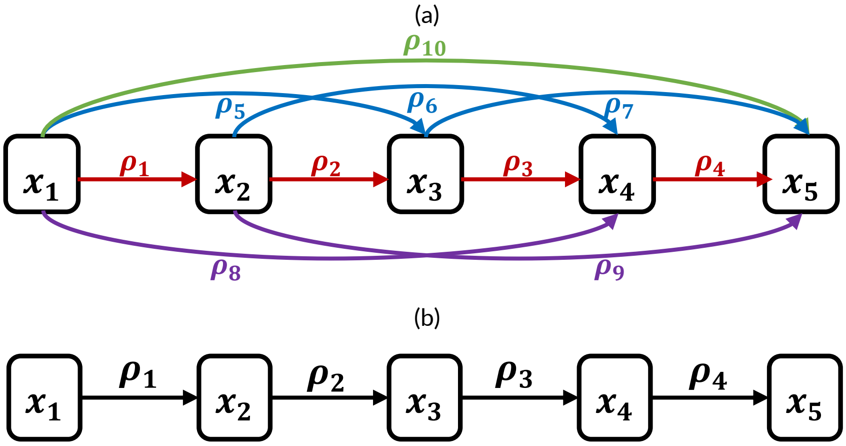
In Figure 2.3.4 (b), in contrast, constraints for standardised path coefficients are very different due to the temporal ordering of the variables in the DAG, where the direction of cause and effect is explicit. If we label all possible path coefficients, as in Figure 2.3.5 (a), this allows algebraic exploration of the constraints implied. Examining the relationship between the first (x_1) and last (x_5) variable, for instance, using Wright's path rules for all causal paths between them, we have the following constraint:

$$\rho_{x_1 x_5} = \rho_1\rho_2\rho_3\rho_4 + \rho_5\rho_3\rho_4 + \rho_5\rho_7 + \rho_1\rho_2\rho_7 + \rho_1\rho_6\rho_4 + \rho_8\rho_4 + \rho_1\rho_9 + \rho_{10} \leq \pm 1$$

This is quite a convoluted expression and reveals the complexity of the constraints imposed by the temporal order of the variables in the DAG.

If x is a longitudinal measure that exhibits modest within-subject variation over time – which is true of many growth measures band also many homeostatic biological variables – the expected size of the standardised path coefficients between successive measures (ρ_1, ρ_2, ρ_3 and ρ_4 in Figure 2.3.5) might be as large as ~ 0.8, which represents reasonably strong serial autocorrelation.

Figure 2.3.5: (a) DAG depicting temporal dependence among five longitudinal variables that crystalise sequentially with all potential path coefficients labelled; and (b) DAG depicting temporal dependence among five longitudinal variables that crystalise sequentially with only the sequential path coefficients linking labelled.



Simplifying the DAG in Figure 2.3.5 (a) to that in Figure 2.3.5 (b), and considering only the sequential path coefficients linking the longitudinal measures (each determined only by its immediately prior value, i.e. an autocorrelation structure of AR1), the following constrain is implied for the relationship between the first (x_1) and last (x_5) measures:

$$\rho_{x_1 x_5} = 0.8 \times 0.8 \times 0.8 \times 0.8 \approx 0.4.$$

In the naïve instance of no other causal influences of x_1 and x_5, and no other mediating variables, their bivariate correlation is not trivial.

With other variables at play, e.g. baseline features of sex and ethnicity, as well as longitudinal (time-varying) features, e.g. socioeconomic background, all of which requiring multiple path coefficients linking them to each other and to x_1 and x_5, it becomes apparent that most path coefficients must be relatively small to avoid generating unrealistic bivariate correlations between temporally distal measures.

Understanding the DMG and depicting this by a DAG translates the implied temporal structure of restrictions to the simulated data that then more realistically reflects the context under investigation. This approach to simulation is fundamentally different from starting with the (observed) correlation structure of a dataset that depicts only the consequences of the (unmodelled) DMG. This cross-sectional view of consequence cannot reflect structure the same insight as the longitudinal view of unfolding (DGM) process. This can matter enormously in the range of potential datasets generated and subjected to various evaluations, affecting the hypothetical 'solution space' of all possible simulations. If the DGM is not captured accurately, the results of the simulation exercise may be at risk of being misleading.

Discussion

The standard simulation process has, for many, involved simulating data that follows a particular covariance (i.e. correlation) structure. The challenge with this approach is that simulations may fail to capture realistic scenarios of how observed data come into being, and the corresponding propensities for different dataset structures. The observed correlation at any one specific timepoint has many potential different underlying causal generating mechanisms, yet only one is true. It is not sufficient to specify the covariance structure without reflecting upon the data generating mechanism.

2.4 PROPENSITY SCORE APPROACHES

Learning objectives:

- Understand the core idea behind a 'propensity score analysis'
- Recognise why propensity score methods and traditional 'single model' methods, if conducted well, should give similar if not identical results
- Appreciate the potential benefits of a propensity score approach

A popular and unpopular approach

Besides traditional theory-free approaches, **propensity score methods** are arguably the most popular method for estimating causal effects in health and medical research. At the least, they are certainly some of the most hyped, with many authors claiming that their propensity score analysis has enabled their study to 'mimic' or 'replicate' a randomised experimental study. Such extreme exaggeration has in-turn prompted cynicism and antagonism among many statistical experts. Unsurprisingly, propensity score methods are neither miracle nor malady but tell a familiar tale of promising tools that are typically used badly. In this regard, as in many ways, propensity score methods are not unlike the traditional alternative of '**directly adjusting for confounding**'.

A propensity score approach

The aim of a propensity score 'approach', like a conventional 'single model' analytical approach, is to obtain conditionally exchangeable units of analysis from which to estimate the causal effect of an exposure (or regime) on an outcome. The difference with a propensity score approach is that these two aspects are explicitly separated into stages; the first of which focusses on obtaining conditionally exchangeable units and the second of which then estimates the exposure-outcome relationship.

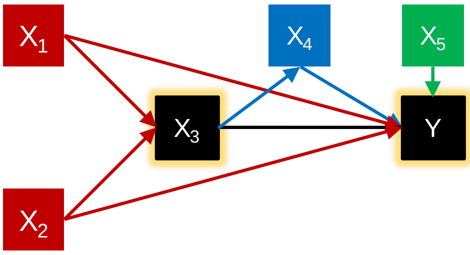
The primary difference between this approach and a conventional 'single model' approach is therefore more practical than theoretical. In fact, a good propensity score analysis should produce very similar (if not identical) results to those generated by a good conventional 'single model' analysis.

The two stages

The two stages of a propensity score analysis consist of:

1. Estimating the propensity of exposure, known as the propensity score; and
2. Estimating the exposure-outcome *association*, conditional on the propensity score.

To demonstrate, consider the following DAG and suppose we are interested in the causal effect of $\mathbf{X_3}$ on **Y**:

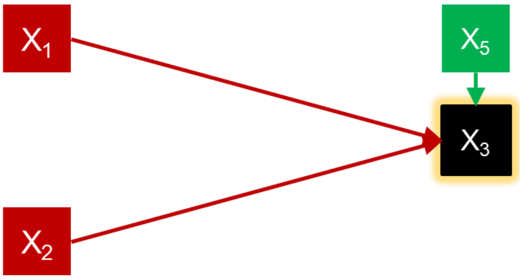


In a conventional 'single-model' that included the exposure and the outcome, this might be estimated from:

$$Y = f(X_3, X_1, X_2, X_5, \varepsilon)$$

Our DAG tells us to include $\mathbf{X_1}$, $\mathbf{X_2}$, as confounders; and (optionally) $\mathbf{X_5}$, because it is a competing exposure.

In a propensity score analysis, we first estimate the propensity of our exposure (i.e. propensity score). The model should *include* **confounders** and **competing exposures** but **NOT mediators** for the relationship between the exposure and the *outcome* (even though the outcome is not explicit in the propensity score model). In theory, the model should not include **instrumental variables**, but in reality these are likely to retain some residual association with the outcome and should therefore be included where accuracy is prioritised over precision.[56]

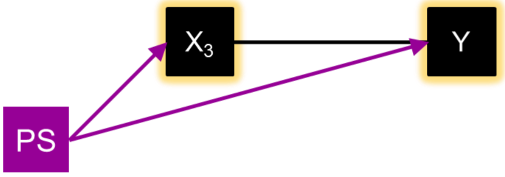


$$X_3 = f(X_1, X_2, X_5, \upsilon) = PS + \upsilon$$

The **propensity score (PS)** obtained from this model captures our best estimate of the parts of X_3 that are determined by the confounders X_1 and X_2. Conditioning on competing exposure X_5 will also remove any small 'random confounding' from X_5 within this sample.

Note: Although X_5 is not a cause of X_3, in any finite sample it will have a small random association with X_3. This non-causal association will be passed on through $X_5 \rightarrow Y$ to introduce error in the relationship between X_3 and Y, which we can remove by conditioning.

Now we have an estimate of the propensity score, we estimate the association between the exposure and outcome, conditional on the propensity score:



$$Y = f(X_3, PS, \varepsilon)$$

The propensity score captures variation in X_3 due to confounders X_1 and X_2 and random confounder X_5. The association between X_3 and Y, conditional on PS, therefore represents our estimate of association that is *not* due to **full** or **random** confounding by X_1, X_2, X_5; and estimates the total causal effect of X_3 on Y.

Conditioning on the propensity score in the outcome model

Because of the long history of propensity score methods, a range of approaches are popular for conditioning on the propensity score in the second model of the exposure-outcome relationship.

Propensity score matching

Propensity score matching describes the analytical technique whereby a subsample is constructed from 'exposed' and 'unexposed' participants matched into pairs/groups based on their propensity score. This method remains popular in clinical communities, probably because of the conceptual similarity with top-down study designs like randomised experimental studies and case-control studies. The approach allows statisticians to produce a table showing the 'covariate balance' between the 'exposed' and 'unexposed' participants.

Other than these translational benefits, however, there are few justifications for matching since it requires subjective decision-making on the best approach to identifying matches and often requires discarding a substantial minority of the sample (i.e. all those participants who end up 'mismatched'). Perhaps more importantly, the final analytical sample is also unlikely to represent the original sample or population from which the participants were drawn, leading to potential selection bias. Indeed, propensity score matching is often described as providing a different casual effect from the (usually desired) *average causal effect in the* *population* (or average treatment effect in the population, ATP) known as the *average causal effect in the exposed* (or average treatment effect in the treated, ATT).

Propensity score weighting

At the other extreme, **propensity score weighting** uses **inverse probability weighting (IPW)** to create a **pseudo-population** with the same distribution as the original sample, but where the 'exposed' and 'unexposed' groups now have equal overall propensity. Unlike matching, this approach can be extended to suit continuous exposures and can also provide the average causal effect in the population or any subset desired. The primary downside of propensity score weighting is that the estimates can become unstable if the sample includes participants with extreme propensity score values (who end up carrying extreme weights).

Propensity score covariate adjustment

By far the most common approach to propensity score analysis is **propensity score covariate adjustment**, in which the propensity score is simply included as a covariate in the exposure-outcome regression model. Depending on the approach used to estimate the propensity score, this is likely to produce a very similar if not identical estimate to a conventional 'single model' approach '**directly adjusting for confounders**'.

Why bother?

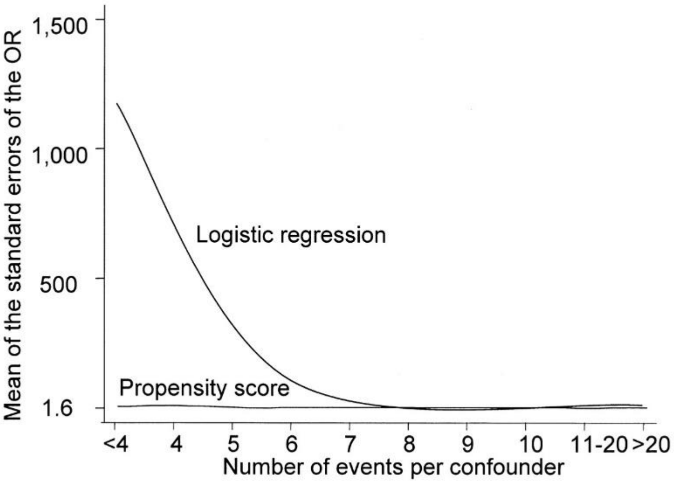
The realisation that a good propensity score analysis will produce similar if not identical results to good 'single model' analysis raises the question, why bother at all with a propensity score analysis? Here are some arguments in favour:

Propensity score analyses discourage the Table 2 Fallacy

Since propensity score models are explicitly built for each exposure-outcome relationship and the final exposure-outcome model does not include any confounders and competing exposures, then there is far less risk of these being misinterpreted.

Propensity score analyses have superior power for rare outcomes.

In a single model of a rare outcome, there may be too few events to accurately estimate the confounding relationships for all confounding strata leading to residual confounding. A propensity score approach avoids this by focussing on modelling the confounder-exposure relationships, where more granular information will likely be available. As the number of outcome events per confounder falls below ten, a propensity score approach will provide increasingly superior power to a single model approach.[57] (Figure reproduced from Cepeda *et al* 2003[57])



Propensity score analyses encourage checking for positivity violations

The structural positivity condition requires that all values of the exposure must be *possible* within all confounding strata. Ideally, this means that all values of the exposure are directly *observed* in all confounding strata, else you will at least have a random positivity violation.

Propensity score methods have historically always encouraged analysts to study the 'balance' of their confounders across exposure groups, and indeed to check that all exposure values are possible for all propensities. This process, known as '**estimating overlap**', is technically a test of random positivity. Where the propensity score distributions overlap, the units are clearly exchangeable (on the measured confounders). Where they do not overlap, however, there are no counterfactuals in your data and these should arguably be excluded (or '**trimmed**') from your analysis. The trimmed distribution may not however accurately reflect your population. Therefore, you must consider carefully whether it is more important to maximise positivity or representativeness. If you are happy that the positivity violation is not structural, you may choose to analyse both the complete sample and the trimmed subsample as a sensitivity analysis.

Propensity score analyses are suitable for flexible machine learning algorithms

Machine learning methods are not currently well suited to estimating causal effects in observational data as they are generally best suited to theory-free predictive modelling. A propensity score approach resolves this incompatibility by reframing the challenge of modelling the propensity score as a predictive one. Machine learning methods can then be used to build the best possible model of the relationship between the confounders and the exposure, reducing the risk of residual confounding from imperfect parameterisation.[58]

It is suggested that **high dimensional propensity score models** (i.e. propensity score models with many confounding variables), if modelled correctly, can reduce *unobserved confounding*;[59] presumably because the greater the number of confounders that are conditioned on, the more confounding pathways are blocked. This should not be relied upon to resolve unobserved confounding from important but unavailable confounders. In such instances, **quantitative bias analysis** approaches should be preferred.[60]

The most obvious downside of a propensity score approach is that by removing the confounders from the exposure-outcome model, it is not possible to examine confounder-exposure interactions; nor likewise to estimate counterfactual contrasts at specific confounding levels. If one or more confounders is of interest in this way, they should ideally be brought into the exposure regime; estimation may then require G-methods.

For further reading see Austin 2011.[61]

DAY 3

3.1 Introduction to Collider Bias

Learning objectives

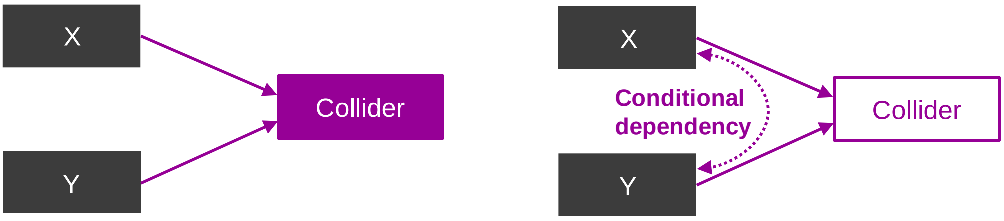
- Understand what is meant by "collider bias"
- Recognise how collider bias can lead to bizarre associations and apparent "paradoxes"

Collider bias and paradoxical consequences of statistical adjustment

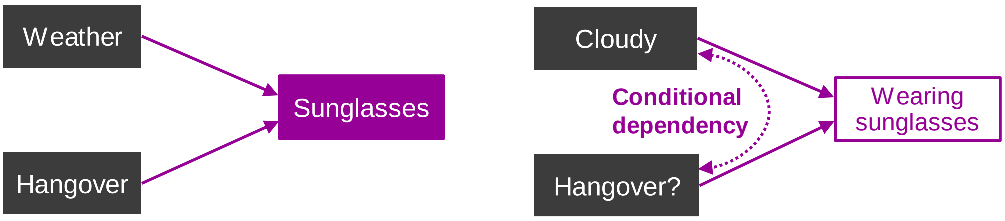
Previous sections have focussed on the benefits of identifying and conditioning for confounders, while ignoring mediators. In part, this is because adjusting for mediators will block some of the total causal effect of the exposure on the outcome. Of more concern, adjusting for mediators also risks introducing serious analytical artefacts and apparent paradoxes. The chief culprit behind these, and a similar but less visible bias, is a phenomenon known as **collider bias**.

Collider bias is a difficult concept to understand. Directed acyclic graphs help to shine a light on this phenomenon but, in our experience, it still takes some time to grasp. For this reason, recognising sources of collider bias has (to date) probably been the biggest source of revelation to emerge from causal inference methods.

Colliders are variables that are caused by two (or more) separate causal paths. They are described as colliders because two or more paths (and their arrowheads) '**collide**' at that node. In a DAG, colliders operate like reverse confounders. In other words, no association is normally transmitted across a collider. If two independent variables cause a collider, we do not therefore expect them to carry an association. However, if we inappropriately and/or inadvertently condition on the collider, we create a spurious **conditional dependency** between all variables on the previously unrelated paths, and this is known as **collider bias**. These dependencies can appear quite paradoxical.



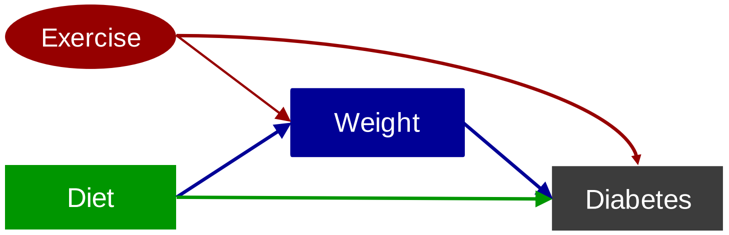
For example, consider two reasons for wearing sunglasses: sunny weather and alcohol hangover. Suppose these are unrelated in the general population. Now suppose we observe someone wearing sunglasses (i.e. conditioning on their sunglasses status) there will now be a conditional dependency between the weather and their likelihood of having a hangover. Thus, if we observed someone wearing sunglasses on a cloudy day, then their likelihood of being hungover is greater.



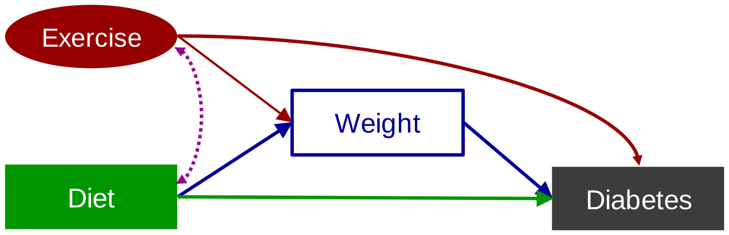
Similar conditional dependencies will be observed between any two variables with a common consequence. If you condition on the consequence, you will create an (inverse) dependency between the two variables. If these two variables were your exposure and your outcome, and you conditioned on a common descendent, then your estimated effect would be biased by collider bias.

Conditioning on a mediator risks introducing collider bias whenever the mediator is caused by other unobserved variables that in turn cause the outcome; these are often called **mediator-outcome confounders**. Without mediator adjustment, there may be no association between your exposure and the unobserved mediator-outcome confounder. Conditioning on the mediator would create a conditional dependency between your exposure, the unobserved mediator-outcome confounder and, in turn, the outcome. Your estimated effect of the exposure on the outcome would thus be distorted by collider bias.

For example, consider the effect of diet on diabetes, mediated through weight. You have no information on exercise, which also causes weight and diabetes. The total effect of diet on diabetes would be estimated without conditioning on weight. If you conditioned on weight, however, this generates an inverse conditional dependency between diet and exercise and in turn between exercise and diabetes. Your estimate of the effect of diet on diabetes is thus distorted by collider bias, as you now introduce (rather than eliminate) confounding from exercise.



Diet → **Weight** ← **Exercise** → **Diabetes** is **closed** and not transmitting association



Diet → **Weight** ← **Exercise** → **Diabetes** is **open** and transmitting **collider bias**

Paths connecting the exposure, mediator, mediator-outcome confounder, and outcome need not be causal for collider bias to occur. Conditioning on a collider will create spurious conditional dependencies between all variables and all open paths (whether causal or not), either side of the collider.

Smoking during pregnancy and infant mortality

We explore a real example of collider bias through the so-called **(low) birthweight paradox**. Here, the association between **smoking** during pregnancy (exposure) and **infant mortality** (outcome) is examined whilst 'adjusting' for **birthweight** (an alleged 'confounder'). The 'paradox' emerges as findings from the (mis-specified) multivariable model are contrary to expectation, showing that:

- Mean birthweight is lower amongst mothers who smoke during pregnancy compared to mothers who do not;
- Overall infant mortality is *higher* amongst mothers who smoke during pregnancy compared to mothers who do not; whilst 'paradoxically',
- When examining birthweight subgroups, infant mortality rates appear *lower* amongst mothers who smoke during pregnancy than those who do not.

This was first exposed as a consequence of poor comprehension of causal inference by Hernandez-Diaz et al.[62] and Wilcox.[63] In October 2014, an entire edition of the IJE was dedicated to this topic.

If data corresponding to this problem are categorised (Table 1.3.1), the phenomenon is recognised as Simpson's paradox; and if data are continuous and considered within a multivariable model (Table 1.3.2), the phenomenon is recognised more generally as the reversal paradox.

We illustrate Simpson's paradox with simulated data: one million mother and child pairs with data on birthweight, mothers' smoking behaviour during pregnancy, and infant mortality. Table 1.3.1 shows that the rate ratio *within* birthweight groups is always <1.0 whilst *overall* (across groups) it is >1.0.

The reversal paradox is demonstrated in the multivariable regression model presented in Table 1.3.2, where the linear model that is not adjusted for birthweight yields *elevated* odds of infant mortality amongst mothers who smoke during pregnancy (OR = 1.07, 95%CI = 0.98-1.17), whilst the model adjusted for birthweight yields *reduced* odds (OR = 0.70, 95%CI = 0.64-0.77).

Table 1.3.1: Simulated data to illustrate the birthweight paradox: birthweight, mother's smoking behaviour during pregnancy, and infant mortality for 1 million mother and child pairs

Birth weight Range (Kg)	**Mothers who smoked**			**Mothers who did not smoke**			**Rate Ratio**
	Live Births	**Infant Deaths**	**Mortality Rate[1]**	**Live Births**	**Infant Deaths**	**Mortality Rate[1]**	
(0.5,1]	2	1	500.0				
(1,1.5]	64	2	31.3	68	6	88.2	0.35
(1.5,2]	1,394	30	21.5	2,250	59	26.2	0.82
(2,2.5]	10,360	127	12.3	30,018	524	17.5	0.70
(2.5,3]	30,318	188	6.2	158,876	1,453	9.1	0.68
(3,3.5]	36,694	143	3.9	329,896	1,528	4.6	0.84
(3.5,4]	17,406	26	1.5	275,228	692	2.5	0.59
(4,4.5]	3,510	3	0.9	91,288	102	1.1	0.76
(4.5,5]	244	0	0.0	11,768	12	1.0	
(5,5.5]	8	0	0.0	600	0	0.0	
(5.5,6]				8	0	0.0	
Total	100,000	520	5.2	900,000	4,376	4.9	1.07

[1] per 1000 live births

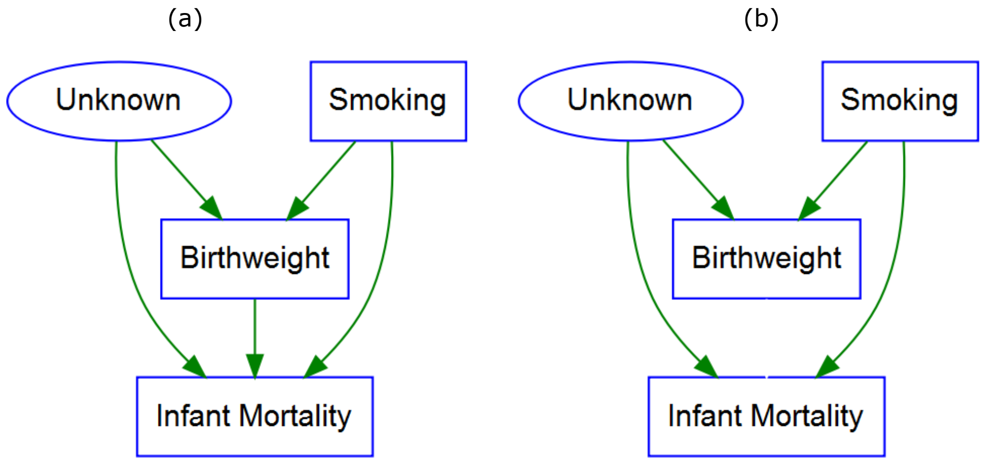
Table 1.3.2: Regression model of infant mortality (outcome) on mother's smoking behaviour during pregnancy (exposure) unadjusted and adjusted for infant birthweight for the simulated data summarised in Table 1.3.1

Model	**Estimate**	**95% CI**
Smoking exposure during pregnancy (unadjusted for birthweight)		
Non-exposed mortality rate[1]	4.86	4.72, 5.01
Smoking exposure odds ratio	1.07	0.98, 1.17
Smoking exposure adjusted for birthweight		
Base mortality rate[1,2]	3.32	3.19, 3.45
Smoking exposure odds ratio	0.70	0.64, 0.77
Birthweight odds ratio[3]	0.25	0.23, 0.26

[1] per 1000 live births; [2] centred on birthweight of 3.5 Kg; [3] per 1 Kg increase in birthweight

The problem lies in treating birthweight as a 'confounder'. Two potential causal relationships are given in the DAGs of Figure 7.

Figure 7: DAGs for relationships amongst mothers smoking behaviour during pregnancy, their infant birthweight and risk of infant mortality, with a common unknown cause of birthweight and infant mortality: (a) infant birthweight is causally related to risk of mortality; (b) infant birthweight is NOT causally related to risk of mortality



There is evidence that lower birthweight children are more at risk of infant mortality due to causal antecedents that affect both foetal health (perhaps leading to premature birth) and infant health, thereby causing a greater risk of infant mortality (Figure 7a). There is also strong evidence that smoking during pregnancy causes lower birthweight and is hence a descendant of the smoking exposure (Figure 7a & 7b). Whether birthweight causes infant mortality (Figure 7a) or not (Figure 7b) is irrelevant, since either:

- Birthweight is a **mediator** (Figure 7a) and should not be adjusted for as the effect sought is the *total* causal impact of smoking during pregnancy on infant mortality; or
- Birthweight is a **collider** that is also a proxy for (one or more unknown) antecedents of infant mortality (Figure 7b) that then become correlated with the smoking exposure when birthweight is conditioned upon; thereby opening a backdoor path to these antecedent, leading to a biased estimate of the causal effects of smoking on infant mortality.

Summary

There are many instances in observational research where the adjustment set of covariates in the multivariable model employed is not carefully and robustly justified within a causal framework. By conditioning on a collider, estimated effect sizes may be seriously misleading. Failure to recognise this has given rise to considerable misunderstanding in the literature and may sometimes lead to ongoing controversies that are described as a paradox. In practice, there is no such paradox; just the failure to think and model observational data in a causal framework.[64,65]

3.2 Selection Bias

Learning objectives

- Understand how an unrepresentative sample can lead to biased estimates due to implicit conditioning on selection
- Recognise M-Bias as a common potential risk in observational studies
- Understand how M-Bias can even bias prospective studies

Collider bias as a universal problem

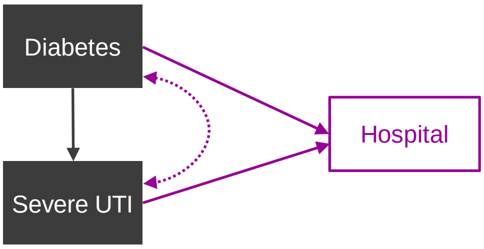
In the previous section, we learnt how **spurious associations** can be introduced between two or more variables by inappropriately and/or inadvertently conditioning on a **collider**. Here we will learn why all observational studies are probably affected to some degree by collider bias, even if adopting robust analytical strategies.

Selection bias and Berkson's paradox

Selection bias is a form of collider bias that occurs because selection into your study acts as a collider for open paths that include your exposure and outcome. Selection (the 'collider') is implicitly conditioned in any subsequent analyses because your analyses are 'restricted' to the collected sample.

Selection bias happens most obviously – and probably most severely – in situations where your exposure and your outcome are both in some way directly responsible for selection. The most well-known example of this is called **Berkson's paradox**, in which a paradoxical inverse association is introduced between two health conditions. It arises in studies that seek to explore the association between two competing reasons for visiting hospital (although it could equally apply to any healthcare setting).

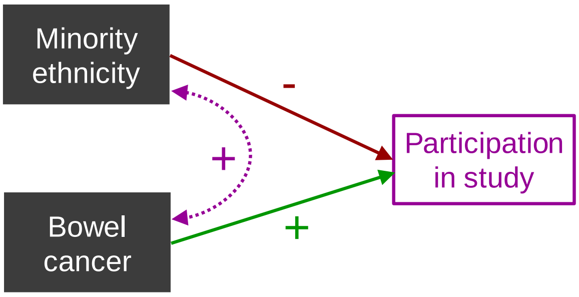
For example, the investigator might be interested in the relative risk of a severe urinary tract infection (UTI) in people with diabetes. The investigator knows that UTIs are generally more common in people with diabetes but wants to know by how much. They attempt to answer the question by looking at hospital admissions data and are surprised to find that the occurrence of severe UTIs is *lower* in people with diabetes! What they haven't realised is that because both diabetes and severe UTI are separate reasons for hospital admission, when they are only studying those who attend hospital (thus conditioning on hospital admission), they create a conditional dependency.



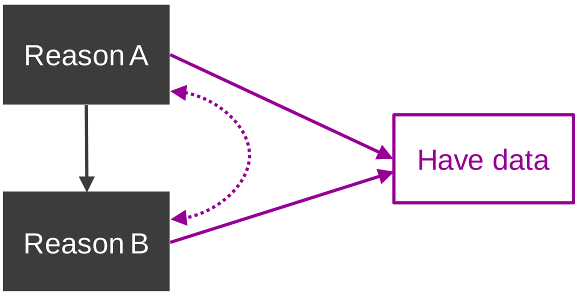
This is clearly a risk for all studies in electronic health records because an entry is only made when someone attends a healthcare appointment, which is usually due to illness.[66]

This phenomenon is also a problem in case-control studies (where exposure information for people with a condition or disease is compared to exposure information for people without that condition or disease), because cases are generally much keener to participate. This will create a spurious association for any exposure that is itself associated with participation.

For example, in a case control study of the risk of bowel cancer by ethnicity, people with bowel cancer are more likely to participate. Similarly, people from minority ethnic groups are less likely to participate. Because your sample implicitly conditions on likelihood of participation, a spurious association would therefore be created to imply a higher risk of bowel cancer in people from minority ethnic groups.



More generally, this issue essentially applies to any dataset with non-random entry; a problem known as **informed presence bias**.[66] Put simply, informed presence bias says, "there are reasons that we have these data for these people that make them different to the people on whom we don't have data".



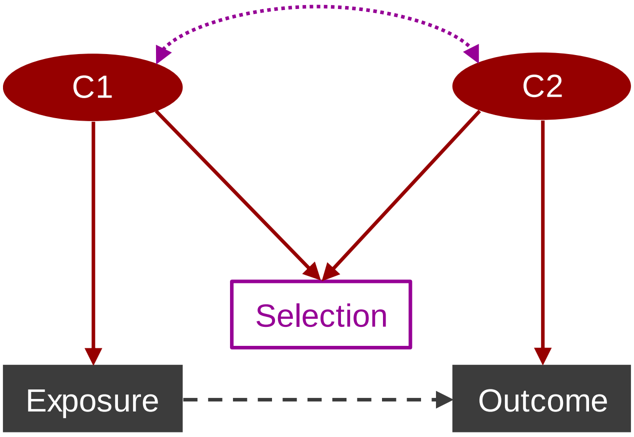
One reason for informed presence bias is our **sampling strategy**, i.e. the approach that we have used to identify participants. If we study purchasing data from Waitrose and how that varies by season, for example, it would not likely be the same as data from Asda. Even with a truly random and representative sampling strategy, however, peoples' **willingness** and **ability** to participate and provide data are influenced by a large range of factors, including their **health**, **education**, **beliefs**, **psychology**, **personality**, and **economic circumstances.**[67]

Selection and prospective studies

There is heavy discussion about whether a study or sample needs to be representative of the population to draw valid inferences about the effect of an exposure on an outcome. While it is recognised that unrepresentative studies provide poor occurrence estimates (e.g. prevalence, incidence, etc.), it is argued by some that representativeness is not necessary for robustly estimating associations.

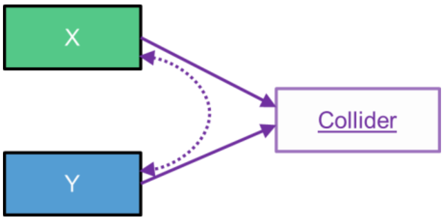
For example, the UK Biobank study approached around 5-10 million people to enrol in their study to recruit a sample of 500,000. The 5-10% who agreed to participate were not representative of the general population in terms of "sociodemographic, physical, lifestyle and health-related characteristics".[68] UK Biobank advises data users to state that "valid assessments of exposure-disease relationships are nonetheless widely generalizable and do not require participants to be representative of the population at large". The most likely reason for this statement is that the outcome – and often the exposure – occurs after the decision to enter the study; therefore, participation cannot *cause* the outcome. Unfortunately, direct causation is not required for differential selection bias to occur; only that the exposure and the outcome are *associated* with participation. Two unobserved confounders could therefore introduce collider bias if they associate the exposure and outcome with study selection. Conducting your analysis in the sample would condition on the participation node and create a spurious association. Such a scenario is known as M-bias.

Real-examples of M-bias are extremely difficult to identify because it requires that you know about the people who did *not* participate. Nevertheless, Munafo *et al.* conducted simulations based on the participant profile of UK Biobank and other cohorts and found M-bias effects for the effect of personality and cognitive function on a range of physical and mental health outcomes that were indistinguishable from what had been found in the published literature.[69]

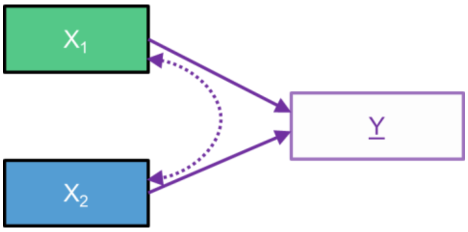


Outcome truncation bias

So far we have considered the effect of collider bias between two variables that are related to the collider, e.g. between X and Y in the figure below:



A similar bias can occur when the collider is the outcome itself, as in X_1 or X_2 and Y in the figure below:

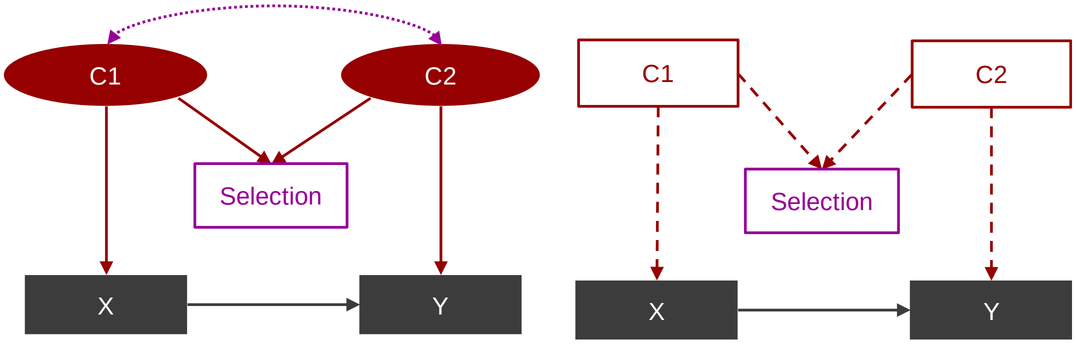


This situation occurs when you analyse an extreme non-random sample of your outcome, a phenomenon known as **conditioning-on-the-outcome**. In this outcome truncated sample, the effect of your exposure on the outcome is now in competition with all other causes of the outcome. E.g. in the figure above, the apparent effect of X_1 on Y would be biased by the open path X_1 <…> X_2 → Y. This bias may be known either as conditioning-on-the-outcome or **outcome truncation bias**. An example might be studying the effect of diet on weight only in obese adults. Here, the outcome (weight) has been truncated, meaning that biasing paths will be opened between diet and weight through all other causes of weight.

Solutions to selection bias

It is likely that all studies are affected to some degree or another by differential selection bias. Unfortunately, due to low recognition of the problem, it is not always possible to solve. It would, however, be a huge advance to simply recognise and highlight the **potential** for differential selection bias, as it may help to explain apparently paradoxical results and other heterogeneity.

Alternatively, several potential solutions are available, provided one has some contextual data. It may be possible, for example, to reduce M-bias if one has information on the key determinants of participation, allowing (partial) closure of the spurious paths by conditioning.[70] In this regard, it has been shown that adjusting for the number of healthcare encounters likely reduces some of the informative data bias in electronic health records.[66]



If we don't know the exact causal paths involved, we can still attempt to remove the spurious path using inverse probability weighting. Reweighting the sample to remove any association between your exposure and the likelihood of participation would reduce differential selection bias.[70] Ideally, this would require additional information on the determinants of participation to be collected, e.g. by using double consenting methods. However, it may still be possible to estimate this information if similar socio-demographic information can be collected on the source population.[71]

Summary

Selection bias is a common variety of collider bias that occurs due to non-random sampling and/or participation into a study sample. The best way of reducing such biases is to carefully consider where they may arise during data collection and try to collect as much information as possible on non-participants (to enable corrective treatment through inverse probability weighting).

Even where no solution is obvious or possible, as with some routinely collected data, it would be a substantial scientific advance to recognise and highlight how our data may be affected by non-trivial selection bias.

3.3 Regression to the Mean and Conditioning on the Outcome

Learning objectives

- Describe **regression-to-the-mean**
- Use DAGs with error terms to understand issues relating to RTM
- Explain how **conditioning-on-the-outcome** and outcome truncation can introduce **collider error**

Regression to the mean (RTM)

Regression-to-the-mean is a universal phenomenon that arises whenever two variables are imperfectly correlated (i.e. are not simply scale changes of each other). First described by Sir Frances Dalton in 1886 - and termed 'regression towards mediocrity',[72] and later termed 'centripetal drift' by Jung in 1938,[73] the phenomenon is most well known in the context of repeated measures, where we are told that after any unusual/extreme event, some 'correction' back towards the mean is likely.

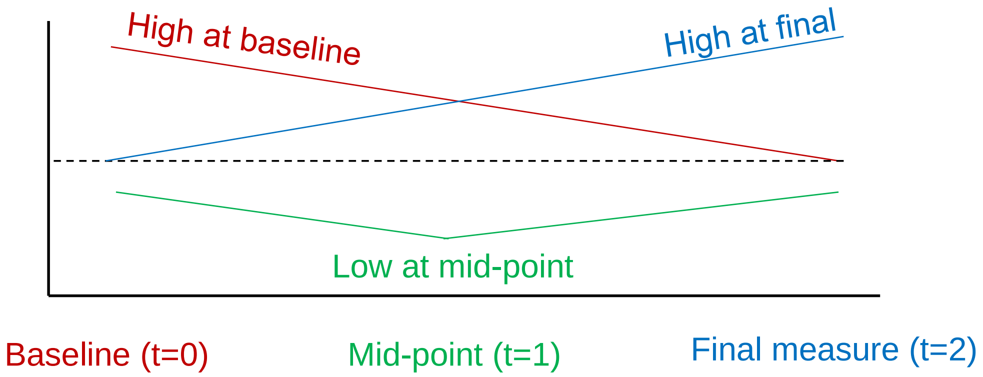
A recent high-profile example, is the performance of the Leicester City Football Club under manager Claudio Ranieri during 2015/16 and 2016/17. Claudio Ranieri took over managing the team for the 2015/16 season after they narrowly escaped relegation in the previous year. Under their new manager, the team shocked the world by winning the English Premier League in 2015/16 having started the season with odds of 5000-to-1. In the following year, however, the club once again struggled to avoid relegation and the manage Claudio Ranieri was sacked. This astonishing 'regression towards mediocrity' exemplifies the phenomenon that can be more formally described as follows:

> Following an extreme *random* event, the next *random* event is likely to be less extreme. Thus, for two independent normally distributed variables, X and Y with correlation zero: Given any extreme (e.g. unusually high) value of X, the value for Y is likely to be less extreme.

Regression-to-the-mean is often understood as a consequence of 'measurement error', but is actually a consequence of all forms of variation, natural (i.e. biological) variation, enigmatic (i.e. obscure) variation, all forms of error, and all other forms of randomness.

Why does it matter

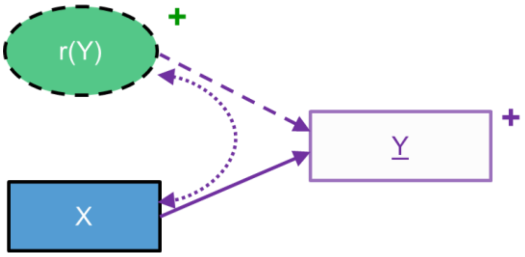
It is important to understand regression-to-the-mean because it can be seriously misleading; particularly when it is mistaken for causal or organic change. Such 'change' will appear to occur whenever we naively analyse an extreme subgroup. This commonly occurs when we condition-on-the-outcome, usually by chosing to study a group of subjects with unusually high starting values (e.g. obese individuals) or unusually high final values (e.g. individuals who end up being diagnosed with a disease). In longitudinal data, when we move away from the extreme timepoint that we conditioned on, we will appear to see a diluting trajectory,[74] as shown below:



Collider error and DAGs with error terms

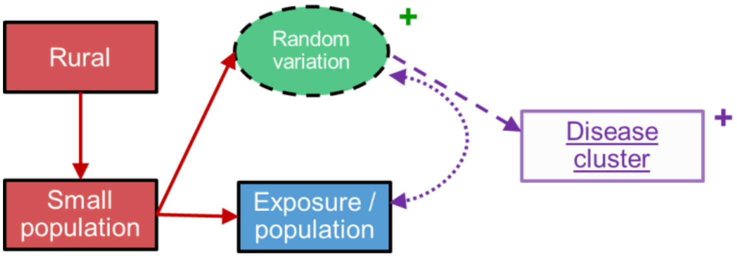
The problems with regression-to-the-mean and conditioning-on-the-outcome can be understood by depicting random terms within DAGs.

In Section 3.1, we learnt how conditioning-on-the-outcome could introduce outcome truncation bias, because the effect of the exposure on the truncated outcome would be biased by all other competing causes of the outcome. What we didn't say was that this would include all random reasons for the outcome variable taking an extreme value. This is shown in the simple DAG below:



In this example, only 'high' levels of Y are being considered (Y+). This means that the effect of X on Y+ will be *errored* by the path X<…> r(Y)+→Y+; a phenomenon we call **collider error**. Depending on the contribution of random variation to Y, this may have a very large and distorting impact.

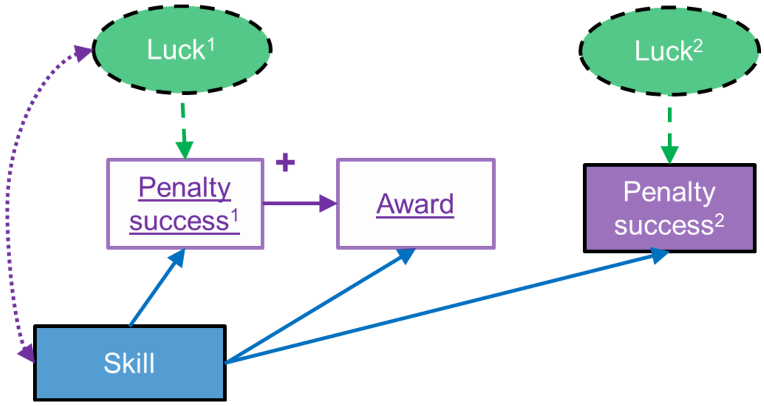
Any exposure that is correlated with random variation will be particularly distorted. In general, random variation is more common in small samples (due to increased sampling volatility), thus any exposure related to a small sample will likely be correlated with the truncated outcome. For example, a ratio variable with a population denominator is likely to be inversely related to the outcome if only extreme values are examined, as in the following example DAG:



This problem is described in more detail in Berrie et al.[75]

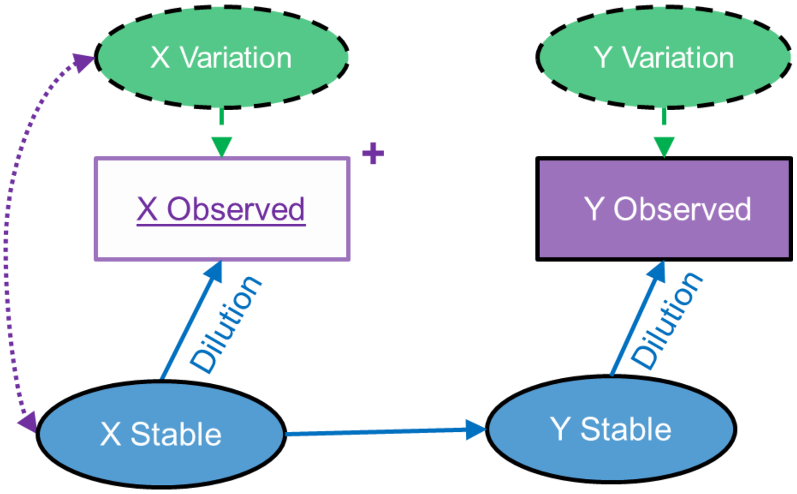
Regression-to-the-mean can similarly be understood by drawing a DAG with error terms. The problem occurs because by conditioning on the outcome at baseline, we introduce a dependency with all random causes of the outcome at baseline. All subsequent associations in the same sample, will not benefit from this non-random 'boost', leading to RTM.

For example, a recent paper studied the penalty performance of various football stars before and after winning an award. Before the award, the players successfully scored on 89% of occasions. After the award, this declined to 65%. The authors attributed this to the extra pressure of having won the award. It is more likely an artefact of regression-to-the-mean. By studying award winning footballers, they conditioned on the players having performed at 'award winning' levels at baseline. In any subsequent sample, all random determinants of their success will return to having a zero average, resulting in the apparent decline in performance. This is show in the figure below:



The general case can be understood by considering all observed variables as being caused by two components, the stable element of that variable and the variation element of that variable. This approach was introduced by Glymour et al with what has been termed 'measurement error DAGs';[76] which describe the 'observed' variable being caused by a latent 'true' variable and an error term.

For any random sample of X, we can expect the effect of X on any other variable Y to be diluted by the imperfect correlation between the stable and observed values of X and Y. For any extreme sample of X, this will be additionally errored by the conditional dependency between the stable value of X and the variation in X. This is show in the figure below:



3.4 Natural Experiment Approaches

Learning objectives:

- Understand what is meant by a 'natural experiment'
- Understand the principles, caveats, and assumptions behind the four most common 'natural experiment' approaches

Triangulating evidence

Estimating causal effects, in the absence of randomisation, is notoriously difficult. This creates a fundamental problem for health and social science researchers, since most exposures are simply not suitable for experimental study.

The methods discussed in the other chapters provide a framework to help improve our understanding of observational data and our ability to accurately estimate causal effects. But even with deep scholarship, high-quality data, and diligent data analysis, we cannot eliminate bias. After years and decades of low-quality research, faith in observational data analysis is relatively low; particularly in the health sciences, where it is widely taught that observational data cannot be used to estimate causal effects.

One solution that is popular in econometrics is to seek out circumstances where experimental conditions have been approximated by external forces. Although still 'observational studies' these '**natural experiments'** offer an alternative means to estimating causal effects and are therefore an important aspect of **triangulating** evidence of causation.

True natural experiments

A 'true' natural experiment occurs when 'nature' or some other exogenous force assigns an exposure 'as random' to approximate unconditionally exchangeable units of analysis.

By far the most famous example of a true natural experiment is John Snow's 'Grand Experiment' of 1984. The 'experiment' sought to estimate the effect of exposure to 'ingesting contaminated water' on contraction of Cholera. This was possible thanks to both a coincidence of circumstances and Snow's ability to recognise the opportunity that this provided.

Snow observed that two competing water companies (the Lambeth Company and Southwark and Vauxhall Company) were in direct competition in the same region of South London but sourced their water from entirely different places. The consequence was that, "*three hundred thousand people of both sexes… every age and occupation… every rank and station… were divided into two groups without their choice and… (or) their knowledge (into those) supplied with water containing… sewerage… (or) water… free from such impurity*". By comparing the number of cholera cases per household, Snow was hence able to estimate the causal risk ratio of exposure to ingesting contaminated water on risk of cholera.[77]

Unfortunately, true natural experiments are very unusual; particularly where the naturally-assigned exposure is the same exposure we are interested in studying. Unsurprisingly, it is also very rare to encounter circumstances where the units of analysis are truly exchangeable, and where all values of the exposure are possible for all participants.

Most 'natural experiments' are therefore better described as '**quasi natural experiments**' or (more accurately) observational studies with **exogenous exposures**. With a little methodological massaging, these unique circumstances still provide an opportunity to study causal effects in a slightly different way to the 'conventional' approach described elsewhere. Exogenous exposures that provide an opportunity for 'natural experiment' include:

- 'Acts of God' (e.g. weather, climate, disasters); which involve 'no human agency'
- Geo-political events (e.g. war, famine, recession, Brexit)
- Government or policy changes (e.g. smoking ban, sugar tax, austerity)

- Other ‘exogenous’ changes (e.g. staff moved to new open-plan offices)

Interrupted time series

In many natural experiment scenarios, there may be no obvious external reference that can be used to compare and subtract away external trends.

If the outcome has been repeatedly measured with suitable intensiveness (i.e. constitutes **time-series** data), it may however be possible to determine and model any secular and seasonal trends from the pre-post data alone.

Interrupted time-series methods aim to provide improved pre-post counterfactual estimates by explicitly modelling the **autocorrelative** patterns in the outcome over time. Once these have been accounted for, it is more likely that any change in the outcome following an exposure/interruption can be attributed thereto.

There are many different methods for conducting interrupted time series analyses, but a popular classical method is an **autoregressive moving-average** (or **ARMA**) regression. Here, autoregressive (AR) and moving-average (MA) terms are iteratively identified and introduced by inspecting the autocorrelative patterns within the data (an approach known as the **Box-Jenkins method**).[78]

Once the autocorrelation has been removed, the instant and lagged effect of the exposure interruption can then be examined with a simple interaction between **time** ($\boldsymbol{T}$) and the **exposure** ($\boldsymbol{X}$):

$$Y_t = \beta_0 + \beta_1 \boldsymbol{T} + \beta_2 \boldsymbol{X}_t + \beta_3 \boldsymbol{T}\boldsymbol{X}_t + [ARMA\ terms] + \varepsilon_t$$

For interpretation:

- β_1 = the trend in the outcome over time, before the interruption;
- β_2 = the instantaneous effect of the exposure (also known as the **step change**); and
- β_3 = the lagged effect of the exposure (also known as the **slope change**).

Assumptions and caveats

Interrupted time series models assume that the pre- and post-interruption periods are conditionally exchangeable. In other words, after modelling the secular trends, seasonal trends, and any relevant confounders, there should be no other differences between the pre- and post-interruption period. This may be more reasonable over short intervals than over longer intervals, however longer intervals are often necessary to provide enough information about the autocorrelative patterns.

This leads onto the next and most significant limitation of interrupted time series: the need for intensive repeatedly-measured data, i.e. time-serial data. Unfortunately, such data are impractical or impossible in many situations, although they may be increasingly common thanks to automated data collection and smart devices.

As with all parametric methods, the accuracy of the estimate depends on having accurately modelled the autocorrelation structure.

Because of these problems, many readers will likely remain unconvinced of the validity of a simple pre-post interrupted time series. Increasingly, investigators are therefore encouraged to attempt to find, or construct, an external reference that is not exposed to the exposure interruption. One approach is to build a ‘synthetic’ control from external populations by taking a representative sample of units.

Regression discontinuity designs

Some natural experiment opportunities are created when an exposure is assigned mechanistically according to levels of some ‘**assignment variable**’. For example, in Peru free primary care (the exposure) is available to anyone below a certain wealth/poverty threshold (the assignment variable). Similarly, certain drugs (e.g. statins, the exposure) are prescribed when an individual crossed a certain biological threshold (e.g. LDL cholesterol above 4.9 mmol/L, the assignment variable).

Formally, this introduces a **positivity** violation, because the exposure is being determined by background variables that are likely to be related to the outcome. However, in theory, there is a **bandwidth** around the threshold where individuals are exchangeable since whether they fall above or below the threshold will be entirely down to random variation (such as measurement error).

To estimate the effect of the exposure, you could therefore directly compare the outcomes of those who fall *just* below the threshold and those who fall *just* above the threshold. The problem is that this requires an extremely large sample size to have sufficient individuals who fall just either side of the threshold. As a compromise, you could increase the bandwidth to include more people on either side, but this stretches the assumption of exchangeability.

A **regression discontinuity design** offers a superior solution, by relaxing the assumption for unconditional exchangeability to permit conditional exchangeability. They allow the relationship between the assignment variable and the outcome to be modelled explicitly. If the relationship between the assignment variable and the exposure is perfectly determined by the threshold then the discontinuity is described as '**sharp**'. If the relationship between the assignment variable and the exposure is imperfect, then the discontinuity is described as '**fuzzy**'. Fuzzy situations occur when protocols are not followed strictly, e.g. if an individual with an LDL cholesterol level over 4.9 mmol/L does not get prescribed a statin. Failure to account for a fuzzy threshold will bias the effect towards the null. To resolve this, the relationship between the assignment variable and the propensity of exposure should be modelled overall; this then acts as a weighting variable in the regression model.

A simple regression discontinuity equation, with no confounding, can be written as follows, where:

- $\boldsymbol{X}$ = propensity of exposure (e.g. statin);
- $\boldsymbol{Z}$ = **assignment variable** (e.g. LDL cholesterol); and
- $\boldsymbol{T}$ = **threshold value** (e.g. 4.9 mmol/L).

$$Y = \beta_0 + \beta_1\boldsymbol{X} + \beta_2(1 - \boldsymbol{X})(\boldsymbol{Z} - \boldsymbol{T}) + \beta_3\boldsymbol{X}(\boldsymbol{Z} - \boldsymbol{T}) + \varepsilon$$

For interpretation:

- β_1 = effect of the exposure (e.g. statin);
- β_2 = effect of the assignment variable below the threshold (e.g. LDL cholesterol concentration under 4.9 mmol/L); and
- β_3 = effect of the assignment variable above the threshold (e.g. LDL cholesterol concentration ≥4.9mmol/L).

Assumptions and caveats

Regression discontinuity designs assume that the units observed above and below the threshold are (conditionally) exchangeable. In other words, after modelling the effect of the assignment variable (and any relevant confounders) the exposure should be effectively assigned 'as random'. In addition to unobserved confounding, this will be violated if there are non-random selection processes operating e.g. if doctors are more likely to repeat cholesterol tests with 'borderline' results.

The estimate may be sensitive to the choice of bandwidth; as the exchangeability assumption will become increasingly stretched with wider bandwidths. In most situations, however, a wider bandwidth is necessary to ensure sufficient power and precision.

As with all parametric methods, the accuracy of the effect estimate is dependent on having accurately modelled the propensity of exposure and the relationship between the assignment variable and the outcome.

Further reading: Natural Experiments;[79] Instrumental Variables;[80] Mendelian Randomisation;[81] Difference-in-Differences;[82] Interrupted Time Series;[83] and Regression Discontinuity Designs.[84]

DAY 4

4.1 Exposure Regimes, Causal Mediation, & 'Interactions'

Learning objectives

- Understand the limitations of traditional approaches to mediation analyses
- Understand how counterfactual methods can theoretically be used to examine the distinct joint effects of exposures, mediators, or time varying exposures
- Understand why redefining the mediator as the exposure is often the best solution

Why you might want to adjust for mediators

We have learnt about the dual problems of adjusting for mediators: 1) conditioning on mediators will partition part of the total causal effect of the exposure; and 2) in the presence of intermediate confounding this will risk introducing collider bias. For these reasons we would generally advise against conditioning on mediators or interpreting coefficients in models where mediator adjustment is present. There are however some circumstances where the exposure is very difficult to influence directly (e.g. socioeconomic circumstances), and we may be interested in discovering to what extent it is still possible to intervene later down the causal path, and how much benefit this may have. Similarly, we may be interested in unpicking the temporal effects of a **time varying exposure** measured on several occasions, perhaps to identify how later changes can undo earlier effects? Finally, we might hypothesise that the exposure has multiple, distinct effects that we wish to unpick.

Traditional mediation analyses

Since the 1980s, questions like these have traditionally been dealt with using **path analyses** or **structural equation modelling** approaches, which are increasingly available within routine statistical software.[85] These methods aim to partition the **total (causal) effect** into the effect that is mediated through one or more later variables (the '**indirect effect**') and the effect that acts separately from the included mediators variables (the '**direct effect**').

The approach has several assumptions, although an increasing number of workarounds are now available. First, they traditionally assumed **multivariate normality**, although many packages now allow generalised parameterisations. Second, they assume linearity, although you could add your own non-linear parameterisations. Third, they assume that the exposure(s) and mediator(s) have purely additive effects, i.e. the total effect of the exposure should be the simple sum of the direct effect (independent of the mediator) + the indirect effect (through the mediator).

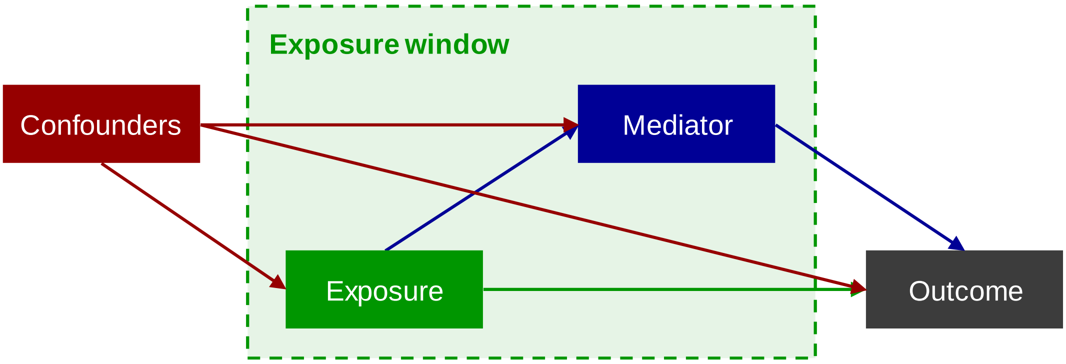
This third assumption is particularly restrictive, as it forbids any interaction between the exposure and mediator. The most obvious solution is to consider whether you really need to understand the **joint effects**, or whether it is sufficient to shift your focus to the 'mediator' – in effect making it your true exposure of interest. You can then easily stratify your analyses across levels of the original exposure and interpret the strata-specific causal effect of the mediator accordingly.

For example, you might be interested in the effect of **socio-economic circumstances** on **cancer survival**, mediated through **engagement with healthcare services**. You observe that the effect of engagement with healthcare services varies substantially by socio-economic circumstances. You therefore decide that your real focus is on the effect of engagement with healthcare services and prioritise estimating this effect in different strata of socio-economic circumstances.

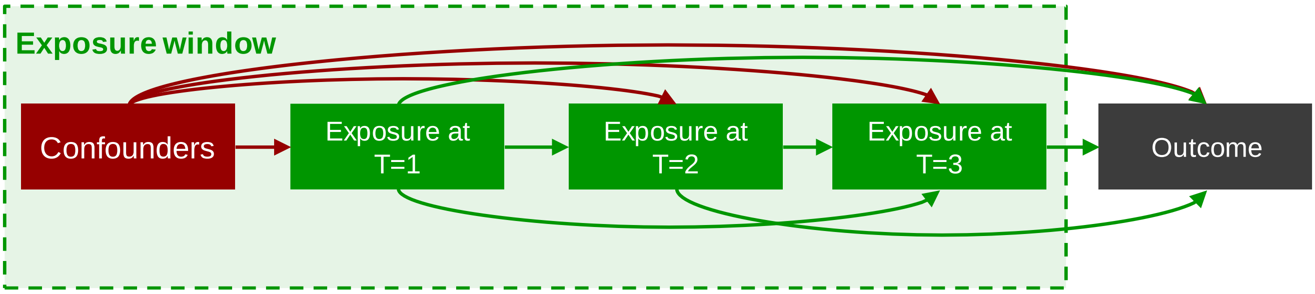
This approach does not however provide you with a formal estimate of the **joint effects** of the exposure and mediator on the outcome. As an alternative, you could include an interaction term, but these are not especially easy to interpret, and it is equally unclear how you would then partition this into the 'direct' and 'indirect' effects.

Counterfactual mediation analyses

A counterfactual approach offers a simple solution to conducting mediation analyses in the presence of complex parameterisation such as interactions. Rather than considering each of the exposure(s) and outcome(s) as separate variables, you extend your **exposure window** to consider the overall exposure pattern.



Participants are now classified by their overall **exposure regime** according to their pattern of exposure(s) and mediator(s) values. For example, in the very simple case where you have a binary exposure (E) and binary mediator (M), there are three obvious regimes of interest: (E=0, M=0) where the exposure and mediator are absent, (E=1, M=0), where the exposure is present but the mediator is absent, and (E=1, M=1), where the exposure and mediator are both present. This can easily be extended to include multiple exposures and/or mediators (provided all confounding occurs before the first exposure; known as **time-invariant confounding**), although the number of possible regimes will increase dramatically with an increasing number of variables.



Going back to the simple exposure and mediator scenario, to determine the '**controlled direct effect**' of the exposure you would now simply compare the **potential outcomes** between the exposure regimes where the exposure is present (E=1, M=0) and the alternative regime where the exposure is absent (E=0, M=0), fixing the mediator to be absent (hence '**controlled**'). But this is not the only potential 'direct effect' of interest. If your mediator is continuous, or you are interested in group level effects (i.e. not individuals), you might be interesting in the 'direct effect' of the exposure in the presence of 'typical' levels of the mediator because it might be very unusual for the mediator to be absent. The '**natural direct effect**' would thus contrast (E=1, M=m($_{E=1)}$, i.e. the average value of the mediator in the presence of the exposure) with (E=0, M=m$_{(E=1)}$). With a continuous exposure, you will also have to make a choice about what sort of change you are interesting in comparing. You might choose 1 unit, or 1 standard deviation, or some other difference that you deem meaningful (such as a 10% decrease in the exposure). One of the strongest aspects of counterfactual analyses of this nature is precisely that it encourages you to consider meaningful comparisons, rather than necessarily focussing on single-unit regression coefficients.

Vanderwheele's Four Way Decomposition

It might be obvious from the existence of more than one direct effect – even for the simple binary exposure and binary mediator scenario – that there are a large number of potential effects that you could chose to calculate, and these quickly balloon once you move beyond from such simple scenarios (comprising a single binary exposure and a single binary mediator) to include continuous variables and an increase in the number of these.

If you are particularly interested in partitioning your total effect into different components, the best approach is probably Tyler Vanderwheele's **'four-way decomposition'**,[86] which describes the joint effects of exposure and mediator in terms of:

- **Controlled direct effect,** the effect of the exposure in the absence of the mediator;
- **Pure indirect effect**, the effect of the exposure that acts purely through the mediator;
- **Reference interaction,** the changing 'direct' effect of the exposure for changing values of the mediator; and
- **Mediated interaction,** the changing mediated effect of the exposure for changing values of the mediator.

The specific contrasts are described in Vanderwheele's 2014 paper with add-on commands in R, SAS, and Stata[86] (for further reading see[51,87]). Their utility however depends on your ability to make sense of the reference interaction and mediated interaction!

Causal interactions and effect modification

Vanderweele's four way interaction is part of a wider interest in examining '**causal interactions**' and '**effect modification**' in observational data. Causal interactions are non-linear interactions between elements of your exposure regime (e.g. the exposure and mediator), meaning that they do not produce additive effects. Effect modification is non-linear interactions between other variables and one or more element of your exposure regime (e.g. between a confounder and your exposure).

Caution is advised for those interested in identifying and studying causal interactions, because they can easily be confused with statistical interactions. Statistical interactions are required in statistical models when there appear to be non- multiplicative relationships between two or more model covariates. The apparent presence and benefit of statistical interactions do not necessarily imply causal interactions for two important reasons. First, statistical interactions may arise when there is unmodelled confounding or unmodelled non-linear effects (e.g. if your exposure has a curvilinear effect on your outcome, but you have not modelled this). Second, statistical interactions are scale dependent. If you model your outcome as a continuous variable using linear regression, an interaction term may be necessary. However, if you transform the same variable into a binary variable, and model this using logistic regression, no interaction term may be necessary. The reverse is also true, where there is no apparent interaction in the linear model, but an interaction appears for the binary model. Finding and interpreting these interactions as indicative of underlying causal mechanisms is therefore somewhat controversial and should be approached cautiously.

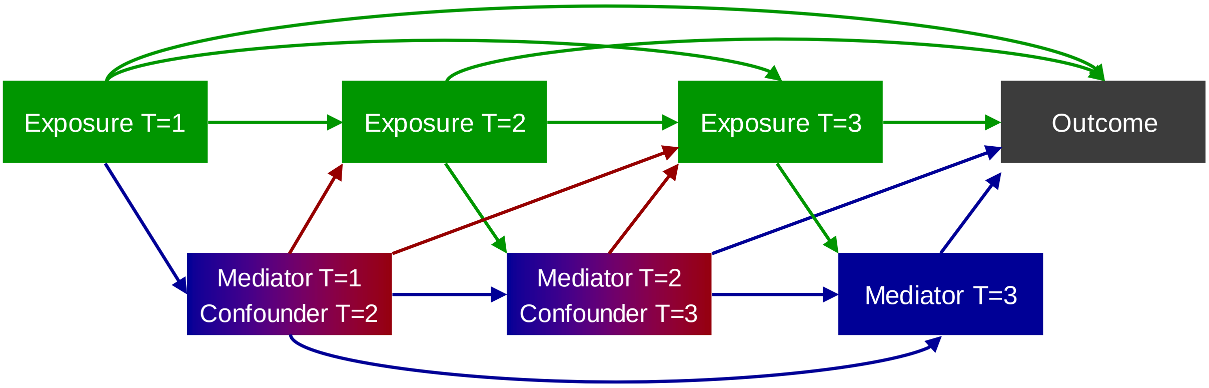
Positivity

In the simple binary exposure and binary mediator example above, we mentioned that you might want to choose a non-zero value of the mediator if it is unlikely to ever be absent. This aptly highlights one of the three core 'identifiability' assumptions of causal inference methods, known as the positivity criteria. Essentially, it must be *possible* for every participant to receive every treatment regime. Thus, if certain exposure values mean you cannot have certain mediator values, then you clearly should not try to estimate these scenarios! This criterion extends to your confounders. If a certain combination of confounding values determines the range of available exposure or mediator values, then you have a positivity violation. Recall that you are trying to achieve conditional exchangeability; thus, conditional on your confounders, your exposure should be assigned 'as random' between exchangeable units.

Intermediate confounding and time-varying confounding

So far, we have considered the scenario where all confounding occurs before the (first) exposure. However, if your context includes **intermediate confounding**, then classical methods may be unsuitable. Recall that intermediate confounding occurs when your mediator is caused by a variable that also causes the outcome and presents a risk of collider bias if you adjust for the mediator. In a mediation analysis, you are deliberately adjusting for the mediator, so your estimate of the causal effect of the exposure is subject to collider bias. If the intermediate confounder it is not caused by the exposure, then measuring the confounder and conditioning for it should correct the collider bias. Here, it is not really acting as a true 'intermediate' confounder but is simple an ordinary time-invariant confounder that has crystallised before the exposure. But if the intermediate confounder is caused by the exposure, then there is no solution within classical methods, since adjusting for the variable will block part of the true causal effect of the exposure.

This is a risk with time varying exposures, because any variable that is caused by the exposure at one point in time is likely to cause the exposure at future time points. **Time-varying exposures** of this nature are common in the contexts of longitudinal data. If you are then interested in understanding the joint effect of the exposure across multiple time points, there are no conditioning solutions, so you will need to use **g-methods** as described in the next section.



Summary

Traditional methods for mediation analyses (as implemented in traditional SEM software and packages) assume an absence of interactions. The easiest option is to reframe your question and consider carefully whether you are truly interested in understanding the joint effects of the exposure and mediator, or whether analysing the mediator alone may be sufficient.

Counterfactual methods provide a possible solution by encouraging/requiring us to extend our ‘exposure window’ to cover all our exposures (and mediators) of interest. Bespoke causal effects can then be estimated by comparing potential outcomes between different exposure regimes.

If you wish to break down your total effect into different parts, Vanderweele’s four-way decomposition offers one of the simplest summaries, although effects and decompositions are available. Mediation methods unfortunately fall down in the presence of intermediate confounding and time-varying confounding, as it becomes impossible to keep all causal paths open, while controlling for confounding and collider bias. In such circumstances, g-methods need to be used.

4.2 Introduction to G-methods

Learning objectives

- Understand how the three 'identifiability conditions' may be used to estimate causal effects from longitudinal observational data
- Become familiar with the three 'g-methods', and how they may be used to estimate the causal effects of time-varying exposures

The 'G-methods'

The g-methods are a family of methods that include the g-computation algorithm formula (the '**g-formula**'), **inverse probability of treatment weighting (IPTW)** of marginal structural models (MSMs), and **g-estimation** of structural nested models (SNMs). The 'g' in g-methods stands for generalised, because they may be used to estimate the causal effect of any form of hypothetical intervention without introducing bias due to inappropriate adjustment for time-dependent confounding.[1] There is no prohibition on using these methods for scenarios involving time-invariant exposures, but they are theoretically and computationally more challenging than traditional stratification/regression methods. Therefore, these methods are usually reserved for scenarios involving time-varying exposures, in which traditional methods result in bias. The true power of the g-methods lies in their ability to estimate the causal effects of time-varying exposures *even in the presence of time-dependent confounding*.

To demonstrate each of the three g-methods, we will use an example scenario involving metformin treatment and fasting blood glucose, which has been adapted from Naimi, et al.[2] as outlined in the following section. We also introduce the three 'identifiability conditions', which allow us to estimate causal effects and are integral in demonstrating each of the g-methods.

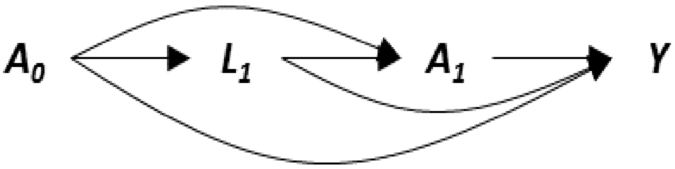
Example scenario

Type 2 diabetes is a condition in which a person's blood glucose levels become too high due to either insufficient insulin production within their body or the cells in their body failing to respond to insulin.[3] Due to the considerable costs associated with treating and managing diabetes once it has developed, public health practitioners are increasingly advocating early interventions to prevent the onset of diabetes. One such intervention that has been proposed is treatment with metformin amongst individuals with 'prediabetes'. Whilst a fasting blood glucose level below 5.6 mmol/L is considered normal and above 7.0 mmol/L indicates diabetes, anything between 5.6 and 7.0 mmol/L is considered indicative of prediabetes. Treatment with metformin to prevent diabetes onset is not widespread, though might be implemented more widely if it could be demonstrated that such treatment is effective at preventing fasting blood glucose levels from reaching the diabetic threshold.

To study this, we collect data from an observational cohort consisting of 100,000 newly-diagnosed prediabetic individuals considered to be at 'high risk' for developing diabetes because their fasting blood glucose levels are severely elevated (>6.5 mmol/L). Treatment with metformin ($A = 1$ for treatment, $A = 0$ otherwise) may be prescribed upon initial diagnosis (A_0) or at a six-month follow-up appointment (A_1). The sole covariate (L_1) is whether or not the individual remains 'high risk' at six-month follow-up ($L_1 = 1$ for 'high risk', $L_1 = 0$ otherwise), as this is affected by whether they received treatment at baseline (A_0), and in turn affects whether they receive treatment at six-month follow up (A_1). The outcome (Y) is fasting blood glucose level at 1-year follow-up.

This scenario, depicted by the DAG in Figure 4.2.1, represents the presumed data-generating process.

Figure 4.2.1: DAG representing the presumed relationship amongst treatment with metformin at baseline (A_0), 'high risk' status at six-months follow-up (L_1), treatment with metformin at six-months follow-up (A_1), and fasting blood glucose level at 1-year follow-up (Y).



For simplicity, we assume: no other confounders of the relationship between metformin treatment and fasting blood glucose; full compliance with the prescribed treatment; and no loss to follow-up over the study period.

A summary of the data is provided in Table 4.2.1, with the outcome fasting blood glucose level at 1-year follow-up (Y) averaged over the individuals for each combination of treatment and confounder, and the number of participants for each combination (N) recorded in the rightmost column.

The data from Table 4.2.1 are presented as a decision tree[4] in Figure 4.2.2, in which all 100,000 individuals start at the left and progress over time towards the right. All individuals begin as 'high risk' at baseline (i.e. $L_0 = 1$) and are prescribed treatment with metformin ($A_0 = 1$) with a probability of 0.401. At their six-month follow-up appointment, individuals are assessed to determine whether they remain 'high risk' ($L_1 = 1$); among those who did or did not receive metformin at baseline, the probability of being 'high risk' at six months is 0.861 or 0.983, respectively. This, in turn, affects the probability of receiving metformin treatment at six months. The mean value of fasting blood glucose levels at 1-year follow-up (Y) amongst those individuals defined by each combination of treatment and confounder are given at the ends of each branch.

Table 4.2.1: Summary data for 100,000 initially 'high-risk' prediabetic individuals illustrating the number of individuals (N) within each possible combination of metformin treatment at baseline (A_0), 'high risk' status at six-month follow-up (L_1), and metformin treatment at six-month follow-up (A_1). The outcome column Y corresponds to the mean fasting blood glucose (in mmol/L) level at 1-year follow up within levels of A_0, L_1, A_1. Note that 'high risk' status (L_0) at baseline is true for everyone by design (i.e. $L_0 = 1$).

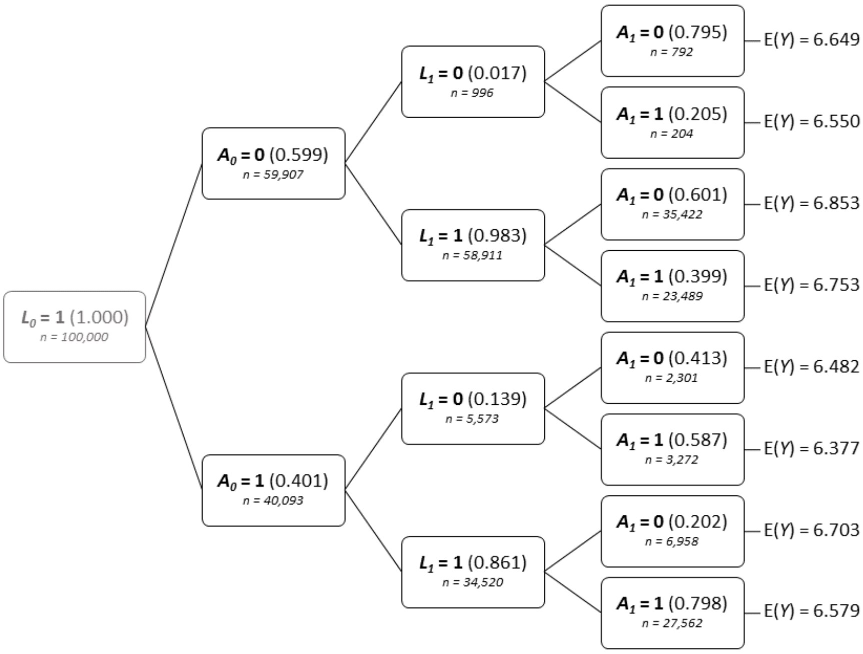
A_0	L_1	A_1	$E(Y)$	N
0	0	0	6.649	792
0	0	1	6.550	204
0	1	0	6.853	35,422
0	1	1	6.753	23,489
1	0	0	6.482	2,301
1	0	1	6.377	3,272
1	1	0	6.703	6,958
1	1	1	6.579	27,562

The *average potential outcome* for exposure/treatment regime $a = (a_0, a_1)$ is denoted $E(Y^{a_0,a_1})$, where the lowercase letters a_0 and a_1 are used to denote specific values of the random variables A_0 and A_1. We seek an estimate of the total causal effect (TCE) of always receiving treatment (i.e. $a = (1,1)$) compared to never receiving treatment (i.e. $a = (0,0)$). Thus, we define the TCE to be:

$$TCE = E(Y^{1,1}) - E(Y^{0,0}) = E(Y^{1,1} - Y^{0,0})$$

The TCE represents an average of the difference in outcomes that *would be observed* had *all* individuals been prescribed treatment with metformin at both baseline and six-month follow-up versus had *no* individuals been prescribed treatment with metformin at either baseline or six-month follow-up. The TCE is considered to be a *marginal effect*, as it averages (or marginalises) over all individual-level effects in the population.[2]

Figure 4.2.2: The data from Table 4.2.1 represented as a decision tree. Individuals move from left to right along branches; the probability of moving along a particular branch is given in parentheses, and the number of individuals on each branch at each point is given below (n).



Unfortunately, those who actually received metformin at both time points are not exchangeable with those who did not. From the DAG in Figure 4.2.1, we can see that the population begins as (unconditionally) exchangeable because there is no confounding of the relationship between A_0 and Y. However, at six months follow-up, those who receive treatment are no longer exchangeable with those who do not, because treatment at that time depends on current ‘high risk’ status (i.e. those receiving treatment at six months follow-up are more likely to be ‘high risk’ than those who are not). To create (conditional) exchangeability at six months follow-up, adjustment for ‘high risk’ status (L_1) is necessary; however, this is a mediator for the relationship between A_0 and Y, and traditional methods of adjustment for L_1 will bias the estimated relationship between them.

Identifiability conditions

Under the three so-called ‘identifiability conditions’, we are able to estimate (i.e. *identify*) the total causal effect – which is a function of *counterfactual quantities* – using observational data. We introduced one of the identifiability conditions (i.e. exchangeability) in a previous Lecture. Here we briefly introduce the other two and provide more formal definitions of all three.

Together, the three conditions imply that an observational study can be conceptualised as a sequentially randomised experiment, which is a randomised experiment in which the exposure value at each time point t (i.e. A_{t}) is randomly assigned with known randomisation probabilities that, by design, may depend on an individual’s past exposure ($\bar{A}_{t-1}$) and confounder history ($\bar{L}_t$) through time t (overbars denote history).[5] For our example scenario, this means that under the three identifiability conditions we can operate as if treatment with metformin were randomised both: (1) at baseline (A_0); and (2) at six months follow-up (A_1), with a probability that may depend on A_0 and L_1 (*but not on any unobserved factors*, in which case there would exist unmeasured confounding).

Consistency

The first identifiability condition is **consistency**. Consistency states that for an individual who received a particular exposure regime $a = (a_0, a_1)$, their potential outcome Y^{a_0,a_1} is equal to their observed outcome Y and is therefore known (though their other potential outcomes remain unknown). This assumption allows us to equate the observed outcomes amongst those who received a particular exposure regime with the counterfactual outcomes that would be observed for them under the same exposure regime. For our example scenario, this assumption is translated as:

$$Y|A_0 = a_0, A_1 = a_1 \quad = \quad Y^{a_0,a_1}|A_0 = a_0, A_1 = a_1$$

Because there are two time points at which each individual may or may not receive treatment with metformin in our example, each individual has four potential outcomes $Y^{0,0}$, $Y^{0,1}$, $Y^{1,0}$, and $Y^{1,1}$. An individual who received treatment with metformin at baseline ($A_0 = 1$) but not at follow-up ($A_1 = 0$), for example, would have three unknown counterfactual outcomes (i.e. $Y^{0,0}$, $Y^{0,1}$, and $Y^{1,1}$) and one known counterfactual outcome $Y^{1,0}$ which is equal to their observed outcome.

Conditional exchangeability

The second identifiability condition is (sequential) **conditional exchangeability**, which states that the potential outcome Y^{a_0,a_1} associated with the particular exposure regime $a = (a_0, a_1)$ is independent of (i.e. not affected by) the exposure regime $A = (A_0, A_1)$ that was actually observed. That is, the potential outcome for any given exposure regime will be the same, regardless of the exposure regime that was actually received. As different exposure-outcome combinations may be affected by different sets of confounders, we make this exchangeability assumption within all levels of covariate history $\bar{L}_t$ (i.e. *conditional* on $\bar{L}_t$) and at each time point separately (i.e. *sequential*). For our example scenario, this can be expressed formally as:

$$\begin{aligned} Y^{a_0,a_1}|A_0 \quad &= \quad Y^{a_0,a_1} \\ Y^{a_0,a_1}|A_1, L_1, A_0 \quad &= \quad Y^{a_0,a_1}|L_1, A_0 \end{aligned}$$

for baseline and six months follow-up, respectively. Alternately, we can express this assumption as:

$$\begin{gathered} Y^{a_0,a_1} \coprod A_0 \\ Y^{a_0,a_1} \coprod A_1 | L_1, A_0 \end{gathered}$$

where $\coprod$ denotes independence. The conditional exchangeability assumption implies that those who received treatment at a particular time point would, had they not received treatment at that time point, experience the same distribution of outcomes as those who actually did not receive the treatment at that time point (i.e. given that they had the same exposure and covariate history up until that point). Conditional exchangeability is often referred to as the assumption of *no unmeasured confounding*.

Positivity

The third identifiability condition is **positivity**. Positivity is the requirement that the exposure at each time point t (i.e. A_t) was not deterministically allocated within any level of past exposure ($\bar{A}_{t-1}$) and covariate history ($\bar{L}_t$) through time t. In other words, this condition states that there is a non-zero chance of being exposed (or unexposed) at every time point, regardless of prior exposure and confounder history. Thus, positivity is satisfied when there are both exposed and unexposed individuals within all levels of prior exposure and of each confounder, which can be easily evaluated empirically (for categorical variables, at least). For our example scenario, we can see that there are both treated and untreated individuals at baseline, and that there are again both treated and untreated individuals at six months follow-up within all levels of A_0 and L_1 (Table 4.2.1, Figure 4.2.2); thus, positivity is satisfied.

The g-formula

We first provide a derivation of the g-formula, and then demonstrate its implementation.

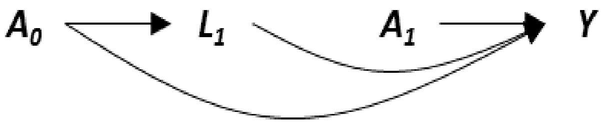
To begin, we factor the joint probability density function $f(\cdot)$ of our observed data in a way that respects the temporal ordering of the variables (as depicted in the DAG in Figure 4.2.1) by conditioning each variable on its history (i.e. employing the 'chain rule'):[6]

$$f(y, a_1, l_1, a_0) = f(y|a_1, l_1, a_0) \cdot P(A_1 = a_1|L_1 = l_1, A_0 = a_0) \cdot P(L_1 = l_1|A_0 = a_0) \cdot P(A_0 = a_0)$$

We can further factor the joint density by representing the conditional mean of Y (i.e. $f(y|a_1, l_1, a_0)$) in terms of its conditional density, and use the law of total probability to marginalise over the distribution of A_1, L_1, and A_0 to produce the marginal mean of Y:

$$E(Y) = \sum_{a_1,l_1,a_0} E(Y|A_1 = a_1, L_1 = l_1, A_0 = a_0)P(A_1 = a_1|L_1 = l_1, A_0 = a_0)P(L_1 = l_1|A_0 = a_0)P(A_0 = a_0)$$

Figure 4.2.3: Modified DAG from Figure 4.2.1, after intervening to set $A_0 = a_0$ and $A_1 = a_1$.



This equation represents the average value of the outcome (Y) in the population as a weighted sum of the average outcome amongst the subgroups defined by all possible combinations of A_0, L_1, and A_1. Because we are interested in the marginal mean of Y that would be observed in the population if we *intervened* to set the exposures A_0 and A_1 to some values a_0 and a_1, respectively, we are able to drop the terms $P(A_1 = a_1|L_1 = l_1, A_0 = a_0)$ and $P(A_0 = a_0)$ from the equation – the probabilities of A_0 and A_1 taking on the values a_0 and a_1 are now equal to one. By intervening (even if only hypothetically), we have, in effect, eliminated the ability of A_0 and A_1 to 'naturally' respond to other variables in the DAG which cause them.

Our 'intervention' produces the modified DAG in Figure 4.2.3, in which no time-dependent confounding is present. By the consistency assumption, we are able to equate the observed mean (Y) with the counterfactual mean (Y^{a_0,a_1}), and our marginal mean equation may therefore be written as:

$$E(Y^{a_0,a_1}) = \sum_{l_1} E(Y|A_1 = a_1, L_1 = l_1, A_0 = a_0)P(L_1 = l_1|A_0 = a_0)$$

This equation is the **g-formula**, which represents the expected value of the potential outcome Y^{a_0,a_1} that *would be observed* post-intervention as a weighted sum of the average *observed* outcome for those who, in fact, received exposures a_0 and a_1.

To illustrate, we use the metformin treatment data from Table 4.2.1 to calculate the average value of the potential outcomes that would be observed were we to intervene to set both A_0 and $A_1$to one (i.e. force everyone to receive treatment with metformin at baseline and six-month follow-up):

$$\begin{aligned}\boldsymbol{\hat{E}(Y^{1,1})} &= \sum_{l_1} E(Y|A_1 = 1, L_1 = l_1, A_0 = 1)P(L_1 = l_1|A_0 = 1) \\ &= E(Y|A_1 = 1, L_1 = 0, A_0 = 1)P(L_1 = 0|A_0 = 1) + E(Y|A_1 = 1, L_1 = 1, A_0 = 1)P(L_1 = 1|A_0 = 1) \\ &= (6.377)\left(\frac{2{,}301 + 3{,}272}{2{,}301 + 3{,}272 + 6{,}958 + 27{,}562}\right) + (6.579)\left(\frac{6{,}958 + 27{,}562}{2{,}301 + 3{,}272 + 6{,}958 + 27{,}562}\right) \\ &= (6.377)(0.139) + (6.579)(0.861) = \mathbf{6.551}\end{aligned}$$

We can also calculate the average value of the potential outcomes that would be observed were we to intervene to set both A_0 and A_1 to zero (i.e. prohibit anyone from receiving treatment with metformin at either baseline or six-month follow-up).

$$\begin{aligned}\boldsymbol{\hat{E}(Y^{0,0})} &= \sum_{l_1} E(Y|A_1 = 0, L_1 = l_1, A_0 = 0)P(L_1 = l_1|A_0 = 0) \\ &= E(Y|A_1 = 0, L_1 = 0, A_0 = 0)P(L_1 = 0|A_0 = 0) + E(Y|A_1 = 0, L_1 = 1, A_0 = 0)P(L_1 = 1|A_0 = 0) \\ &= (6.649)\left(\frac{792 + 204}{792 + 204 + 35{,}422 + 23{,}489}\right) + (6.853)\left(\frac{35{,}422 + 23{,}489}{792 + 204 + 35{,}422 + 23{,}489}\right) \\ &= (6.649)(0.017) + (6.853)(0.983) = \mathbf{6.850}\end{aligned}$$

Thus, our estimate of the total causal effect (TCE) of metformin treatment among 'high-risk' prediabetic individuals is calculated to be:

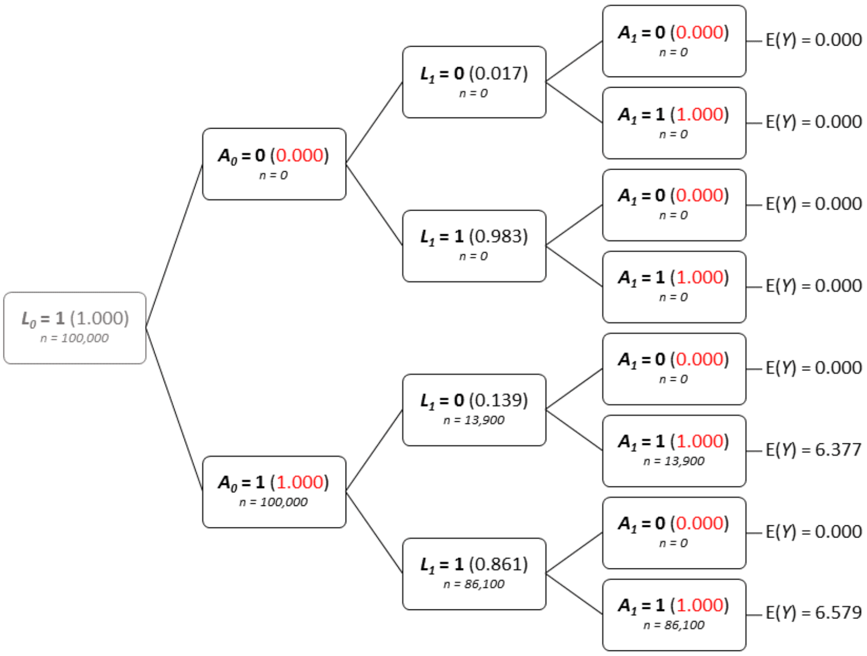
$$\widehat{TCE} = \hat{E}(Y^{1,1}) - \hat{E}(Y^{0,0}) = 6.551 - 6.850 = \mathbf{-0.299}$$

Had all individuals in our study received treatment with metformin at both baseline and six-month follow-up, we would expect the average level of their fasting blood glucose to be 0.299 mmol/L lower than if they had not received treatment with metformin at either time point.

By weighting the conditional expectation of the observed outcome (i.e. $E(Y|A_1 = a_1, L_1 = l_1, A_0 = a_0)$) by the conditional probability of the confounder (i.e. $P(L_1 = l_1|A_0 = a_0)$), we have accounted for the fact that L_1 simultaneously mediates the effect of A_0 on Y whilst confounding the effect of A_1 on Y^2 *without explicitly adjusting for* L_1 (as in standard methods of analysis). The weighting used in our calculations creates conditional exchangeability between those who received treatment at both time points and those who didn't receive treatment at either time point; we have accounted for the fact that those receiving treatment at six months are more likely to be 'high risk' than those not receiving treatment. In the absence of weighting, the two groups would not be exchangeable.

We can also demonstrate implementation of the g-formula more intuitively using decision trees. To estimate the effect of any (hypothetical) intervention, we can alter the tree for our observed data (Figure 4.2.1) by setting the probabilities of exposure status to those that we desire (Figures 4.2.4 and 3.3.5).

Figure 4.2.4: Modified decision tree from Figure 4.2.2, in which the probability of receiving treatment at both baseline and follow-up is set to one (i.e. $P(A_0 = A_1 = 1) = 1.000$).



For example, to estimate the mean counterfactual outcome $E(Y^{1,1})$, we set the probabilities of $A_0 = 1$ and $A_1 = 1$ to 1.000 (i.e. everyone receives treatment at baseline and six-month follow-up with a probability equal to one) and calculate the expected number of individuals for each branch, as demonstrated in Figure 4.2.4. The estimated mean counterfactual outcome under 'treatment' intervention is thus given by:

$$\hat{\boldsymbol{E}}(\boldsymbol{Y}^{\mathbf{1,1}}) = (6.377)\left(\frac{13{,}900}{100{,}000}\right) + (6.579)\left(\frac{86{,}100}{100{,}000}\right) = \mathbf{6.551}$$

As expected, this is equivalent to the estimate that was produced previously by the g-formula.

Similarly, to estimate the mean counterfactual outcome $E(Y^{0,0})$, we set the probabilities of $A_0 = 0$ and $A_1 = 0$ to 1.000 (i.e. no one receives treatment at baseline or six-month follow-up with probability equal to one), as in Figure 3.3.5. The estimated mean counterfactual outcome under 'no treatment' intervention is given by:

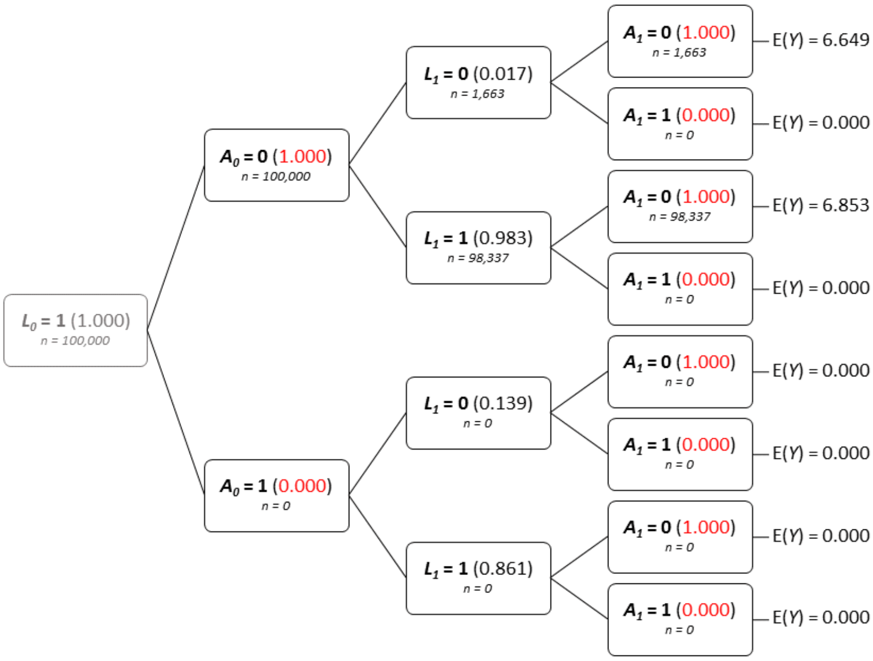
$$\widehat{\boldsymbol{E}}(\boldsymbol{Y}^{\mathbf{0,0}}) = (6.649)\left(\frac{1{,}663}{100{,}000}\right) + (6.853)\left(\frac{98{,}337}{100{,}000}\right) = \mathbf{6.850}$$

which again is equivalent to the estimate that was previously produced by the g-formula, and our estimate of the TCE of metformin treatment among 'high risk' prediabetic individuals is:

$$\widehat{TCE} = \hat{E}(Y^{1,1}) - \hat{E}(Y^{0,0}) = 6.551 - 6.850 = \mathbf{-0.299}$$

Thus, the g-formula functions by estimating the expected value of the counterfactual outcome $E(Y^{a_0,a_1})$ by simulating the joint distribution of the variables that *would have been observed* in a hypothetical scenario in which every individual received exposures $A_0 = a_0$ and $A_1 = a_1$, based on the joint distribution of the variables that was actually observed.[5]

Figure 3.3.5: Modified decision tree from Figure 4.2.2, in which the probability of receiving treatment at both baseline and follow-up is set to zero (i.e. $P(A_0 = A_1 = 0) = 1.000$).



We can see clearly the importance of the three identifiability conditions in this method. The consistency assumption allows us to equate the distribution of observed outcomes among those who received a particular exposure regime (e.g. $a = (1,1)$) with the counterfactual outcomes that would have been observed for them under that exposure regime. Moreover, by conditional exchangeability, we are able to draw the conclusion that had *everyone* been exposed to the same regime, they would experience the same distribution of outcomes as the subset of individuals who actually were exposed. Lastly, the positivity condition ensures that we are, in fact, able to 'exchange' individuals receiving differing exposure regimes; if the probability of receiving treatment at any time point were zero amongst any subgroup of prior exposure and confounder history, then it would not be valid to exchange those who did not receive treatment (i.e. everyone) with those who did (i.e. no one).

Inverse probability of treatment weighting (IPTW) of marginal structural models (MSMs)

The g-formula requires that we explicitly model each component of the joint density of A_0, L_1, A_1, and Y, which can lead to bias if any of these models are mis-specified. Marginal structural models (MSMs) avoid this potential problem by using a single model to target the causal effect of interest[2]. To accommodate time-varying confounding, inverse probability of treatment weighting (IPTW) is used to create a 'pseudo population' in which treatment at each time point depends only on prior treatment history and not on values of the time-dependent confounder (i.e. as would be expected if it had, in fact, been sequentially randomised based only on prior exposure history).

Inverse probability of treatment weighting (IPTW): Creating the 'pseudo population'

To create the pseudo population, we calculate the (conditional) probability of each individual receiving the treatment that they actually received, and weight them by the inverse of that probability (hence, **'inverse probability of treatment weighting'**). The weight (w) for each individual is given by:

$$w = \prod_{t=0}^{T} \frac{1}{P(A_t|\bar{A}_{t-1}, \bar{L}_t)}$$

The denominator of the expression above represents *the probability of receiving the treatment actually received at time* t, given exposure and confounder history up to time t (denoted using overbars). Thus, for our scenario, the weight for each individual can be calculated as:

$$w = \frac{1}{P(A_0)} \cdot \frac{1}{P(A_1|A_0, L_1)}$$

For example, the weight for an individual who did not receive metformin treatment at baseline ($A_0 = 0$), was not classified as 'high risk' at six months follow-up ($L_1 = 0$), and did not receive metformin treatment at six months follow-up ($A_1 = 0$) is calculated to be:

$$w = \frac{1}{P(A_0 = 0)} \cdot \frac{1}{P(A_1 = 0|A_0 = 0, L_1 = 0)} = \frac{1}{0.599} \cdot \frac{1}{0.795} = 2.099$$

Thus, each individual with this combination of A_0, L_1, and A_1 counts for 2.099 individuals in the pseudo population, and the total number of individuals with $A_0 = 0$, $L_1 = 0$, and $A_1 = 0$ in the pseudo population is equal to $792 * 2.099 = 1{,}663$. The weights (w) for all combinations of A_0, L_1, and A_1 are given in Table 4.2.2, as are the total number of individuals in the pseudo population (*Pseudo* N) for each combination.

As is evident from Table 4.2.2, the total number of individuals in the pseudo population created by weighting by w is equal to 400,000, which is quadruple the number of individuals in the original, unweighted population. Weighting the original population by the inverse probability of treatment essentially simulates what *would have happened* had all individuals simultaneously received *and* not received metformin treatment at both baseline and six-month follow-up;[5] thus, each individual has four potential outcomes (one of which is, in fact, observed). This process can be demonstrated using a decision tree, as in Figure 3.3.6, and is correct under the assumption of sequential conditional exchangeability.

Table 4.2.2: Summary data from Table 4.2.1, with both *unstabilised* and *stabilised* inverse probability of treatment weights (w) calculated for each combination of A_0, L_1, and A_1 and the resulting number of individuals in the 'pseudo population' (*Pseudo* N) calculated as $N * w$ and $N * sw$, respectively.

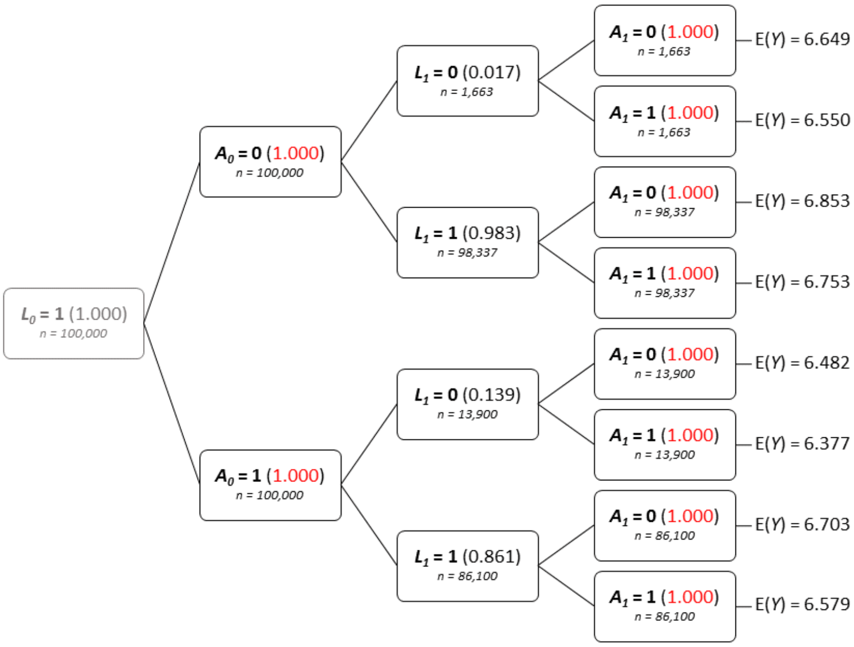
					IPTW (unstabilised weights)		IPTW (stabilised weights)	
A_0	L_1	A_1	$E(Y)$	N	w	Pseudo N	sw	Pseudo N
0	0	0	6.649	792	2.099	1,663	0.760	602
0	0	1	6.550	204	8.150	1,663	1.931	394
0	1	0	6.853	35,422	2.776	98,337	1.005	35,612
0	1	1	6.753	23,489	4.187	98,337	0.992	23,299
1	0	0	6.482	2,301	6.041	13,900	0.559	1,287
1	0	1	6.377	3,272	4.248	13,900	1.310	4,286
1	1	0	6.703	6,958	12.374	86,100	1.146	7,972
1	1	1	6.579	27,562	3.124	86,100	0.963	26,548

The weights (w) calculated in Table 4.2.2 are generally referred to as the **'unstabilised' weights**. However, it is generally preferable to use **'stabilised' weights** (sw), which are given by the formula:

$$sw = \prod_{t=0}^{T} \frac{P(A_t|\bar{A}_{t-1})}{P(A_t|\bar{A}_{t-1}, \bar{L}_t)}$$

As previously, the denominator of the expression above represents *the probability of receiving the treatment actually received at time* t, given past exposure and confounder history. However, this value is 'stabilised' by the numerator, which represents the marginal probability of the exposure received across the entire population. It can be shown that the mean counterfactual outcome $E(Y^{a_0, a_1})$ does not actually depend on the numerator, and thus both stabilised and unstabilised weights can be used to obtain an unbiased estimate of the total causal effect.[5] However, using stabilised weights has several benefits, including: a pseudo population that is equal in size to the original population, less-volatile weights (thereby minimising extreme up-weighting), a pseudo population in which the probability of each exposure regime is equal to those of the original population, and narrower confidence intervals for the estimates of causal effects produced in the MSM[5].

Figure 3.3.6: Modified decision tree from Figure 4.2.2, in which each individual both receives treatment and does not receive treatment at both baseline and follow-up (i.e. $P(A_0 = A_1 = 0) = 1.000$ and $P(A_0 = A_1 = 1) = 1.000$).



L0 = 1 (1.000) n = 100,000
A0 = 0 (1.000) n = 100,000
A0 = 1 (1.000) n = 100,000
L1 = 0 (0.017) n = 1,663
L1 = 1 (0.983) n = 98,337
L1 = 0 (0.139) n = 13,900
L1 = 1 (0.861) n = 86,100
A1 = 0 (1.000) n = 1,663
A1 = 1 (1.000) n = 1,663
A1 = 0 (1.000) n = 98,337
A1 = 1 (1.000) n = 98,337
A1 = 0 (1.000) n = 13,900
A1 = 1 (1.000) n = 13,900
A1 = 0 (1.000) n = 86,100
A1 = 1 (1.000) n = 86,100
E(Y) = 6.649
E(Y) = 6.550
E(Y) = 6.853
E(Y) = 6.753
E(Y) = 6.482
E(Y) = 6.377
E(Y) = 6.703
E(Y) = 6.579

For our scenario, the stabilised weight for each individual can be calculated as:

$$sw = \frac{P(A_0)}{P(A_0)} \cdot \frac{P(A_1|A_0)}{P(A_1|A_0, L_1)}$$

For example, to calculate the stabilised weight for an individual who did not receive metformin treatment at baseline ($A_0 = 0$), was not classified as 'high risk' at six months follow-up ($L_1 = 0$), and did not receive metformin treatment at six months follow-up ($A_1 = 0$) we compute:

$$sw = \frac{P(A_0 = 0)}{P(A_0 = 0)} \cdot \frac{P(A_1 = 0|A_0 = 0)}{P(A_1 = 0|A_0 = 0, L_1 = 0)} = \frac{0.599}{0.599} \cdot \frac{0.605}{0.795} = 0.760$$

The stabilised weights (sw) for all combinations of A_0, L_1, and A_1 are also given in Table 4.2.3, as are the total number of individuals in the pseudo population (*Pseudo* N) for each combination. Because the numerator of the stabilised weight represents the marginal probability of the exposure within the entire population and the denominator represents the marginal probability of the exposure within the stratum of past exposure and confounder history, individuals with exposure regimes that are more prevalent in the population than in the stratum will be up-weighted ($sw > 1$) and those with exposure regimes that are more prevalent in the stratum than in the population will be down-weighted ($sw < 1$).

The total number of individuals in the pseudo population created using stabilised weights is equal to 100,000, as in the original, unweighted population (Table 4.2.2). Moreover, the distribution of the exposure in this pseudo population is equal to that in the original population (Table 4.2.3).

Table 4.2.3: Distributions of the four exposure regimes for A_0 and A_1 in: the observed population (Table 4.2.1); the pseudo population constructed using unstabilised weights (Table 4.2.2); and the pseudo population constructed using stabilised weights (Table 4.2.2).

Exposure regime	Probability of exposure regime in…		
	Observed population	Pseudo population (unstabilised)	Pseudo population (stabilised)
$A_0 = 0, A_1 = 0$	0.362	0.250	0.362
$A_0 = 0, A_1 = 1$	0.237	0.250	0.237
$A_0 = 1, A_1 = 0$	0.093	0.250	0.093
$A_0 = 1, A_1 = 1$	0.308	0.250	0.308

We can confirm that treatment with metformin at six months follow-up (A_1) is conditionally independent of 'high risk' status at six months follow-up (L_1) in each of the pseudo populations. First, the *unstabilised* pseudo population: Among those who did not receive treatment at baseline ($A_0 = 0$), the conditional probability of receiving treatment at six months follow-up ($A_1 = 1$) is the same regardless of whether or not that individual was deemed to be 'high risk', i.e.:

$$P(A_1 = 1|A_0 = 0, L_1 = 0) = \frac{1{,}663}{1{,}663 + 1{,}663} = 0.5$$
$$P(A_1 = 1|A_0 = 0, L_1 = 1) = \frac{98{,}337}{98{,}337 + 98{,}337} = 0.5$$

Similarly, among those who did receive treatment t baseline ($A_0 = 1$), the conditional probability of receiving treatment at six months follow-up ($A_1 = 1$) is equal regardless of whether or not that individual was deemed to be 'high risk', i.e.:

$$P(A_1 = 1|A_0 = 1, L_1 = 0) = \frac{13{,}900}{13{,}900 + 13{,}900} = 0.5$$
$$P(A_1 = 1|A_0 = 1, L_1 = 1) = \frac{86{,}100}{86{,}100 + 86{,}100} = 0.5$$

These equivalencies also hold true in the *stabilised* pseudo population. Among those who did not receive treatment at baseline ($A_0 = 0$), the conditional probability of receiving treatment at six months follow-up ($A_1 = 1$) is the same regardless of whether or not that individual was deemed to be 'high risk', i.e.:

$$P(A_1 = 1|A_0 = 0, L_1 = 0) = \frac{394}{602 + 394} = 0.396$$
$$P(A_1 = 1|A_0 = 0, L_1 = 1) = \frac{23{,}299}{35{,}612 + 23{,}299} = 0.396$$

Among those who did receive treatment t baseline ($A_0 = 1$), the conditional probability of receiving treatment at six months follow-up ($A_1 = 1$) is equal regardless of whether or not that individual was deemed to be 'high risk', i.e.:

$$P(A_1 = 1|A_0 = 1, L_1 = 0) = \frac{4{,}286}{1{,}287 + 4{,}286} = 0.769$$
$$P(A_1 = 1|A_0 = 1, L_1 = 1) = \frac{26{,}548}{7{,}972 + 26{,}548} = 0.769$$

Thus, weighting by either w or sw creates a pseudo population in which treatment with metformin at six months follow-up is not dependent upon 'high risk' status, and so no time-dependent confounding is present. We may then fit a marginal structural model (MSM) using the data from each of these pseudo populations to target the effect of interest. In contrast to the g-formula, IPTW effectively simulates the data that would have been observed had, contrary to fact, exposure been unconditionally randomised at each time point.[5]

Note: Stabilised and unstabilised weights are, for all intents and purposes, interchangeable for estimating the causal effects of *non-dynamic* (i.e. *static*) treatment regimes, in which the treatments considered in the hypothetical interventions are independent of the time-varying covariates.[7] However, for *dynamic* treatment regimes, in which the treatment strategy at each time point depends either deterministically or probabilistically on the evolution of the individual's measured time-dependent confounders and, possibly, treatment history, only the unstabilised pseudo population should be used. These issues (and others) are far beyond the scope of this lecture and are covered more thoroughly by Robins and Hernan.[5]

Marginal structural modelling: Estimating the causal effect

We have previously defined the average potential outcome for exposure regime $a = (a_0, a_1)$ to be $E(Y^{a_0, a_1})$. For our scenario, this average potential outcome may be expressed as a sum of the individual effects of each of a_0 and a_1, their two-way interaction, plus some constant, i.e.:

$$E(Y^{a_0, a_1}) = \alpha + \beta_0 a_0 + \beta_1 a_1 + \beta_2 a_0 a_1$$

This model is referred to as a **marginal structural model** because it models the marginal mean of the counterfactual outcome Y^{a_0, a_1}, and models for counterfactuals are often referred to as 'structural models'[5]. Above, β_0 represents the average effect of a_0, β_1 represents the average effect of a_1, and β_2 represents the average additional joint effect of a_0 and a_1 on the outcome. Thus, the average potential outcome for each of the four exposure regimes $a = (0,0)$, $a = (0,1)$, $a = (1,0)$, and $a = (1,1)$ may be written as:

$$\begin{aligned}
E(Y^{0,0}) &= \alpha + \beta_0(0) + \beta_1(0) + \beta_2(0)(0) = \alpha \\
E(Y^{0,1}) &= \alpha + \beta_0(0) + \beta_1(1) + \beta_2(0)(1) = \alpha + \beta_1 \\
E(Y^{1,0}) &= \alpha + \beta_0(1) + \beta_1(0) + \beta_2(1)(0) = \alpha + \beta_0 \\
E(Y^{1,1}) &= \alpha + \beta_0(1) + \beta_1(1) + \beta_2(1)(1) = \alpha + \beta_0 + \beta_1 + \beta_2
\end{aligned}$$

Because there exists no time-dependent confounding in either of the pseudo-populations, each of the four groups of exposure regimes are **unconditionally exchangeable**. We may equate their potential outcomes Y^{a_0,a_1} to their observed outcomes $Y|A_0 = a_0, A_1 = a_1$ by the consistency assumption to estimate the desired average potential outcome $E(Y^{a_0,a_1})$ for each of the four exposure regimes. For example, the average potential outcome that would be observed were everyone prohibited from receiving treatment at both baseline ($A_0 = 1$) and six months follow-up ($A_1 = 0$) can be estimated (using data from the *stabilised* pseudo population in Table 4.2.2) as:

$$\begin{aligned}
\hat{E}(Y^{0,0}) &= \hat{E}(Y|A_0 = 0, A_1 = 0) \\
&= \hat{E}(Y|A_0 = 0, L_1 = 0, A_1 = 0)P(L_1 = 0|A_0 = 0, A_1 = 0) \\
&\qquad + \hat{E}(Y|A_0 = 0, L_1 = 1, A_1 = 0)P(L_1 = 1|A_0 = 0, A_1 = 0) \\
&= (6.649)\left(\frac{602}{602 + 35{,}612}\right) + (6.853)\left(\frac{35{,}612}{602 + 35{,}612}\right) \\
&= 6.850
\end{aligned}$$

Estimates of the average potential outcome for each of the four exposure regimes are given below, utilising data from the *stabilised* pseudo population (Table 4.2.2):

$$\begin{aligned}
\hat{E}(Y^{0,0}) &= \hat{E}(Y|A_0 = 0, A_1 = 0) = 6.850 = \alpha \\
\hat{E}(Y^{0,1}) &= \hat{E}(Y|A_0 = 0, A_1 = 1) = 6.750 = \alpha + \beta_1 \\
\hat{E}(Y^{1,0}) &= \hat{E}(Y|A_0 = 1, A_1 = 0) = 6.672 = \alpha + \beta_0 \\
\hat{E}(Y^{1,1}) &= \hat{E}(Y|A_0 = 1, A_1 = 1) = 6.551 = \alpha + \beta_0 + \beta_1 + \beta_2
\end{aligned}$$

Thus, we compute our estimate of the TCE of metformin treatment among 'high-risk' prediabetic individuals as:

$$\hat{E}(Y^{1,1}) - \hat{E}(Y^{0,0}) = 6.551 - 6.850 = \mathbf{-0.299}$$

We can alternately use ordinary least-squares regression on our IPTW pseudo population to target the causal effect of interest. To illustrate, consider that:

$$TCE = E(Y^{1,1}) - E(Y^{0,0}) = \alpha - (\alpha + \beta_0 + \beta_1 + \beta_2) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} + \boldsymbol{\beta_2}$$

The above parameters can be easily estimated using standard statistical software. For our scenario, a model summary (again utilising the *stabilised* pseudo population) is provided below:

$$\hat{E}(Y^{a_0,a_1}) = 6.850 - 0.178a_0 - 0.100a_1 - 0.021a_0a_1$$

Thus, our estimate of the TCE is:

$$\widehat{TCE} = (-0.178) + (-0.100) + (-0.021) = \mathbf{-0.299}$$

This regression method is preferable to computing the individual conditional expectations when the exposure regimes under consideration are more complex than those in our simple illustrative example. The same estimate could also be obtained by using the *unstabilised* pseudo population (as the conditional expectations of Y are equivalent for both pseudo populations), though for brevity we do not go through these calculations.

G-estimation of structural nested models (SNMs)

G-estimation is a process which is based on solving equations that result directly from the assumption of conditional exchangeability.[2] Of the three g-methods, g-estimation of structural nested models is the most flexible method but also the most conceptually and practically challenging to implement. In this section, we illustrate how g-estimation may be used to estimate the total causal effect for our metformin treatment example scenario.

Before we consider structural *nested* mean models for a time-varying exposure $A = (A_0, A_1)$ – more commonly referred to as simply **structural nested models (SNMs)** – we first introduce *structural mean models (SMMs)* for a time-invariant exposure *A*. An SMM is a model for the conditional causal effect $E(Y^a - Y^0|L = l)$; that is, it encodes the average effect of setting an exposure A to a particular value a versus setting that exposure to zero, amongst individuals within a particular stratum of confounder L.[8]

For example, we may express the average potential outcome within strata of L $E(Y^a|L = l)$ as the (linear) effect of A on Y plus some constant, i.e.:

$$E(Y^a|L = l) = \alpha + \gamma a$$

Note, however, that $E(Y^0|L = l) = \alpha$; thus, we may rearrange the equation above to create the SMM:

$$E(Y^a - Y^0|L = l) = \gamma a$$

This SMM implies that the causal effect of A on Y is both linear and equivalent across all levels of the confounder L. We could also consider more complex parameterisations (and, indeed, this is one of the strengths of SMMs), but these are not required for the purposes of illustration. In the equation above, γ represents the average causal effect of setting the exposure A to a particular value a versus setting that exposure to zero, amongst individuals within strata of L.

For a time-varying exposure $= (A_0, A_1)$, in which individuals may be either exposed or unexposed at each time point, we extend this concept. A structural nested (mean) model describes, *at each time point* t, the average effect of setting an exposure A to a particular value a versus setting that exposure to zero, in the hypothetical scenario in which the exposure is set to zero after time t; this is done within levels of past exposure ($\bar{A}_{t-1}$) and confounder history ($\bar{L}_t$) through time t.[8]

To illustrate, consider the DAG in Figure 4.2.1 as a general example of an exposure A measured once at time 0 (A_0) and once at time 1 (A_1), a time-varying covariate measured once at time 1 (L_1), and an outcome measured once at time 2 (Y). Because individuals may receive treatment at two separate time points, our SNM is made up of the following two models:

$$E(Y^{a_0,0} - Y^{0,0}) = \beta_0 a_0$$
$$E(Y^{a_0,a_1} - Y^{a_0,0}|L_1 = l_1, A_0 = a_0) = \delta_0 a_1 + \delta_1 a_1 l_1 + \delta_2 a_1 a_0 + \delta_3 a_1 l_1 a_0$$

The first model encodes the average effect (β_0) of setting A_0 to a_0 at time 0 when treatment at time 1 is withheld ($A_1 = 0$); because there are no preceding exposures or confounders, this contrast is not conditional on anything. The second model encodes the average effect of setting A_1 to a_1 at time 1 when the value for A_0 remains constant, within levels of L_1 and A_0.

For our specific example scenario, in which A_0, L_1, and A_1 are all binary, the above SNM can be reduced to:

$$E(Y^{1,0} - Y^{0,0}) = \beta_0$$
$$E(Y^{a_0,1} - Y^{a_0,0}|L_1 = l_1, A_0 = a_0) = \delta_0 + \delta_1 l_1 + \delta_2 a_0 + \delta_3 l_1 a_0$$

The first model states that the average effect of receiving treatment at baseline ($A_0 = 1$) when treatment at six months follow-up is withheld ($A_1 = 0$) (i.e. the *direct effect* of A_0 on Y) is equal to β_0. The four parameters $\delta_0, \delta_1, \delta_2, \delta_3$ in the second model express the effect of receiving treatment at six months follow-up ($A_1 = 1$) within the four possible levels A_0 (treatment at baseline) and L_1 ('high risk' status at six-month follow-up). The average effect of receiving treatment at six months follow-up is: (i) δ_0 when $A_0 = 0$ and $L_1 = 0$; (ii) $\delta_0 + \delta_1$ when $A_0 = 0$ and $L_1 = 1$; (iii) $\delta_0 + \delta_2$ when $A_0 = 1$ and $L_1 = 0$; and (iv) $\delta_0 + \delta_1 + \delta_2 + \delta_3$ when $A_0 = 1$ and $L_1 = 1$. Thus, our SNM models the degree to which the effect of metformin treatment at six months follow-up is modified by metformin treatment at baseline and 'high risk' status at six months.[5]

The utility of this representation for estimating the TCE of A on Y becomes clear if we consider that the total effect of A on Y may be thought of as the sum of two effects: (i) the average effect of A_1 on Y, given that treatment at A_0 was also received; and (ii) the average (direct) effect of A_0 on Y. Thus, we may alternately express the TCE as:

$$TCE = E[(Y^{1,1} - Y^{1,0}|L_1 = l_1, A_0 = 1) + (Y^{1,0} - Y^{0,0})]$$

This expression of the TCE is equivalent to our original expression $TCE = E(Y^{1,1} - Y^{0,0})$, and comprises the two components of our previously-defined SNM.

Therefore, estimating the TCE requires that we solve for the unknown parameters $\beta_0, \delta_0, \delta_1, \delta_2, \delta_3$. To this end, we first use the second equation of our SNM to express the counterfactual quantity $E(Y^{A_0,0})$ in terms of the unknown parameters $\delta_0, \delta_1, \delta_2, \delta_3$ for each strata of A_0, L_1, and A_1 (i.e. $E(Y^{A_0,0}|L_1 = l_1, A_0 = a_0) = E(Y^{A_0,0}|A_1 = a_1, L_1 = l_1, A_0 = a_0)$, by conditional exchangeability); these are given in Table 4.2.4.

Table 4.2.4: Summary data from Table 4.2.1 with the counterfactual quantity $E(Y^{A_0,0}|A_1 = a_1, L_1 = l_1, A_0 = a_0)$ calculated within each possible combination of metformin treatment at baseline (A_0), 'high risk' status at six-month follow-up (L_1), and metformin treatment at six-month follow-up (A_1).

A_0	L_1	A_1	$E(Y)$	N	$E(Y^{A_0,0}\|A_1, L_1, A_0)$
0	0	0	6.649	792	6.649
0	0	1	6.550	204	6.550 - δ_0
0	1	0	6.853	35,422	6.853
0	1	1	6.753	23,489	6.753 - δ_0 - δ_1
1	0	0	6.482	2,301	6.482
1	0	1	6.377	3,272	6.377 - δ_0 - δ_2
1	1	0	6.703	6,958	6.703
1	1	1	6.579	27,562	6.579 - δ_0 - δ_1 - δ_2 - δ_3

To illustrate the derivations of the expressions for $E(Y^{A_0,0}|A_1,L_1,A_0)$ in Table 4.2.4, consider the case in which metformin treatment was not received at baseline ($A_0 = 0$), individuals were not classified as 'high risk' at six months follow-up ($L_1 = 0$), and metformin treatment was not received at six months follow-up ($A_1 = 0$). The average potential outcome $E(Y^{A_0,0}|A_1,L_1,A_0)$ is equal to $E(Y^{0,0}|A_1 = 0, L_1 = 0, A_0 = 0)$. By consistency, we can conclude that $E(Y^{0,0}|A_1 = 0, L_1 = 0, A_0 = 0) = E(Y|A_1 = 0, L_1 = 0, A_0 = 0)$, and therefore the mean counterfactual outcome $E(Y^{0,0}|A_1 = 0, L_1 = 0, A_0 = 0)$ is equal to the mean observed outcome 6.649.

As another example, consider the case in which metformin treatment was received at baseline ($A_0 = 1$), individuals were classified as 'high risk' at six months follow-up ($L_1 = 1$), and metformin treatment was subsequently received at six months follow-up ($A_1 = 1$). By consistency, the observed mean $E(Y|A_1 = 1, L_1 = 1, A_0 = 1) = 6.579$ is equal to the counterfactual mean $E(Y^{1,1}|A_1 = 1, L_1 = 1, A_0 = 1)$. To calculate $E(Y^{A_0,0}|A_1,L_1,A_0) = E(Y^{1,0}|A_1 = 1, L_1 = 1, A_0 = 1)$, we use the second equation of our SNM, with $A_0 = 1$, $L_1 = 1$, and $E(Y^{1,1}|A_1 = 1, L_1 = 1, A_0 = 1) = 6.579$:

$$6.579 = E(Y^{1,0}|A_1 = 1, L_1 = 1, A_0 = 1) + \delta_0 + \delta_1(1) + \delta_2(1) + \delta_3(1)(1)$$
$$\text{Thus, } E(Y^{1,0}|A_1 = 1, L_1 = 1, A_0 = 1) = 6.579 - \delta_0 - \delta_1 - \delta_2 - \delta_3.$$

Our derivations thus far have relied on the consistency condition for identifiability. We now exploit the conditional exchangeability condition to produce estimates for the unknown parameters $\delta_0, \delta_1, \delta_2, \delta_3$. As defined previously, (sequential) conditional exchangeability states that the potential outcome Y^{a_0,a_1} for a given treatment regime $a = (a_0, a_1)$ is independent of the treatment actually received at both baseline (A_0) and six months follow-up (A_1, within levels of A_0 and L_1). This implies that the average potential outcome $E(Y^{A_0,0}|A_1,L_1,A_0)$ should be the same regardless of whether $A_1 = 0$ or $A_1 = 1$ (i.e. $E(Y^{A_0,0}|A_1,L_1,A_0) = E(Y^{A_0,0}|L_1,A_0)$). This may be exploited to produce estimates for $\delta_0, \delta_1, \delta_2, \delta_3$.

Within the strata defined by $A_0 = 0$ and $L_1 = 0$, we may equate $6.649 = 6.550 - \delta_0$, which implies $\delta_0 = -0.099$. Within the strata defined by $A_0 = 0$ and $L_1 = 1$, we may equate $6.853 = 6.753 - \delta_0 - \delta_1 = 6.753 - (-0.099) - \delta_1$, which implies $\delta_1 = -0.001$. Within the strata defined by $A_0 = 1$ and $L_1 = 0$, we may equate $6.482 = 6.377 - \delta_0 - \delta_2 = 6.377 - (-0.099) - \delta_2$, which implies $\delta_2 = -0.006$. And finally, within the strata defined by $A_0 = 1$ and $L_1 = 1$, we may equate $6.703 = 6.579 - \delta_0 - \delta_1 - \delta_2 - \delta_3 = 6.579 - (-0.099) - (-0.001) - (-0.006) - \delta_3$, which implies $\delta_3 = -0.018$.

Thus, our SMM can now be written as:

$$E(Y^{1,0} - Y^{0,0}) = \beta_0$$
$$E(Y^{a_0,1} - Y^{a_0,0}|L_1 = l_1, A_0 = a_0) = -0.099 - 0.001l_1 - 0.006a_0 - 0.018l_1a_0$$

To estimate β_0, we can again exploit conditional exchangeability, which further implies that the average potential outcome $E(Y^{0,0})$ (from the first equation above) is independent of the observed value of exposure A_0, i.e. $E(Y^{0,0}) = E(Y^{0,0}|A_0 = 0) = E(Y^{0,0}|A_0 = 1)$. We can estimate this quantity from the data in Table 4.2.4 as:

$$E(Y^{0,0}) = E(Y^{0,0}|A_0 = 0) = (6.649)\left(\frac{996}{59{,}907}\right) + (6.853)\left(\frac{58{,}911}{59{,}907}\right) = 6.850$$

We can estimate $E(Y^{1,0})$ using the same process, because $E(Y^{1,0}) = E(Y^{1,0}|A_0 = 0) = E(Y^{1,0}|A_0 = 1)$ by conditional exchangeability:

$$E(Y^{1,0}) = E(Y^{1,0}|A_0 = 1) = (6.482)\left(\frac{5{,}573}{40{,}093}\right) + (6.703)\left(\frac{34{,}520}{40{,}093}\right) = 6.672$$

Therefore, we can compute $\beta_0 = E(Y^{1,0}) - E(Y^{0,0}) = 6.672 - 6.850 = -0.178$. Our final SNM is given below:

$$\hat{E}(Y^{1,0} - Y^{0,0}) = -0.178$$
$$\hat{E}(Y^{a_0,1} - Y^{a_0,0}|L_1 = l_1, A_0 = a_0) = -0.099 - 0.001l_1 - 0.006a_0 - 0.018l_1a_0$$

The first equation above represents the average direct effect of metformin treatment at baseline (A_0) on fasting blood glucose level at 1-year follow-up (Y). The second equation above represents the average effect of metformin treatment at six months follow-up (A_1) on fasting blood glucose level at 1-year follow-up (Y), within strata defined by metformin treatment at baseline (A_0) and 'high risk' status at six months follow-up (L_1).

We now use our SNM to estimate the TCE of metformin treatment on fasting blood glucose level:

$$\widehat{TCE} = \hat{E}(Y^{1,1} - Y^{1,0}|L_1 = l_1, A_0 = 1) + \hat{E}(Y^{1,0} - Y^{0,0})$$

The first quantity of the TCE (i.e. $\hat{E}(Y^{1,1} - Y^{1,0}|L_1 = l_1, A_0 = 1)$) represents the estimated effect of metformin treatment at six months follow-up (A_1) on fasting blood glucose level at 1-year follow-up (Y), only within the strata in which $A_0 = 1$ (i.e. in which treatment was received at baseline). To compute this quantity for the population, we compute the weighted average across the two strata defined by $L_1 = 0, A_0 = 1$ and $L_1 = 1, A_0 = 1$:

$$\begin{aligned}
&\hat{E}(Y^{1,1} - Y^{1,0}|L_1 = l_1, A_0 = 1) \\
&\quad = \hat{E}(Y^{1,1} - Y^{1,0}|L_1 = 0, A_0 = 1)P(L_1 = 0|A_0 = 1) + \hat{E}(Y^{1,1} - Y^{1,0}|L_1 = 1, A_0 = 1)P(L_1 = 1|A_0 = 1) \\
&\quad = (-0.099 - 0.006)(0.139) + (-0.099 - 0.001 - 0.006 - 0.018)(0.861) \\
&\quad = -0.121
\end{aligned}$$

The second quantity of the TCE (i.e. $\hat{E}(Y^{1,0} - Y^{0,0})$) equals -0.178 from the first equation of our SNM. Thus, we estimate the TCE as:

$$\widehat{TCE} = -0.121 - 0.178 = \mathbf{-0.299}$$

Summary

For the example scenario considered throughout this section of these lecture notes, all three g-methods gave identical estimates of the total causal effect (TCE). This equivalency is expected to hold for nonparametric analyses (i.e. when estimates are based only on conditional probabilities and expectations). However, for scenarios in which parametric modelling is required (e.g. if L_1 were continuous such that calculating the weights for IPTW would require fitting a logistic regression model to estimate the probability of treatment), they will likely give different estimates of the TCE. The example scenario we considered was purposely oversimplified in order to demonstrate the fundamental principles underlying each of the three g-methods. Real-world scenarios are likely to consist of a greater number of longitudinal time points, baseline covariates, and time-varying confounders; moreover, few real-world scenarios deal only with binary variables. The methods outlined previously may be extended to deal with such complexity, but this is beyond the scope of this lecture.

4.3 Deterministic Variables and Tautological Associations

Learning objectives

- Describe the concept of deterministic variables within a DAG
- Introduce the concept of mathematical coupling (MC) and resultant tautological associations
- Explain tautological associations within the analysis of change with respect to initial value
- Explain tautological associations in the analysis of constructed ratio variables
- Discuss some ways to avoid tautological associations

DAGs for deterministic relationships

Distinguishing causal effect from statistical association is imperative to health and social science research, where observational data scientists have to identify, measure, and compensate for every source of non-causal association between the two variables of interest.[88] Directed acyclic graphs (DAGs) have become increasingly popular because they provide an intuitive aid to help identify and understand different types of non-causal associations.[3,13] However, little attention has yet been given to describing or considering tautological associations – self-fulfilling associations that arise when a variable is analysed in relation to: itself, a part thereof, or a 'sibling' variable with shared algebraic components.[89] Tautological associations have received little attention in DAGs because they arise from deterministic relationships rather than the probabilistic relationships for which DAGs are more commonly utilised.[22] Indeed, many of the familiar 'rules' for identifying causal effects break down in the presence of deterministic relationships, in which a variable is fully explained (i.e. determined) by one or more parent variable(s).

Deterministic variables can be defined as variables that are fully determined by one or more parent variables via a functional non-probabilistic relationship. Depending on how they occur, variables arising from deterministic relationships can be summarised as the following:

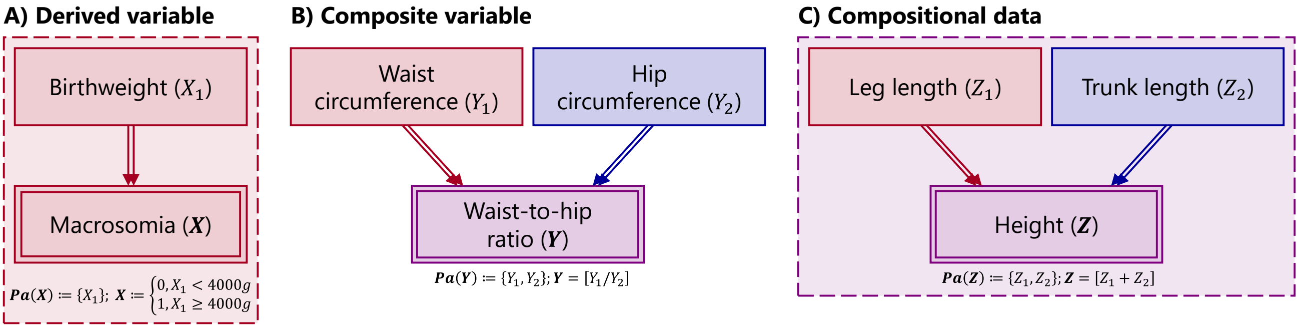
- **Derived variables**: Variables functionally created from, and fully explained by, a single parent variable, e.g. macrosomia is a variable derived from and fully explained by birthweight.
- **Composite variables**: Variables functionally created from, and fully explained by, two or more parent variables, e.g. waist-to-hip ratio is a composite variable derived from and fully explained by waist and hip circumference.
- **Compositional data**: Hierarchical data containing distinct 'part' variables that sum to a 'whole' variable, e.g. trunk and leg length sum up to total height.

To illustrate deterministic relationships in DAGs, we use double-outlined nodes to depict fully determined variables[90] and signify the deterministic relationship using double-lined arcs. We additionally enclose all variables that share a deterministic relationship within a dashed-outline box, to highlight that they cannot be separated in time. Examples of this notation are given in Figure 3.4.1, where Figure 3.4.1a depicts a derived variable, Figure 3.4.1b depicts a composite variable, and Figure 3.4.1c depicts compositional data.

Identifying Tautological Associations

Perhaps the most straightforward benefit to depicting deterministic relationships within DAGs is the ability to identify and avoid misinterpreting tautological associations. A tautological association is a 'self-fulfilling' association that arises when a variable is analysed in relation to itself, a part thereof, or a 'sibling' variable with shared algebraic components.[91] Inferential bias occurs when tautological associations are mistakenly interpreted as causally meaningful, rather than resulting from their functionally deterministic relationship. Although such mistakes are probably more common with higher order composite variables, tautological associations between derived variables and their parents can still be overlooked at aggregate level, as in a recent study that compared the average systolic blood pressure and prevalence of hypertension between areas.[92]

Figure 3.4.1. Directed acyclic graphs using deterministic notation to depict (a) a derived variable (b) a composite variable and (c) compositional data.


A) Derived variable
Birthweight (X_1)
Macrosomia ($\boldsymbol{X}$)
$\boldsymbol{Pa}(\boldsymbol{X}) \coloneqq \{X_1\};\ \boldsymbol{X} \coloneqq \begin{cases} 0, X_1 < 4000g \\ 1, X_1 \geq 4000g \end{cases}$
B) Composite variable
Waist circumference (Y_1)
Hip circumference (Y_2)
Waist-to-hip ratio ($\boldsymbol{Y}$)
$\boldsymbol{Pa}(\boldsymbol{Y}) \coloneqq \{Y_1, Y_2\}; \boldsymbol{Y} = [Y_1/Y_2]$
C) Compositional data
Leg length (Z_1)
Trunk length (Z_2)
Height ($\boldsymbol{Z}$)
$\boldsymbol{Pa}(\boldsymbol{Z}) \coloneqq \{Z_1, Z_2\}; \boldsymbol{Z} = [Z_1 + Z_2]$

Mathematical Coupling

In its simplest form, **mathematical coupling** (MC) is the phenomenon where the *null hypothesis* is distorted due to an **algebraic relationship** between two or more variables that are analysed by correlation or regression. Due to this distortion, any test of the null hypothesis (i.e. that the regression coefficient is zero) will be biased,[93] as will any corresponding inferences.[94–96] While regular null hypothesis significance testing becomes invalid since coupled variables are no longer independent, more importantly, any inferences drawn from the statistical evaluation are likely to be seriously misleading.

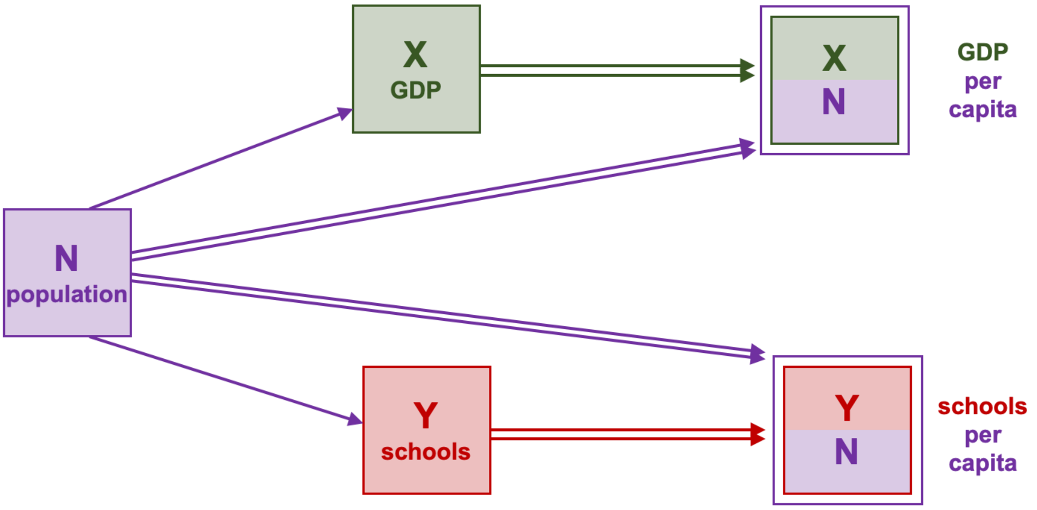
MC most noticeably occurs when a new variable is constructed from a mathematical transformation of another, e.g. through addition, subtraction, multiplication or division.[97] Examples include change variables (e.g. change between baseline and follow-up) and ratio variables, either where one variable is divided by another (e.g. prevalence proportions) or divided by a function of another (e.g. body mass index [BMI], where weight in kilograms is divided by height in meters squared). MC then arises if these constructed variables are analysed with respect to any of their component variables using correlation or regression (e.g. comparing two prevalence rates, which share the same denominator, or predicting BMI from height).

The effects of MC are well known and have been stated for some time, beginning with Pearson's warning for would-be-analysts in 1897 to be wary of what he termed a 'spurious correlation' that can arise between two ratio variables (e.g. X/N and Y/N) with a common denominator parent (N).[98] Pearson proposed a general solution[99] that was later reinforced by Fisher[100] in 1947 and Neyman[101] in 1979. A warning around the analysis of change with respect to baseline was made most prominently by Oldham[102] in 1962, where he proposed a basic solution for the correlation of two variables that later became known as the Bland and Altman test[103] of 1986 fame, though the basis of this solution has been alluded to as early as 1939 by Morgan.[104] Although these problems came to light many years ago, the term *mathematical coupling* did not appear in much of the literature until the term was coined by Archie in 1981;[105] and despite this history, the phenomena of MC and its proclivity to generate misleading tautological associations has either been overlooked or a source of confusion for many (including statisticians),[97] and still leads to inaccurate claims.[106]

Tautological associations such as mathematical coupling can be easily identified from a DAG that contains the ratio variables (e.g. X/N and Y/N) as well as their three parent variables (X, Y, and N).

Figure 3.4.2 depicts a situation in which it is clear that N (i.e. population) is a common ancestor of the other parent and ratio variables, and also a confounder for the relationship between X/N and Y/N (i.e. GDP per capita and schools per capita).

Figure 3.4.2. Directed acyclic graph showing the tautological association between two composite ratio variables that share a common denominator parent variable.



Analysis of change with respect to initial value

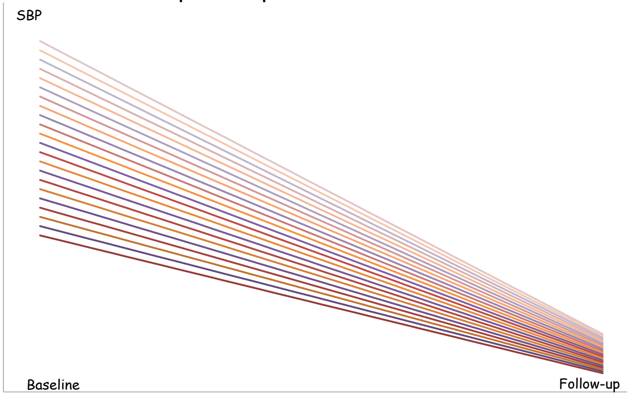
The most widely recognised illustration of MC and the generation of tautological associations arises in the analysis of change with respect to initial value. The relation between initial disease status and change following an intervention has attracted considerable interest in clinical research. What seems a relatively simple issue is deceptively complex, and the obvious strategies for analysing such data – i.e. simple correlation between change and baseline, or the regression of change on baseline – are highly problematic.[97] At first glance, it is far from clear what the problem is, which is perhaps why so many continue to make the same analytical mistakes.

Many (statisticians included) who at least recognise there to be problems with analysing change with respect to baseline often mistakenly attribute the problem entirely to *regression to the mean* (RTM); the issue of mathematical or causal coupling is then overlooked completely. It is not only RTM that is present in the analysis of change with respect to initial value, and the tautological associations generated by MC cannot be solved by seeking to minimise RTM.[97]

For instance, we ask: *Do individuals with higher initial systolic blood pressure (SBP) experience greater SBP reduction following intervention?* We are therefore asking if there is an **intervention differential effect**, where changes in SBP (due to pharmacological treatment of hypertension) depend on patients' initial SBP level (see figure 3.4.3).

Despite the many warnings against correlating or regressing change on initial value, researchers (including statisticians) have published confusing analyses that fall foul of these warnings.

Figure 3.4.3: Response to hypertension treatment: an *intervention differential effect* means that change in SBP following treatment depends upon an individual's initial SBP



Analysis of Ratio Variables

A *ratio variable* is where a new variable is derived from the division of one variable by another. In medicine and health, many variables are generated as ratios to capture a measure of one human feature (e.g. obesity), while acknowledging that people vary in genetic predisposition for others (e.g. height). Hence, ratios seek to capture a *relative* construct (e.g. BMI as a measure of weight relative to height-squared). In epidemiology, one is often concerned with prevalence and incidence (counts of total cases per population and counts of new cases per population per unit time, respectively), which are also ratios that capture the *relative* extent of a condition (e.g. prevalence of obesity, incidence of mortality) by accounting for differences in population sizes. The concept of what is *relative* in both contexts is seeking to standardise a measure with respect to a perceived 'norm', such as average body height or a typical cross-section of society.

The implications of MC amongst ratio variables are numerous and far reaching. Some of the most important outcomes and exposures in epidemiology are ratios, which is why the impacts of potential tautological associations generated from these variables are so crucial. The most ubiquitous examples are perhaps found in health geography, where the common denominator of *population at risk* is investigated for the population rates of two or more properties (e.g. proportion of unemployment and limiting long-term illness), or if investigating the relationship between the incidence of a disease and the prevalence of an exposure for that disease (e.g. proportion of children with asthma and the proportion of households that are overcrowded). A sizeable association can be observed, even if the relationship is entirely artefact.

To illustrate, consider three random variables (x, y and z) that are uncorrelated with each other and have identical standard deviations. It can be shown that the correlation of x/z with $y/z \approx$ 0.5.[98] A strong correlation will exist between any two variables when divided by the same denominator, even if they otherwise have nothing in common. As with the analysis of change, the assumption that the null is zero for the correlation or regression of ratio variables that share a common denominator is entirely false. Any estimated relationship will comprise an element of true effect (if non-zero) plus artefact; the latter will often be sizeable and could dominate.

To tackle the problem, Pearson suggested that analysts should calculate the *partial correlation* between numerators (disease counts) whilst 'adjusting' (within a regression model) for the common denominator (population counts), rather than analysing the two ratios directly.[99] Poisson regression automatically advocates this approach by encouraging analysts to model counts with a denominator 'offset' included as a model covariate (typically logged to match the Poisson log-link). Consequently, Poisson multivariable regression avoids the adverse impacts of MC, though this is fortuitous due to how Poisson regression is applied in epidemiology, rather than by design to deliberately avoid tautological associations – it does not however 'adjust' for cluster population size, as perhaps intended, but that is a separate issue. Fisher[100] and Neyman[101] both later reiterated how numerators and denominators of a ratio variable should be separated and analysed as Pearson suggested (not as a logged offset).

It should be noted that MC arises amongst ratios analysed by correlation or regression whenever each construct possesses common elements as numerator *or* denominator in either, i.e. x/z is also coupled with z/y and z/x is also coupled with z/y. Coupling will similarly occur if either numerator or denominator is a function of common elements, as for instance w/h^2 is coupled to any expression of w or h (where $w = \mathbf{weight}$ and $h = \mathbf{height}$, $w/h^2 = \mathbf{body\ mass\ index}$). We must therefore be vigilant in recognising the potential for tautological associations when examining such constructs using correlation or regression, since the association between such coupled ratios will yield, at least in part, spurious findings. Although we have a proposed solution for the situation of **common denominators**, other instances of MC between ratios will need different approaches and a solution may not always be apparent.

A causal framework of ratio variables

Within a causal inference framework, the analysis of change and investigation of variables used to generate ratios can be extremely challenging – even if there is no obvious mathematical coupling – as we will later see. Even for MC due to ratios with common denominators, the proposed solution by Pearson,[99] Fisher,[100] and Neyman[101] may not provide robust causal inferences. This is because there will always be exogenous factors surrounding the inter-variable causal relationships that must be taken into consideration when examining variable components within a casual question, and these restrictions were not fully grasped at the time Pearson, Fisher, and Neyman were writing.

Summary

MC is ubiquitous and yet despite adversely impacting research in many different guises, its consequences (or indeed its very existence) are hardly recognised. Even amongst those who have encountered (some of) the issues with the spurious generation of tautological associations, comprehension is poor, and there is little appreciation of the solutions needed – some already proposed, some in need of development – to overcome the resulting analytical challenges.

Analysis of change is deceptively complex, and although widely employed in research, ratios are extremely problematic and present many challenges. When considering causal questions, it is best to separate components (i.e. follow-up and baseline, or numerator and denominator) to analyse variables separately within a multivariable linear regression model. This nevertheless relies upon the model being meaningfully placed within a causal framework and informed by an appropriate directed acyclic graph.[23]

Further work is needed in both highlighting these problems and, ultimately, providing workable solutions. In the meantime, a limited appreciation of coupling means we continue to encounter erroneous analyses that yield meaningless and/or misleading findings. This is particularly worrying since some results, no matter how biased, will likely be perceived by some as robust.

DAY 5

5.1 Compositional Data

Learning objectives

- Explain what is meant by compositional data
- Understand when collider bias may not actually be a 'bias' with compositional data
- Discuss the challenges of interpreting compositional data in a causal inference framework
- Identify the appropriate modelling strategies to avoid composite variable bias

Dependency amongst covariates in multivariable models

It is important to remember that when we regress Y on X, adjusting for Z, we effectively ask: W*hat is the relationship of X with Y whilst keeping Z constant?* The assumption made is that the X-Y relationship is the same for all values of Z, i.e. the relationship is conditionally 'independent' of Z. We therefore need to think carefully about the implications of holding Z constant. For instance, what if $Z=X^2$, i.e. we have the quadratic model: $Y=\beta_0+\beta_1X+\beta_2X^2$?

Clearly, we cannot interpret the coefficient for X (β_1) as though X^2 were constant; this instead requires the *joint* interpretation of the coefficients for X and X^2, i.e. β_1 and β_2 must be considered simultaneously when seeking to understand the X-Y relationship. This is not too challenging an issue if we are familiar with interpreting *curvilinear* relationships; but there are more complex scenarios that often go unnoticed where the same issue arises.

Generally, the dependency induced between independent events when conditioning on a common outcome has the potential to cause serious interpretational problems for causal analyses. Termed 'collider bias', it often produces seemingly paradoxical results which are contrary to intuition (e.g. the Monty Hall problem[3]) and scientific feasibility (e.g. the birthweight paradox[62]). Although this problem has been discussed in various epidemiological contexts,[62,64,69,107] there has been no explicit discussion of collider bias in the context of compositional data, for which additional nuances exist.

Compositional data comprise the parts of some whole, for which all parts sum to that whole;[108] the whole itself may vary across units of analysis (e.g. total energy intake) or remain fixed (e.g. total hours in a day). Almost all data are potentially compositional, in the sense that most concepts can be considered part of a greater whole and/or subdivided into smaller parts, though data are often explicitly conceptualised as compositional when there is interest in understanding the role of one or more component(s) in relation to the whole. Many of the challenges associated with analysing compositional data have been widely discussed,[108–110] though none have sought to demonstrate these challenges within a formal causal framework utilising DAGs. This is despite the utilisation of DAGs becoming increasingly widespread and providing insights into historical 'paradoxes'[111,112], including collider bias in its generic form. Using DAGs to consider the issue of collider bias in the context of compositional data, we describe the nuances therein and provide a systematic approach to thinking about the analytical and interpretational issues that may arise. Moreover, we discuss circumstances in which conditioning on a collider provides meaningful and useful causal effect estimates, as well as circumstances in which it is unavoidable. (These notes are reproduced here from two published papers.[113,114])

Collider 'bias' for compositional data

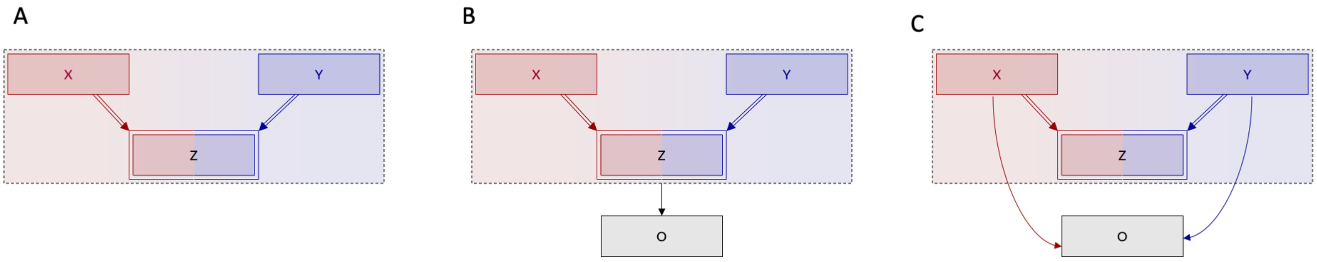
Consider the three random variables – X, Y, and Z – for which X + Y = Z. The relationship amongst these variables is depicted in the DAG in Figure 4.2.1a, where the double circles around Z indicate that it is a deterministic function of X and Y. X and Y are (unconditionally) independent, but become dependent when conditioning on Z. The reason for this is simple – conditioning on Z may be thought of as 'filtering' by Z or holding Z constant; therefore, any change in one of the components (X or Y) must be accompanied by an equal and opposite change in the other.[1] For example, in the absence of conditioning on Z, increasing X by one unit also increases Z by one unit, but crucially does not affect Y. In contrast, increasing X by one unit when holding Z constant means Y must decrease by one unit. This is because Z is a collider on the path between X and Y.

Now, suppose we consider X, Y, and Z in relation to a subsequent outcome O. In the absence of conditioning on Z, increasing either X or Y can only affect changes in O via their influence on increasing Z; Figure 4.2.1b depicts this scenario, in which only indirect causal pathways (i.e. through Z) to O from each of X and Y exist. However, conditioning on Z blocks off these indirect paths, implying that changes in X or Y must affect changes in O directly; this is depicted in Figure 4.2.1c (where the rectangular box around Z indicates conditioning).

Without loss of generality, suppose we are interested in the causal effect of the component X on the outcome O. Figures 4.2.1b and 4.2.1c indicate the existence of two distinct effects, respectively:

1. The 'unbiased' (total) effect of X on O: This estimand captures the effect on O of increasing X (and thereby increasing Z) regardless of Y; it may be estimated with or without conditioning on the component Y.
2. The 'collider biased' effect of X on O: This estimand captures the effect on O of increasing X while simultaneously decreasing Y; it may be estimated by conditioning on the collider Z (with no conditioning on Y).

Figure 4.2.1: DAGs depicting three random variables X, Y, and Z, for which X + Y = Z. Double circles around a variable indicate that it is a deterministic function of its parents, and a rectangular box around a variable indicates that it has been conditioned upon. **A.** X and Y are unconditionally independent. **B.** X and Y are unconditionally independent, and may only affect a subsequent outcome O via their influence on Z. **C.** X and Y are conditionally dependent, and may only affect a subsequent outcome O directly.



In the setting of compositional data, where the collider Z is fully determined by its component parts, both effects may be of interest depending upon the context; indeed, this is contrary to perceived wisdom in the generic case, in which collider bias is considered undesirable for causal analysis (as the word 'bias' suggests). We discuss these two effects in the context of several example scenarios involving compositional data (with both *variable* and *fixed* totals), and the implications for causal analyses involving data of this kind. Note that we will continue to refer to collider 'bias' throughout but emphasise that due to the unique nature of compositional data, it is not truly an interpretational bias as long as the desired effect is appropriately interpreted.

Note: The coefficients for X and Y in the model O ~ X + Y will be identical if the true underlying causal relationship is between Z and O; if X and Y cause O distinctly and separately, the coefficients for X and Y may (by chance) be very similar, but they are not identical and are distinctly estimated.

Compositional data with variable totals

We first consider causal inference for compositional data with variable totals, which are compositional data for which the 'total' can vary across units of analysis. Examples include:

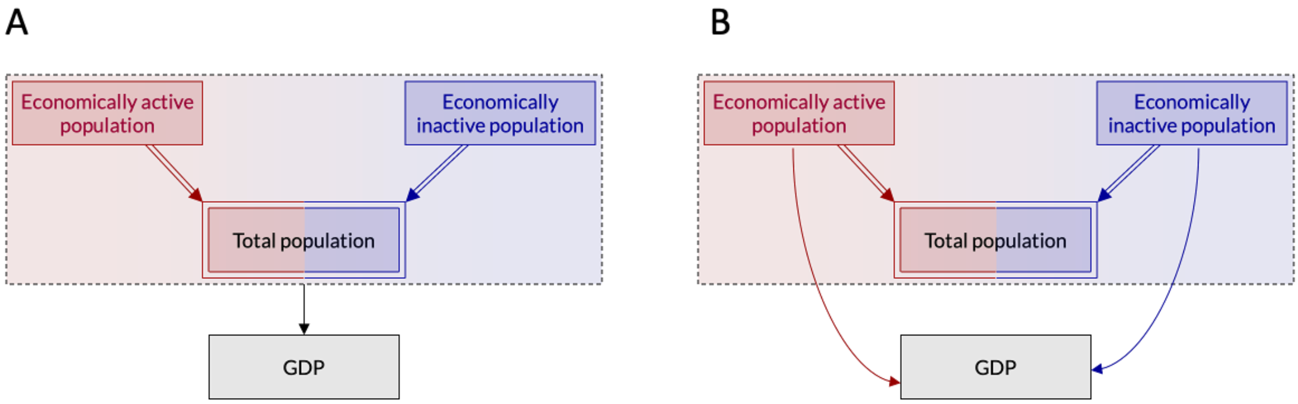
- Total height (decomposed into leg length and trunk length)
- Total fat mass (decomposed into brown fat mass and white fat mass)
- Total population (decomposed into those aged 0-18, 19-35, 36-60, and 61+ years)

We consider the 'unbiased' and 'collider biased' effects for two specific example scenarios, and the resulting implications for compositional data with variable totals.

Scenario 1: Economically active population and gross domestic product (GDP)

Suppose we are interested in estimating the causal effect of the total number of economically active individuals within a geographical area on the area-level gross domestic product (GDP). The DAG in Figure 4.2.2 represents this scenario, which also explicitly depicts the compositional nature of the exposure (i.e. economically active population + economically inactive population = total population); confounders are omitted for ease of illustration. In this scenario, both 'unbiased' and 'collider biased' estimates of the effect of the economically active population on GDP are obtainable, and both may be of utility depending on the context.

Figure 4.2.2: DAGs depicting the total population in relation to gross domestic product (GDP), in which total population is subdivided by economic activity (i.e. total population = economically active population + economically inactive population). Double circles around a variable indicate that it is a deterministic function of its parents, and a rectangular box around a variable indicates that it has been conditioned upon. **A.** Economically active population and economically inactive population are unconditionally independent, and may only affect GDP via their influence on the total population. **B.** Economically active population and economically inactive population are conditionally dependent, and may only affect GDP directly.


A
Economically active population
Economically inactive population
Total population
GDP
B
Economically active population
Economically inactive population
Total population
GDP

The 'unbiased' effect of the economically active population represents the average change in GDP that results from adding economically active individuals to the area, thereby increasing both the number of economically active individuals and the total number of individuals, whilst doing nothing to the population of economically inactive individuals. An estimate of this effect may be of interest if, for example, a government is considering policies aimed at increasing economic immigration.

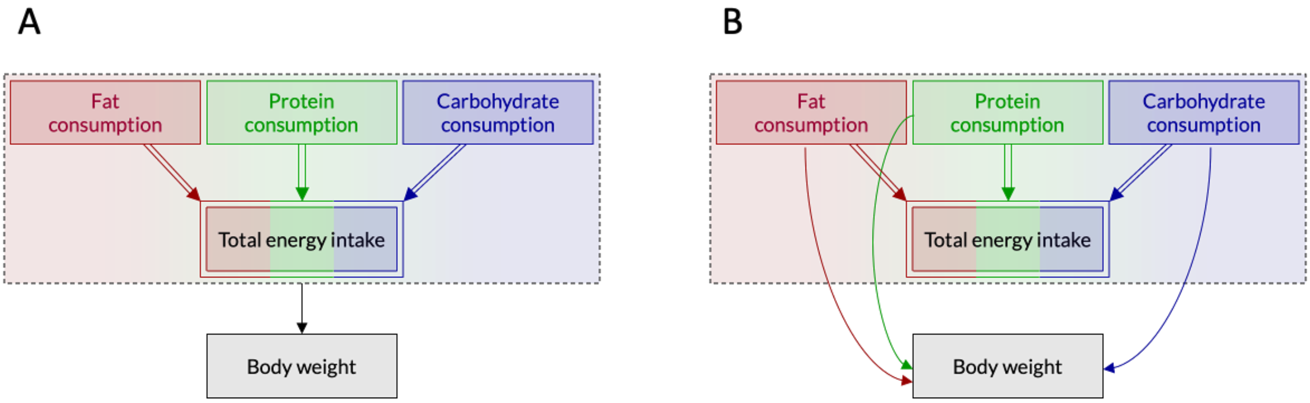
In contrast, the 'collider biased' effect of the economically active population represents the average change in GDP achieved by swapping economically inactive individuals for economically active individuals – either by adding economically active individuals and removing an equal number of economically inactive individuals, or by effectively converting economically inactive individuals to economically active individuals (or some combination thereof). The 'collider biased' effect is therefore a combination of the 'unbiased' effects of both subgroups on GDP – the positive effects of simultaneously increasing the economically active population and decreasing the economically inactive population by equal numbers, thereby retaining the same total population. An estimate of this effect may be of interest if, for example, the government were considering job-training programmes for currently unemployed individuals.

In this scenario, both the 'unbiased' and 'collider biased' effects reflect the population-level average effects of changing the relative numbers (i.e. the proportions) of economically active individuals to alter GDP, but by different mechanisms; they therefore reflect distinct causal quantities, either of which may be of interest depending on the context or proposed intervention.

Scenario 2: Fat consumption and body weight

Suppose we are interested in estimating the causal effect of fat consumption on weight. The DAG in Figure 4.2.3a represents this scenario, which also explicitly depicts the compositional nature of diet (i.e. fat consumption + protein consumption + carbohydrate consumption = total energy intake).

Figure 4.2.3: DAGs depicting total energy intake in relation to body weight, in which total energy intake is subdivided by macronutrient consumption (i.e. total energy intake = fat consumption + protein consumption + carbohydrate consumption). Double circles around a variable indicate that it is a deterministic function of its parents, and a rectangular box around a variable indicates that it has been conditioned upon. **A.** Fat, protein, and carbohydrate consumption are unconditionally independent, and may only affect body weight via their influence on the total energy intake. **B.** Fat, protein, and carbohydrate consumption are conditionally dependent, and may only affect body weight directly.



The 'unbiased' effect of fat consumption represents the average change in weight that results from adding fat to an individual's diet, irrespective of the consumption of all other macronutrients, which consequently increases total energy intake without altering other consumption behaviours. An estimate of this effect may be of interest if, for example, individuals were considering a diet that advocated reducing and/or eliminating fat and not replacing it with other macronutrients (e.g. by reducing or eliminating cooking oil).

The 'collider biased' effect of fat consumption represents the average change in weight that results from replacing 'other' macronutrient consumption (i.e. protein and carbohydrate consumption, in their relative proportions) with fat consumption, thereby increasing fat consumption without increasing total energy intake. An estimate of this effect may be of interest if, for example, individuals were considering a diet which advocated replacing fat from their diet with 'other' macronutrients (e.g. replacing high-fat foods with their lower-fat counterparts).

As with Scenario 1, each effect captures a different mechanism for increasing the relative amount of fat intake, and each may yield radically different quantities according to different contextual interpretations. Whilst both effects arguably have meaningful causal interpretations, each must be considered carefully, and their interpretation made explicit according to the context sought.

Implications

For analyses involving compositional data with variable totals, both the 'unbiased' and 'collider biased' effects of a component may be estimable and meaningful, depending upon context. However, care must be taken when interpreting the 'collider biased' effect for a component of compositional data with a variable total, as the estimates derived represent the effect of one component relative to the component(s) which have been omitted from the analysis.

In the instance that only two components are considered (e.g. Scenario 1), conditioning on the total uses one degree of freedom, meaning that the two components share only one degree of freedom between them and thus represent just one single (binary) variable (i.e. economically active and not economically active). In such a scenario, the 'collider biased' effect of the component of interest is unavoidably interconnected with the effect of the other component; it represents the effect of replacing the first component with the second, which is equal and opposite to the effect of replacing the second component with the first. The two effects are intractably entangled, and the causal effect of each component only has meaning relative to the other.

Where three or more components are considered (e.g. Scenario 2), this means that the 'collider biased' effect represents the effect of one component relative to a combination of the omitted components. Whether this reference level provides a meaningful contrast must be determined by context, though more specific contrasts may be achieved by conditioning on more components. This issue has relevance in nutrition research, where ‘other’ consumption often makes up a substantial part of total energy intake due to the deconstruction of individual diets into smaller and smaller components. In such scenarios it may be more informative to look at the net effect of a component (i.e. the ‘unbiased’ effect) rather than relative effect (i.e. the ‘collider biased’ effect).

Compositional data with fixed totals

We now consider compositional data with fixed totals – i.e. compositional data for which the ‘total’ is fixed to the same value for every unit of analysis. These types of data usually involve some standard unit of measurement (e.g. time or space) that is fixed by nature or convention. Examples include:

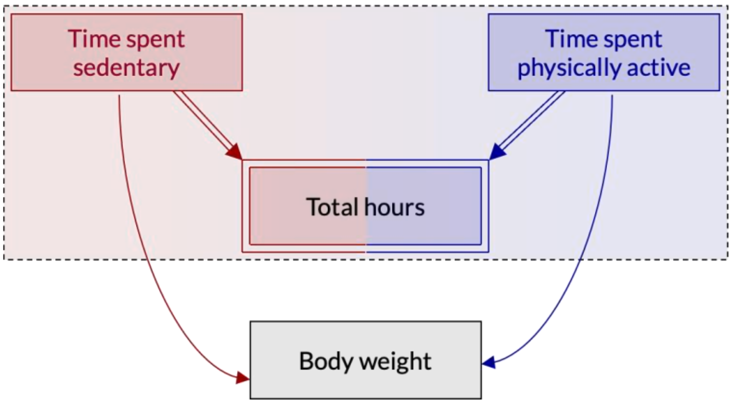
- Hours per week (decomposed into time spent commuting, time spent working, time spent sleeping, and ‘other’)
- Boeing 747 capacity (decomposed into adult passengers, child passengers, and vacant seats)
- Child benefit block grant (decomposed into money spent directly on the child and money not spent directly on the child)

We consider one specific example scenario and discuss the resulting implications for compositional data with fixed totals.

Scenario 3: Time spent sedentary and body weight

Imagine we are interested in estimating the causal effect of total time spent sedentary (i.e. not moving, including: sleeping, sitting, and standing) per day. Because total hours per day is fixed at 24 for every individual, it is more accurately described as a constraint for time spent sedentary and time spent physically active (i.e. time spent sedentary + time spent physically active = 24 hours). Nevertheless, it may be useful to consider the length of day as a ‘variable’ within a causal framework, as though it were possible to condition thereupon, since such a representation neatly illustrates the challenges associated with causal inference for compositional data that are inherently constrained. The DAG in Figure 4.2.4 represents this scenario, with confounders omitted for ease of illustration. Total hours is depicted as a deterministic function of time spent sedentary and time spent physically active which has also been automatically conditioned upon.

Figure 4.2.4: DAG depicting *total hours* in relation to *body weight*, in which *total hours* is subdivided by activity category (i.e. *total hours = time spent sedentary + time spent physically active*). Double circles around a variable indicate that it is a deterministic function of its parents, and a rectangular box around a variable indicates that it has been conditioned upon. *Time spent sedentary* and *time spent physically active* are dependent due to the inherent constraint (‘conditioning’) upon *total hours* (i.e. *total hours* = 24), and thus may only affect *body weight* directly.



It is impossible to estimate the effect on weight of increasing time spent sedentary whilst maintaining time spent physically active (i.e. the ‘unbiased’ effect) because total hours is a fixed quantity. Any increase in time spent sedentary must be accompanied by an equal and opposite decrease in time spent physically active. In this scenario, only the ‘collider biased’ effect of time spent sedentary is estimable; moreover, it is the only effect that has any causal interpretation. The ‘collider biased’ effect respects the inherent constraint imposed by the fixed length of a day and represents the average change in weight produced by swapping time spent physically active for time spent sedentary – which is equal and opposite to the effect of swapping time spent sedentary for time spent physically active.

Implications

For analyses involving compositional data with fixed totals, only the ‘collider biased’ effect of a component is identifiable or potentially causally meaningful. The inherent constraint upon a fixed total operates in a similar fashion to conditioning on a variable total, and results in an estimate that represents the effect of one component relative to the other component(s) which have been omitted from the analysis.

In the instance in which only two components are considered (e.g. Scenario 3), the total constraint (i.e. exactly 24 hours in a day for everyone) means that the two components share only one degree of freedom and are therefore implicitly just one (binary) variable (i.e. time spent sedentary and time spent not sedentary). It makes little sense to even conceptualise the two components as having separate effects, since each variable may only be defined and estimated relative to the other. This is important for discussions regarding the relative merits of decreasing one component versus increasing another (e.g. decreasing sedentary behaviour versus increasing physical activity[6-8]), as the two are not distinct entities from a causal perspective. Where more than two components are considered (e.g. where time spent physically active is further subdivided into light, moderate, and vigorous exercise), care should be taken to consider and select the most meaningful reference component(s).

Composite variable bias

We previously introduced compositional 'totals' or 'wholes' as special cases of composite variables. As such, they carry all the implications of what we term 'composite variable bias', a family of different biases and issues that often arise in the analysis of composite variables. It might then seem unusual to suggest that, for instance, to estimate the relative causal effect we need to adjust for the 'total'. As a type of composite variable, it would conflate the effects of its constituent components and would also suffer from 'information loss' because the different variances of the components cannot be captured by the variance of a single variable. Indeed, we have shown this to be the case with simulations, where the effect of total energy is not equal to the weighted average effect of the nutrient components.[113]

The key to avoiding composite variable bias with compositional totals is in the modelling strategy. Instead of modelling the compositional data as either 'total' or 'other' (i.e. total minus the exposure), a way to avoid using composite variables is to include all constituent components as individual covariates, an approach we term 'the all-components model'. Suppose we have a scenario in which we are seeking to estimate the effect of non-milk extrinsic sugars (NMES) on fasting plasma glucose concentration (GLUC). A model including all constituent components of energy intake would look like this:

$$\widehat{GLUC} = \hat{f}_0 + \widehat{\boldsymbol{f}_1}NMES + \hat{f}_2CRB + \hat{f}_3FBR + \hat{f}_4SF + \hat{f}_5UF + \hat{f}_6PRO + \hat{f}_7ALC$$

We can see that this model adjusts for all competing sources of energy other than the exposure, i.e. all 'other' energy intake. It therefore targets the *total* causal effect of NMES on GLUC. By avoiding the use of the composite 'other', this approach provides an estimate of the total causal effect that is free from composite variable bias. If, instead, the average *relative* causal effect was sought, it can be estimated using this model by subtracting the weighted average of the estimated effects of all other components from the total causal effect of the exposure, i.e. $\hat{g}_1 = \hat{f}_1 - \left[\sum_2^n w_i\hat{f}_i\right]$, where w_i is the proportion of the remaining energy intake contributed by each component $i = \{2, \ldots, n\}$). Therefore, the all-components model does not only estimate unbiased effects by avoiding composite variable bias, but also provides a convenient approach with which both the total and any relative causal effect of interest may be estimated using a single model.

Conclusion

Although collider bias is generally agreed to be problematic for causal analyses, compositional data present a unique situation in which meaningful interpretation may be obtained. Scenario 1 presents one such example, in which compositional data have a variable total and the omitted component provides a meaningful reference category for the component of interest. Where the omitted component(s) do not provide a meaningful reference category, as perhaps argued in Scenario 2, the total effect likely provides a more important estimand. In all situations involving compositional data, it is paramount that researchers explicitly consider and declare which causal effect is sought and how it should be interpreted, since the total (i.e. 'unbiased') and 'collider biased' effect estimates may be radically different, even if both are causally meaningful. For example, the effect on cardiovascular disease of eating red meat on top of an otherwise healthy diet may be drastically different to the effect of replacing 'healthy' dietary components with red meat. Insufficient clarity regarding the distinction between these two effects likely contributes to ongoing confusion due to apparently contradictory results.[9;10]

For compositional data with fixed totals, as in Scenario 3, collider 'bias' is unavoidable. Notably, the relative effects that characterise these types of data are well-recognised in other contexts. For instance, categorical data may be conceptualised as a trivial case of compositional data, in which the total is fixed at one. This notion is implicit in the coding of such variables for statistical analysis – each category is treated as a binary variable with value zero or one, and the sum of all categories for every individual equals one (i.e. each individual may belong to one and only one category). In such situations, one category must be specified as the reference category and all other effect estimates must be interpreted relative to this category.

5.2 Composite Variable Bias

Learning objectives

- Understand the different types of composite variables
- Recognise the 'meaning' of a composite variable, identify the appropriate estimand, and understand composite variable bias
- Explain why use of composite variables may violate the assumptions of exchangeability and consistency

Introduction

Potential analytical and interpretational problems increase substantially when moving from derived variables, where there is just one parent variable, to composite variables, where there are two or more parent variables. This may be because the underlying tautology is simply less obvious once there are multiple variables involved, despite decades of warnings about the analysis of certain composite variables.[98,102,105] For example, at least part of any observed relationship between gross domestic product (GDP) and homicides is tautological when they are analysed on a per-capita basis (i.e. GDP/population and homicides/population) since both share the same population denominator. However, there are further challenges in the analysis of composite variables that have received far less attention.

Composite variables are generally constructed for one of the following purposes: (1) to create a variable that aims to **summarise** multiple related concepts in a convenient or parsimonious way (e.g. metabolic syndrome, deprivation index), perhaps capturing a latent concept (e.g. deprivation), or (2) to **standardise** one variable by another (e.g. body mass index [BMI], GDP per capita). The distinction between these two purposes is not trivial and indeed has important implications for determining the appropriate analytic strategy. Crucially, summarisation implies an interest in modelling and understanding the *average effect* of a series of related concepts on an outcome of interest, whilst standardisation implies an interest in modelling the *conditional effect* of a variable on the outcome, where the variable conditioned upon is deemed to 'confound' the focal relationship. There is a third type of composite variable, distinct from those 'constructed' with a specific purpose in mind, because they occur 'naturally'. These are compositional data **totals** or **sums** (e.g. total population numbers, total energy intake) which are considered composite because they comprise several constituent components. It might not be intuitive to think of compositional totals as sources of 'bias' but all of the issues that we describe also apply to these types of variables.

We have already discussed how deterministic relationships that operate in composite variables give rise to tautological associations. As a reminder, a tautological association is a 'self-fulfilling' association that arises when a variable is analysed in relation to itself, a part thereof, or a 'sibling' variable with shared algebraic components[115], e.g. mathematical coupling may be observed in analyses of change or ratio variables. Tautological associations are relatively straightforward to identify once the composite variables and their parents are depicted in a DAG. However, composite variables may lead to problems beyond mathematical coupling that are less evident from the model itself and require thinking about the causal estimand of interest and what the variables used in the analysis actually mean.

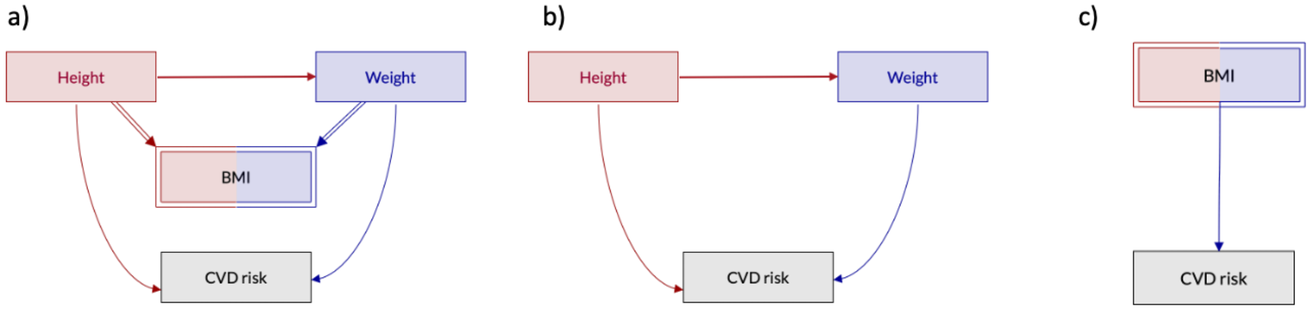
Composite variable bias

Composite variable bias is not a specific single type of bias but instead a range of problems that may arise in the analyses of composite variables. To illustrate how and why each potential problem occurs, we consider the causal effect of BMI on cardiovascular diseases (CVD).

Estimand confusion

Although the use of BMI is ubiquitous in health and medical research, its status as an algebraic construct determined by body mass and height is often overlooked. Using deterministic notation, we can depict our scenario by the DAG in Figure 4.1.1a. If we apply what is known as Deterministic Node Reduction algorithm[116], this would be reduced to the DAG in Figure 4.1.1b. From a strictly correlative perspective, BMI provides no information beyond that provided by height and weight. Indeed, for any DAG containing a variable X that is fully determined by a set of parent variables, denoted $pa(X)$, conditioning on $pa(X)$ renders X independent of *all* other variables in the DAG.[117] This is why previous commentators have argued that "no causal knowledge is gained by estimating a non-existent effect of body mass index".[118] However, this assumes there is no utility in the composite variable; or in dimensionality reduction in general. If we did decide we were less interested in the individual effects of height and weight, or we no longer possessed the source information on height and weight separately, then we might use the so-called 'parent-divorcing' algorithm to provide the DAG in Figure 4.1.1c[114].

Figure 4.1.1. (a) DAG depicting relationships amongst height, weight, body mass index (BMI) and cardiovascular disease (CVD) risk; note that BMI is drawn in double circles to indicate that it is a deterministic function of height and weight. (b) DAG from (a) after applying the Deterministic Node Reduction algorithm and reducing BMI from the diagram. (c) DAG from (a) following the algorithm that removes all deterministic parents of BMI.



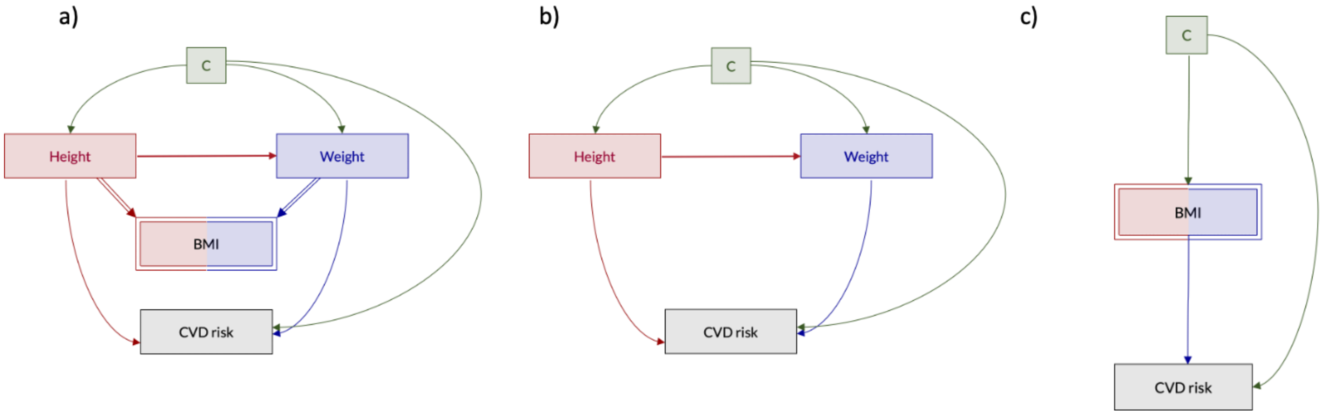
Which of the two DAGs - 1b or 1c - is the more appropriate for causal inquiry? The answer cannot be answered algorithmically. Instead, it depends upon the value and meaning that we give to the average causal effect of the composite exposure, BMI. In other words, it depends on whether or not we think that BMI represents a meaningful summary of height and weight that serves as a useful proxy for another more clearly defined concept (e.g. adiposity) or is simply a measure of weight 'standardised' by height (to account for the fact that taller people are generally heavier).

If we consider BMI to be a valid and useful proxy for adiposity, then analysing the composite as a distinct variable is arguably better, though still not without issues, as we discuss later. Alternatively, if we consider BMI to be a measure of weight 'standardised' by height, then we must carefully consider whether weight/height^2 represents the most effective parameterisation of this relationship, or whether the causal effect of weight is better captured by exploring weight directly while conditioning on height (i.e. through 'statistical adjustment' in a regression model).

The total causal effect of BMI on CVD risk will likely differ from the total causal effect of weight on CVD risk, conditional on height; although both can be theoretically estimated without *statistical* bias, we risk inferential bias if the effect estimate we obtain does not reflect the causal mechanism that we seek to understand and may eventually wish to target for intervention.

Even if we are content with the validity and interpretability of our composite variable, additional complexities may arise if the parent component variables 'crystallise' at different points in time (i.e. if their values are not determined contemporaneously as part of the same causal processes). In Figure 4.1.2, we have added another variable C which confounds the relationship between BMI and CVD risk. Although this variable seems innocuous in the parentless DAG shown in Figure 4.1.2c, DAGs containing the parents of BMI (Figures 4.1.2a-b) reveal that the status of C as a confounder might be more complex than initially understood. Whereas height remains fixed throughout most of adulthood (having crystallised during young adulthood), weight represents an ever-changing accumulation of influences across the lifecourse; therefore, any measurement of weight in adulthood is likely to have crystallised much closer to the time at which it was measured. Consequently, height and weight represent distinct entities in time, and the potential implications of collapsing them into one node (i.e. BMI) has received little consideration.

Figure 4.1.2. DAGs corresponding to those given in Figures 4.1.1a-c, respectively, but with an additional variable C which 'confounds' the relationship between BMI and CVD risk.



For example, if C represents biological sex, we can safely conclude that the direction of causation flows from C to each of height and weight because sex crystallises at conception. If C represents marital status, it is more plausible to draw a causal arrow from C to weight but not to height. If C represents participation in sport, however, we would likely draw causal arrows from C to weight and *from* height to C. In this final scenario, we are presented with a situation in which one of our supposed confounders of the relationship between BMI and CVD risk is simultaneously a confounder for one of the elements of BMI (i.e. weight) and a mediator for the other element (i.e. height). When C more realistically represents multiple variables, each of which may have different relationships with the individual components of BMI, the complexity of the situation multiplies, and the meaning of the 'causal effect' of BMI on risk of CVD becomes ever harder to define, derive, and interpret. In any situation where the parents have different confounders, or different relationships with the same confounders, then it can be argued that the summary effect of the composite child variable cannot ever be identified.

Exchangeability violation

The scenario in which a variable like C is, in relation to BMI, both a parent of weight and a child of height leads to issues that are beyond just confusion of the estimand that is being estimated. As a reminder, units of analysis are considered conditionally exchangeable when they do not differ systematically, i.e. when the propensity of the outcome is independent of the assignment of the exposure, conditional on all confounders. When a composite variable like BMI is used, it conflates the otherwise separate variables, such as weight and height, which might occur (or 'crystallise') at different points in time (i.e. height is relatively stable in adulthood, whereas weight may vary substantially). When a variable is then, say, a cause of weight but caused by height, it would operate as a confounder if weight was the exposure, but a collider if height was the exposure. When the two variables are conflated, it becomes impossible to adjust for all confounders and none of the mediators. On the one hand, we have to adjust for it as a confounder in order to achieve conditional exchangeability, but on the other hand, we should not adjust for it as a collider because we risk invoking collider bias. It is even more problematic when we do not *realise* that BMI conflates two different variables because we would not be able to correctly identify the confounders that need to be conditioned upon, which is required to achieve conditional exchangeability.

With a simple composite variable like BMI, we might be lucky in the rare scenario in which we recognise BMI as being constructed by weight and height and seek to be sure that no other variables occur between them. However, once we move on to more complex composites and use several very different variables to construct them, it is highly unlikely that all of the constituent components would occur at the exact same point in time. It would be therefore inevitable that some variables would be ancestors of some constituent components and descendants of others. If we do not realise that this is the case, we risk not conditioning on important confounder variables, therefore violating the exchangeability assumption. Even if we do realise the nature of the composite variable and its constituent components, it becomes increasingly tricky to condition on all relevant confounders that are not also mediators of other components.

Consistency violation

The interpretation of 'obesity' as a definable exposure with an identifiable causal effect has previously been challenged;[118,119] in particular, there are concerns that obesity fails to satisfy the consistency assumption[120] required for causal inference because it can represent multiple states, including high adiposity and high muscle mass.[121,122] The same concern is clearly relevant for BMI – and indeed all composite variables – since any value of the composite may represent various combinations of the determining component parents. Hypothesising that BMI 'causes' an increased risk of CVD implies that intervening to lower BMI would result in a decreased risk of CVD. While theoretically, for any one individual, we could lower BMI only by either decreasing weight, since we cannot realistically intervene on, the estimate derived for the causal effect of BMI relates to both differences in weight *and* height *between* individuals – the estimate is thus derived with two (hypothetical) mutable targets for intervention. Regardless of our philosophical perspective on the utility and validity of BMI, this realisation suggests that, if we are interested in the causal effect of BMI on CVD risk with a view to modifying weight only, it might be more useful to estimate the causal effect of weight adjusted for height.

This, while for most adults height is relatively fixed throughout life and any change in BMI would come from a change in weight, because this is only true on a within-person level and the causal effect estimate also incorporates the population as a whole, allowing for between-person contrasts, we see that for a single value of BMI, there are infinite possible combinations of weight and height. When parent variables have different values and influence the outcome of interest in different ways, it would be naïve to expect that a summary composite variable, such as BMI, could consistently capture the causal effects of parent variables regardless of their specific values. The 'weight' variable does not differentiate between adipose tissue, muscle tissue, or bone tissue, introducing further (and even more complex) variations that BMI cannot capture. We cannot therefore expect that BMI, or any other composite variable, to have a consistent effect on CVD.

The violation of the consistency assumption is relatively easier to identify in composite variables used for summary or standardisation and may be less obvious in compositional data totals, but exactly the same principles apply. Any 'total' or 'whole' variable (e.g. total energy intake) can be achieved by an infinite number of different food or nutrient combinations. We cannot expect that an average daily intake of 2000 kcal that predominantly comes from fat to have the same causal effect as 2000 kcal that predominantly comes from protein. Compositional totals rarely, if ever, produce consistent causal effects that do not depend on the variation and proportion of the specific components they comprise; the extent to which this is detrimental to the study will vary depending on the example and context considered.

Loss of estimating performance

When composite variables are created with the aim to either summarise or standardise, it is probably hoped that a single composite variable parsimoniously captures all the 'information' that its parent variables carry. For example, it may be argued that metabolic syndrome is a lot more informative than individual 'markers' of metabolic health because it is a summary of five different metabolic conditions. Similarly, BMI is calculated to capture weight and height together and standardise weight against height.

A lot of the 'information' that parent variables possess lies in their variance. More specifically, in how they vary in relation to the outcome. For example, when the variation of one variable sufficiently 'captures' the variation of another variable (e.g. the outcome), we consider it a good predictor because it 'explains' (i.e. gives information on) variation in the outcome. When prediction is not the goal, of interest is the variation in the outcome that is attributed to variation in the exposure. This is key when we are interested in how a variable causally relates to other variables.

When different variables are summarised into a single composite, it becomes evident that most of the variation (and information) that they carry may be lost. This is conceptually similar to dichotomising a variable (i.e. turning an otherwise continuous variable into a binary according to a specified threshold), where all values are collapsed into (usually) 0's and 1's and most of the variation in the variable is 'lost'. Dichotomisation is done for convenience but, while sometimes useful, this inevitably leads to loss of information that may be otherwise key for the analysis sought. This 'information loss' may not always be serious, but it becomes problematic when we expect that a, say, summary variable would truly have a 'summary' effect, e.g. the average effect of the constituent components combined. Or that the effect of BMI would be equal to the effect of weight divided by the squared effect of height.

To illustrate, consider an example using the type of composite variable that is arguably expected to introduce least bias – compositional totals, which occur 'naturally' and are free from many of the complications arising as a result of 'artificially' constructing composites. Suppose we were interested in the effect of diet on systolic blood pressure (SBP) and total energy intake comprised energy from protein, carbohydrates, fat, and alcohol (without breaking down the macronutrients further, for ease of illustration); a 2000 kcal daily intake could be 500 kcal for each nutrient (intentionally oversimplified). Each macronutrient would have an individual causal effect on CVD that may be estimated. For example, assume, for the purpose of demonstration, that the effects of 100g of each nutrient on SBP were as follows: protein, -3 mmHg; carbohydrate, -1 mmHg; fat 3 mmHg, alcohol 4 mmHg. We would expect then that the effect of the overall energy intake would be the average effect of all nutrients, assuming no residual confounding. If we do not expect this, then it is not clear why we would expect the creation of a summary composite variable to provide true summary effects (fallacy of composition). Our expectation would be that the effect of total daily energy intake would be the average of the effects of all macronutrients, i.e. 3 mmHg. However, it has been shown using simulations that this would not be the case.[113] It is likely that even if the effect of total energy is very close to 3, it would not actually equal it. This is because the variation that each macronutrient variable carries is 'lost' when we use total energy intake as a summary of the different sources of energy. It is not possible to perfectly capture all of them, and the more the variations of the constituent components differ from each other, the less the 'total' (or any composite) will be able to provide the true average.

We would similarly expect all composite variables, regardless of how they were constructed, to suffer significant information loss whenever the constituent components have very different variances, because a single variable would not be able to sufficiently capture all of them. This is particularly important, and raises serious concerns, for summary composite variables because they are created with the purpose of providing a 'summarised' effect.

APPENDIX

6.1 Algorithmic Transparency, Explainability & Interpretability

Learning objectives

- Describe transparency, explainability, and interpretability as three distinct aims
- Explain why transparent and/or explainable models are not necessarily interpretable
- Understand why algorithmic interpretability requires causal inference

Introduction

Machine learning algorithms, and artificial intelligence (AI) methods in general, have recently gained considerable attention in the field of medicine and health. They have been used with different levels of success across various settings and contexts, such as diagnostic imaging[123], electronic health records[124], and genetics[125]. Some are well-known and extensively used in healthcare already, such as the QRISK®[126] algorithm for estimating future risk of cardiovascular disease, which is part of the NHS Health Check for adults, and PREDICT[127] – a model that predicts how different breast cancer treatments might improve survival rates after surgery. We are often led to believe that such predictive tools could revolutionize healthcare by predicting patient outcomes, assisting with diagnosis, or determining the most appropriate treatment. In reality, algorithms are fundamentally unsuited to individual-level predictions and cannot meaningfully inform interventions unless models are grounded in causal inference (i.e. counterfactual prediction).

The lack of individual or single event predictions is true even in cases where we know, with complete or near certainty, the exact probabilities of different potential outcomes, such as when rolling a pair of dice. Our ability to 'predict' a single roll of a die is not improved by knowing the probabilities of all potential rolls. Similarly, we may know the differences in probability of developing a certain health outcome (e.g. cancer) between different groups of people, but we can never determine with certainty whether a particular person will develop the outcome or not. Prediction models are excellent at identifying and 'learning' patterns and associations within data, but they are unable to determine what these patterns 'mean' and why they occur.

Meaningful interventions require knowledge of 'what will happen if we do such and such …', not 'what will happen if we do nothing and just watch and wait'. Consequently, without a causal inference framework, our models cannot provide insights and confer meaning to the patterns they detect in the data. We demonstrate this in more detail later.

The main consequence of these limitations, when relying on such algorithms for decision making, is that we risk amplifying any biases or prejudices that were present in the data used for training a model (i.e. building a model that has 'learned' from the data), and this need not at all apparent or even discoverable from careful scrutiny of the model itself. This issue has given rise to a lot of discussions around the need for 'fair' AI and methods for maximising AI fairness. Algorithmic transparency and the overall idea of models being 'transparent', both with regards to how they are built and how they are subsequently used, is often considered a key aspect of ensuring models are fair and unbiased.

What is Transparency?

Algorithmic transparency has been defined as "the disclosure of information about algorithms to enable monitoring, checking, criticism, or intervention by interested parties"[128]. To satisfy these criteria, a model should be transparent in terms of 1) how it was built (e.g., what data were used for training, what was the methodological processes that led to the final model), 2) what the final model (e.g., what variables or features are used, how they are parameterised, what the coefficients/weights of each variable or feature are), and 3) how it is used (e.g., what the model is used for, in what setting, how the model decision/prediction is interpreted). Without such transparency, critiquing a prediction model and the decisions that it makes is extremely challenging.

The opposite of a 'transparent' model is known as a 'black-box' model. Black box models use information from all available inputs (variables) to predict or determine the output (outcome) of interest, but this process does not provide us with the knowledge of how a model arrives at a decision in a way that is interpretable by a human being. For example, neural networks are very powerful prediction and classification tools, but the increased complexity of their multilayer non-linear structures makes it impossible to gain any meaningful insights about the learnt function from the parameters/weights. If we look at a simpler model, say a logistic regression, we can easily understand how the model produces a given output and how tweaking the model coefficients would alter the output. But something like a neural network is so complex that we cannot extract any meaningful information about *how* the output is determined, therefore similar approaches have been termed 'black boxes'.

Without knowledge of how a model operates, 'black box' algorithms may encode biases and prejudices. It has been argued that algorithmic transparency is key for achieving algorithmic fairness[129]. It is no surprise, therefore, that a lot of attention has been directed towards methods for 'opening' such black boxes, which claim to transform 'black-box' models by discovering the learned functional forms into fully transparent models, often termed 'white box'[130–132]. It has to be said, however, that the extent to which such models are actually transparent is questionable. Most approaches aim to extract the relative feature (i.e. variable) *importance*, for example, by using symbolic metamodels[131], but this does not provide the same level of transparency as specific coefficients and model parameterisations that are straightforward to extract from a simpler model. It can be argued that something like feature 'importance' offers enough transparency to make the model *explainable*, because we can see which variables 'drive' the model decision, but it takes more than this to call a model truly 'explainable'.

What is Explainability?

An algorithm or an AI is considered explainable when humans can provide explanations about its behaviour and decision-making process[133]. For example - going back to the QRISK prediction model for estimating future cardiovascular risk – by simply looking at the model we can get information on things like the size of each variable coefficient, any interaction terms, polynomial transformations, etc. We have all the information that we need to *explain how* the model estimates that a young healthy non-smoker's 10-year risk of heart attack or stroke is 0.1%, whereas the risk for a middle-aged heavy smoker with type 2 diabetes is 20%. We can tweak the model parameterisations and immediately observe the difference in how the model behaves, in a way that is not only transparent, but also straightforward to explain. However, this would not be the case if, say, we only had information on the relative 'importance' of variables within the model, which may often be the case when black-box models are 'opened' using some of the recent methodological developments. We can speculate that a feature with very high estimated 'importance' would drive most of the change in the estimated output (without interpreting this as an effect, but simply the result of a highly correlated predictor), but this is not sufficient to explain the exact process by which the model computes a *specific output value.* All that it can do is provide a very vague explanation about the relative weights of features, but a model should not be considered sufficiently explainable if it fails to provide information on how it arrives at a specific outcome for an *individual.*

It is probably evident at this point that transparency and explainability are very closely related. Sufficient explainability requires complete transparency of a model, and a model cannot be explainable if it is not transparent. But they are still separate *aims* – transparency is concerned with making the 'ingredients' of a model fully available, whereas explainability provides detailed information on how the model produces a specific output.

The development of methods for improving (or at least attempting to improve) algorithmic transparency and explainability is an impressive achievement in the field of computer science. However, this approach has been uncritically praised when it comes to its potential real-world application, leading to the misleading perspective that if the resulting models are sufficiently transparent and/or explainable this would increase trust in the users[131]. We have shown that this is not the case, but there is an even more dangerous 'promise' that some computer scientists make – that 'white box' models, by virtue of their improved transparency and explainability, are also *interpretable*[134].

What is Interpretability?

Model interpretability goes a step further than explainability, and instead of asking *how* the algorithm computes an output, it asks *why.* This happens to be not only the most difficult question to answer, but is also the question that everyone wants to know. When a person goes to their GP and undergoes their NHS Health Check, they want to know *why* they are identified as high-risk by the QRISK algorithm instead of *how*, i.e. they might be interested in finding out whether they can do something to *change* their risk, where providing details on model parameterisations and interaction terms would carry very little value. In seeking this distinction, there is a high probability of inadvertently arriving at the wrong conclusions.

For instance, a GP might be tempted to say to the patient that their high-risk is because of a certain variable like their age, BMI, smoking status, or even postcode, if it seems to them that the algorithm's score increases because of these features within the model. But this is fundamentally wrong to do, because the feature weights of a predictive model are completely uninterpretable[135–137]. This might not seem very intuitive to understand when using QRISK as an example, because factors such as adiposity, diabetes, postcode (sometimes used as a proxy of deprivation), and comorbidities are all things that *do* increase the overall risk of a cardiovascular event and many of these would be causally related to the risk of such an event. But not all variables are necessarily causal and their contribution to a prediction model cannot tell you to what extent their prediction is causal, if at all.

Let us assume, for the purpose of demonstration, that we build an alternative model which contains only two variables – baldness and shoe size, and we use it for the same purpose – to try and predict the 10-year cardiovascular risk of a person. Although not perfect, baldness would be a very strong proxy of age and sex, whereas shoe size would be a proxy of sex and height. Age, sex, and height are all variables that causally influence the probability of a cardiovascular event, and the variation of these variables in relation to the risk of an event would be captured by baldness and shoe size quite well. If a hypothetical patient attends their NHS Health Check and is given a high score by this alternative model, it would be quite unusual for the GP to suggest that the patient is at risk of a heart attack *because they are bald*. The two variables in the model might manage to sufficiently capture the other unobserved variables and might even estimate a similar risk, but they cannot be interpreted as the *causes* of increased cardiovascular risk, which is what the question *why* requires.

Similarly, interpreting the individual variables as causes in the QRISK might seem intuitive if they are known to have some causal link to the outcome, but this would be equally wrong. Prediction models are concerned with maximizing the available information from the training data and using a set of covariates that *jointly* predict the outcome with the highest accuracy. The weight of each feature thus completely depends both on the combination of variables *in* the model *and* the combination of variables *not in* the model. The act of inappropriately interpreting the coefficients of all variables in a model and subsequently inferring any causal explanations between the variable(s) and the outcome has been termed 'The Table 2 Fallacy'[138], and although widely recognised in some areas, it is not appreciated enough in the context of prediction modelling.

The QRISK algorithm therefore managed to 'pass' the transparency and explainability criteria, but failed the interpretability criteria by virtue of being a predictive model as opposed to a model built using a causal inference framework (to be able to answer questions such as *why*). In the case of black box models, we cannot even hope to make them interpretable when they are not even sufficiently transparent and explainable to begin with. After all, we have just demonstrated that a fully transparent and explainable model is completely uninterpretable. Even if we can use sophisticated methods to extract the relative feature importance, or even the specific feature weights, this would not, unfortunately, be useful at all, because prediction models are fundamentally unsuited for seeking *interpretation* and asking *why* the model arrives at a decision (i.e. how these variables actually impact outcomes in the real world) as opposed to simply how (i.e. what are the intrinsic model mechanisms leading to a specific output).

Ethical Considerations

To illustrate the actual dangers of conflating the concepts of transparency, explainability and interpretability, and see why such conflation might be considered unethical, we can explore a recent *Nature* article that looked at factors associated with Covid-19-related death using data from OpenSAFELY[139]. This study does not feature fancy 'black box' algorithms but instead uses linear regression, which is completely transparent as the coefficients of all model variables are estimated explicitly. The authors presented all coefficients and took extra care to explain that these do not correspond to any causal effects; instead, they discussed the extent to which a certain patient characteristic is a 'risk factor' for Covid-19-related death.

Among the results were some unexpected findings even if we try not to make causal explanations and stick to strictly discussing 'risk factors' (a vague term that often implies causal association even if explicitly stated otherwise). The results would appear to suggest that smoking and hypertension are protective, which is implausible physiologically. The authors state that smoking status only appears 'protective' in the model that was adjusted for all other 'risk factors', whereas in post-hoc analyses they find that the change in the direction of association is dependent on whether adjustment is made for chronic respiratory disease. The reason why this occurs is the phenomenon termed 'collider bias', which arises if a variable on the causal path between the exposure and the outcome (i.e. a mediator) is conditioned upon, either inadvertently or deliberately[140]. In the *Nature* study, chronic respiratory disease can be conceptualized as a mediator because it is often caused by smoking, and it causes (i.e. affects the risk of) Covid-19-related death. Adjusting for chronic respiratory disease thus invokes spurious associations leading to potential sign reversal, as seen in this study. Collider bias is a serious potential source of bias in many studies that analyse observational Covid-19 data[141].

The study had some unexpected findings, but the authors cautioned against making any causal inferences. However, this does not 'fix' the problem. This illustration demonstrates how the human desire to give meaning to the model coefficients leads to 'The Table 2 Fallacy', but the implications of this are not limited to merely 'unexpected' findings. After the results from the study were published, it was announced that the French government had used the hazard ratios reported in the study to inform their decisions on assigning risk of Covid-19-related death to various categories of workers, which led to inappropriately withdrawing protection from some workers at severe risk of death.[142] This way, coefficients from a study that were not meant to be interpreted, were not only interpreted but actually used in the real world during a pandemic, with potentially serious consequences.

Many models are built with the implicit desire for causal explanation and interpretation, yet they are not explicitly designed to address such queries and thereby fail to provide any robust foundation to solve ethical queries. This demonstrates the ethical implications that results from failing to define properly the concepts of transparency, explainability, *and interpretability*. It is imperative to recognise these terms as distinct, needing both technical and theoretical considerations. Just because algorithmic transparency is achieved, it does not mean that model coefficients are interpretable and/or useful, raising the question about the appropriateness of transparency and/or explainability alone, i.e. devoid of any causal interpretability. Furthermore, making all 'black box' AIs as transparent as a regression model, could (perversely) lead to many more inappropriate interpretations, with potentially unknown ethical implications. How to make models actually interpretable?

Interpretability requires a causal model and a causal understanding that, in turn, requires external theory to frame it. No AI is yet able to infer causal understanding and neither the model or algorithmic weights nor the relative importance of different features in a model represent any meaningful assessment of causal effects[135].

To ensure that a model is interpretable, we should build it using a causal framework. We must not forget that the *only* variable that we can interpret in a model is the exposure, i.e. the coefficient of the exposure that represents its estimated causal effect on the outcome, conditional on all other variables in the model (i.e. the adjustment set) being held constant.

7.1 References

1 Pearl J, Mackenzie D. The book of why: the new science of cause and effect. Basic Books, 2018.

2 Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health* 2018; **108**: 616–9.

3 Pearl J, Glymour MM, Jewell NP. Causal Inference in Statistics: A Primer. London: Wiley, 2016.

4 The Oxford Handbook of Causation. 2019 https://books.google.com/books/about/The_Oxford_Handbook_of_Causation.html (accessed June 12, 2019).

5 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: A classification of data science tasks. *Chance* 2019; **32**: 42–9.

6 Arnold KF, Davies V, de Kamps M, Tennant PW, Mbotwa J, Gilthorpe MS. Reflections on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology* 2020.

7 Stacey T, Tennant P, McCowan L, *et al.* Gestational diabetes and the risk of late stillbirth: a case–control study from England, UK. *BJOG: An International Journal of Obstetrics & Gynaecology* 2019; **126**: 973–82.

8 Browne MW. Cross-validation methods. *Journal of mathematical psychology* 2000; **44**: 108–32.

9 Wynants L, Van Calster B, Collins GS, *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj* 2020; **369**.

10 Huitfeldt A. Is caviar a risk factor for being a millionaire? *BMJ* 2016; **355**: i6536.

11 Weed DL. Commentary: Causal inference in epidemiology: potential outcomes, pluralism and peer review. *Int J Epidemiol* 2016; **45**: 1838–40.

12 Hernan MA, Robins JM. Causal inference. CRC Boca Raton, FL, 2019 https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2019/02/hernanrobins_v1.10.38.pdf.

13 Greenland S, Pearl J, Robins JM, others. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37–48.

14 Hoggart CJ, Parra EJ, Shriver MD, *et al.* Control of confounding of genetic associations in stratified populations. *AmJHumGenet* 2003; **72**: 1492–504.

15 Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–25.

16 Rothman KJ. The estimation of synergy or antagonism. *AmJEpidemiol* 1976; **103**: 506–11.

17 Tennant PWG, Murray EJ, Arnold KF, *et al.* Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology* 2021; **50**: 620–32.

18 Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *IntJEpidemiol* 1986; **15**: 413–9.

19 McNamee R. Confounding and confounders. *OccupEnvironMed* 2003; **60**: 227–34.

20 Greenland S, Morgenstern H. Confounding in health research. *AnnuRevPublic Health* 2001; **22**: 189–212.

21 Jewell NP. Statistics for Epidemiology. London: Chapman & Hall, 2004.

22 Pearl J. Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press, 2009.

23 Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *IntJ Epidemiol* 2016; **45**: 1887–94.

24 Belsky DW, Caspi A, Houts R, *et al.* Quantification of biological aging in young adults. *ProcNatlAcadSciUSA* 2015; **112**: E4104–10.

25 Huxley RR, Neil A, Collins R. Unravelling the fetal origins hypothesis: is there really an inverse association between birthweight and subsequent blood pressure? *Lancet* 2002; **360**: 659–65.

26 Tu YK, West R, Ellison GT, Gilthorpe MS. Why evidence for the fetal origins of adult disease might be a statistical artifact: the ‘reversal paradox’ for the relation between birth weight and blood pressure in later life. *AmJEpidemiol* 2005; **161**: 27–32.

27 Tu YK, Gilthorpe MS, Ellison GT. What is the effect of adjusting for more than one measure of current body size on the relation between birthweight and blood pressure? *JHumHypertens* 2006; **20**: 646–57.

28 Weinberg CR. Invited commentary: Barker meets Simpson. *AmJEpidemiol* 2005; **161**: 33–5.

29 Tu YK, Ellison GT, West R, Gilthorpe MS. Tu et Al. Respond to ‘barker meets simpson’. *AmJEpidemiol* 2005; **161**: 36–7.

30 Lippa RA. Gender, Nature, and Nurture. New York: Taylore & Francis, 2005.

31 Halpern DF. Sex Differences in Cognitive Abilities. New York: Taylor and Francis, 2012.

32 Pearl J. Do Calculus Revisited. In: Freitas N de, Murphy K, eds. . Corvallis, OR: AUAI Press, 2012: 4–11.

33 Pearl J. Interpretation and identification of causal mediation. *PsycholMethods* 2014; **19**: 459–81.

34 Pearl J, Bareinboim E. External validity: From *do*-calculus to transportability across populations. *Statistical Science* 2014; **29**: 579–95.

35 Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *AmJEpidemiol* 2013; **177**: 292–8.

36 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* 2016; **183**: 758–64.

37 Rohman JL. Tweet: 'If we can't think of a target trial, we probably do not have a causal question'. https://twitter.com/JLRohmann/status/1014067407424499712.

38 García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol* 2017; **32**: 495–500.

39 Yee A, Majumdar SR, Simpson SH, McAlister FA, Tsuyuki RT, Johnson JA. Statin use in Type 2 diabetes mellitus is associated with a delay in starting insulin. *Diabetic Medicine* 2004; **21**: 962–7.

40 Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010; **340**: b5087.

41 Voysey M, Clemens SAC, Madhi SA, *et al.* Single-dose administration and the influence of the timing of the booster dose on immunogenicity and efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine: a pooled analysis of four randomised trials. *The Lancet* 2021; **397**: 881–91.

42 Zhou Z, Rahme E, Abrahamowicz M, Pilote L. Survival Bias Associated with Time-to-Treatment Initiation in Drug Effectiveness Evaluation: A Comparison of Methods. *American Journal of Epidemiology* 2005; **162**: 1016–23.

43 Hernán MA. How to estimate the effect of treatment duration on survival outcomes using observational data. *BMJ* 2018; **360**: k182.

44 Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Statistics in medicine* 2013; **32**: 4118–34.

45 Boulesteix A-L, Groenwold RH, Abrahamowicz M, *et al.* Introduction to statistical simulations in health research. *BMJ open* 2020; **10**: e039921.

46 Tarka P. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & quantity* 2018; **52**: 313–54.

47 Pearl J, Verma TS. A theory of inferred causation. In: Studies in Logic and the Foundations of Mathematics. Elsevier, 1995: 789–811.

48 Wright S. Correlation and causation. 1921.

49 Wright S. The method of path coefficients. *The annals of mathematical statistics* 1934; **5**: 161–215.

50 Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics* 2015; **71**: 1–14.

51 VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Press, 2015.

52 Pearl J, Verma TS. A theory of inferred causation. In: Studies in Logic and the Foundations of Mathematics. Elsevier, 1995: 789–811.

53 Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; **82**: 669–88.

54 Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology* 2016; **45**: 1887–94.

55 Ruscio J, Kaczetow W. Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research* 2008; **43**: 355–81.

56 Myers JA, Rassen JA, Gagne JJ, *et al.* Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology* 2011; **174**: 1213–22.

57 Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology* 2003; **158**: 280–7.

58 Tu C. Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation* 2019; **89**: 708–19.

59 Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass)* 2009; **20**: 512.

60 Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International journal of epidemiology* 2014; **43**: 1969–85.

61 Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 2011; **46**: 399–424.

62 Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight 'paradox' uncovered? *AmJEpidemiol* 2006; **164**: 1115–20.

63 Wilcox AJ. Invited commentary: the perils of birth weight--a lesson from directed acyclic graphs. *AmJEpidemiol* 2006; **164**: 1121–3.

64 Cole SR, Platt RW, Schisterman EF, *et al.* Illustrating bias due to conditioning on a collider. *IntJEpidemiol* 2010; **39**: 417–20.

65 Pearl J. Lord's Paradox Revisited (Oh Lord! Kumbaya!). http://ftp.cs.ucla.edu/pub/stat_ser/r436.pdf, 2014.

66 Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American journal of epidemiology* 2016; **184**: 847–55.

67 Keeble C, Baxter P, Barber S, Law G, others. Participation rates In epidemiology studies and surveys: a review 2005-2007. *The Internet Journal of Epidemiology* 2016; **14**: 1–14.

68 Biobank U. Access matter: representativeness of the UK Biobank resource. *Stockport, UK: UK Biobank* 2017.

69 Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *IntJEpidemiol* 2018; **47**: 226–35.

70 Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. *Acta obstetricia et gynecologica Scandinavica* 2018; **97**: 407–16.

71 Thompson CA, Arah OA. Selection bias modeling using observed data augmented with imputed record-level probabilities. *Annals of epidemiology* 2014; **24**: 747–53.

72 Galton F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* 1886; **15**: 246–63.

73 Centripetal Drift: A Fallacy in the Evaluation of Therapeutic Results. https://www.science.org/doi/abs/10.1126/science.87.2264.461 (accessed Sept 5, 2021).

74 Gadd SC, Tennant PW, Heppenstall AJ, Boehnke JR, Gilthorpe MS. Analysing trajectories of a longitudinal exposure: A causal perspective on common methods in lifecourse research. *PloS one* 2019; **14**: e0225217.

75 Berrie L, Ellison GT, Norman PD, *et al.* The association between childhood leukemia and population mixing: an artifact of focusing on clusters? *Epidemiology (Cambridge, Mass)* 2019; **30**: 75.

76 Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *AmJEpidemiol* 2005; **162**: 267–78.

77 Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. New York: Lippincott Williams & Wilkins, 2008.

78 Nelson BK. Time series analysis using autoregressive integrated moving average (ARIMA) models. *Academic emergency medicine* 1998; **5**: 739–44.

79 Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annual review of public health* 2017; **38**: 39–56.

80 Baiocchia M, Chengb J, Smallc DS. Tutorial in Biostatistics: Instrumental Variable Methods for Causal Inference. .

81 Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American journal of clinical nutrition* 2016; **103**: 965–78.

82 Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama* 2014; **312**: 2401–2.

83 Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology* 2017; **46**: 348–55.

84 Bor J, Moscoe E, Mutevedzi P, Newell M-L, Bärnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology (Cambridge, Mass)* 2014; **25**: 729.

85 Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 1986; **51**: 1173.

86 VanderWeele TJ. A unification of mediation and interaction: a four-way decomposition. *Epidemiology (Cambridge, Mass)* 2014; **25**: 749.

87 VanderWeele TJ. Mediation analysis: a practitioner's guide. *Annual review of public health* 2016; **37**: 17–32.

88 Spirtes P, Glymour CN, Scheines R, *et al.* Causation, prediction, and search. MIT press, 2000.

89 Berrie L, Tennant P, Norman P, Baxter P, Gilthorpe M. Mathematical coupling and causal inference through example. *Journal of Epidemiology and Community Health* 2018.

90 Geiger D, Verma T, Pearl J. Identifying independence in bayesian networks. *Networks* 1990; **20**: 507–34.

91 Berrie L, Tennant P, Norman P, Baxter P, Gilthorpe M. Mathematical coupling and causal inference through example. *Journal of Epidemiology and Community Health* 2018.

92 Razak F, Subramanian S, Sarma S, *et al.* Association between population mean and distribution of deviance in demographic surveys from 65 countries: cross sectional study. *BMJ* 2018; **362**: k3147.

93 Andersen B. Methodological errors in medical research. London: Blackwell, 1990.

94 Moreno LF, Stratton HH, Newell JC, Feustel PJ. Mathematical coupling of data: correction of a common error for linear calculations. *Journal of Applied Physiology* 1986; **60**: 335–43.

95 Stratton HH, Feustel PJ, Newell JC. Regression of calculated variables in the presence of shared measurement error. *Journal of Applied Physiology* 1987; **62**: 2083–93.

96 Tu YK, Gilthorpe MS, Griffiths GS. Is reduction of pocket probing depth correlated with the baseline value or is it 'mathematical coupling'? *JDentRes* 2002; **81**: 722–6.

97 Tu YK, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. *StatMed* 2007; **26**: 443–57.

98 Pearson K. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* 1897; **60**: 489–98.

99 Pearson K, Lee A, Elderton EM. On the correlation of death-rates. *Journal of the Royal Statistical Society* 1910; **73**: 534–9.

100 Fisher RA. The analysis of covariance method for the relation between a part and the whole. *Biometrics* 1947; **3**: 65–8.

101 Neyman J. Human cancer: radiation and chemicals compete. *Science* 1979; **205**: 259–60.

102 Oldham PD. A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases* 1962; **15**: 969–77.

103 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–10.

104 Morgan WA. A test for the significance of the differences between two variances in a sample from a normal bivariate distribution. *Biometrika* 1939; **31**: 13–9.

105 Archie JP. Mathematical Coupling: A common source of error. *Annals of Surgery* 1981; **193**: 296–303.

106 Chiolero A, Paradis GP, Rich BD, Hanley JP. Assessing the Relationship between the Baseline Value of a Continuous Variable and Subsequent Change Over Time. *Front Public Health* 2013; **1**. DOI:10.3389/fpubh.2013.00029.

107 Sperrin M, Candlish J, Badrick E, Renehan A, Buchan I. Collider Bias Is Only a Partial Explanation for the Obesity Paradox. *Epidemiology* 2016; **27**: 525–30.

108 Blair A, Hollenbeck A, Schatzkin A, *et al.* Amount of time spent in sedentary behaviors and cause-specific mortality in US adults. *The American Journal of Clinical Nutrition* 2012; **95**: 437–45.

109 Matthews CE, Chen KY, Freedson PS, *et al.* Amount of time spent in sedentary behaviors in the United States, 2003-2004. *AmJEpidemiol* 2008; **167**: 875–81.

110 Owen N, Healy GN, Matthews CE, Dunstan DW. Too much sitting: the population health science of sedentary behavior. *ExercSport SciRev* 2010; **38**: 105–13.

111 McAfee AJ, McSorley EM, Cuskelly GJ, *et al.* Red meat consumption: an overview of the risks and benefits. *MeatSci* 2010; **84**: 1–13.

112 Bronzato S, Durante A. A Contemporary Review of the Relationship between Red Meat Consumption and Cardiovascular Risk. *IntJPrevMed* 2017; **8**: 40-.

113 Tomova GD, Arnold KF, Gilthorpe MS, Tennant PW. Adjustment for energy intake in nutritional research: a causal inference perspective. Epidemiology, 2021 DOI:10.1101/2021.01.20.21250156.

114 Arnold KF, Berrie L, Tennant PW, Gilthorpe MS. A causal inference perspective on the analysis of compositional data. *International journal of epidemiology* in press.

115 Berrie L, Tennant P, Norman P, Baxter P, Gilthorpe M. Mathematical coupling and causal inference through example. *Journal of Epidemiology and Community Health* 2018.

116 Shachter RD. Probabilistic inference and influence diagrams. *Operations research* 1988; **36**: 589–604.

117 Geiger D, Verma T, Pearl J. Identifying independence in Bayesian networks. *Networks* 1990; **20**: 507–34.

118 Shahar E. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American journal of epidemiology* 2009; **170**: 957–8.

119 Shahar E. Causal diagrams for encoding and evaluation of information bias. *Journal of evaluation in clinical practice* 2009; **15**: 436–40.

120 Cole SR, Frangakis CE. The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology* 2009; **20**: 3–5.

121 Hernán MA. Does water kill? A call for less casual causal inferences. *Annals of epidemiology* 2016; **26**: 674–80.

122 Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International journal of obesity* 2008; **32**: S8.

123 Rajpurkar P, Irvin J, Ball RL, *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018; **15**: e1002686.

124 Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016; **6**: 26094.

125 Díez Díaz F, Sánchez Lasheras F, Moreno V, Moratalla-Navarro F, Molina de la Torre AJ, Martín Sánchez V. GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. *Mathematics* 2021; **9**: 654.

126 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017; : j2099.

127 Wishart GC, Azzato EM, Greenberg DC, *et al.* PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010; **12**: R1.

128 Diakopoulos N, Koliska M. Algorithmic Transparency in the News Media. *Digital Journalism* 2017; **5**: 809–28.

129 Rai A. Explainable AI: from black box to glass box. *J of the Acad Mark Sci* 2020; **48**: 137–41.

130 Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. In: Precup D, Teh YW, eds. Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017: 3145–53.

131 Alaa AM, van der Schaar M. Demystifying Black-box Models with Symbolic Metamodels. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F d\textquotesingle, Fox E, Garnett R, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper/2019/file/567b8f5f423af15818a068235807edc0-Paper.pdf.

132 Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 2016: 1135–44.

133 Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—Explainable artificial intelligence. *Sci Robot* 2019; **4**: eaay7120.

134 Poon AIF, Sung JJY. Opening the black box of AI-Medicine. *Journal of Gastroenterology and Hepatology* 2021; **36**: 581–4.

135 Arnold KF, Davies V, de Kamps M, Tennant PW, Mbotwa J, Gilthorpe MS. Reflections on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology* 2020.

136 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: A classification of data science tasks. *Chance* 2019; **32**: 42–9.

137 Shmueli G. To Explain or to Predict? *Statist Sci* 2010; **25**. DOI:10.1214/10-STS330.

138 Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *AmJEpidemiol* 2013; **177**: 292–8.

139 Williamson EJ, Walker AJ, Bhaskaran K, *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; **584**: 430–6.

140 Cole SR, Platt RW, Schisterman EF, *et al.* Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* 2010; **39**: 417–20.

141 Griffith GJ, Morris TT, Tudball MJ, *et al.* Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020; **11**: 5749.

142 Suspension des nouveaux critères de vulnérabilité au covid-19 ouvrant droit au chômage partiel. Affiches Parisiennes. 2020; published online Oct 19. https://www.affiches-parisiennes.com/suspension-des-nouveaux-criteres-de-vulnerabilite-au-covid-19-ouvrant-droit-au-chomage-partiel-11123.html.