Causal inference in the CLS cohort studies

User Guide (Version 2)

October 2026

UCL Centre for Longitudinal Studies

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Liam Wright, Georgia D. Tomova, Martin N. Danka, George B. Ploubidis, Richard J. Silverwood.

How to cite this guide

Wright, L., Tomova, G.D., Danka, M.N., Ploubidis, G.B., Silverwood, R.J. (2026) *Causal inference in the CLS cohort studies User Guide (Version 2)*. London: UCL Centre for Longitudinal Studies.

Data citation and acknowledgement

You should cite the data and acknowledge CLS following the guidance from Citing our data page on our website.

About UCL Centre for Longitudinal Studies

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre. It is home to a unique series of UK national cohort studies. It is part of the UCL Social Research Institute, based at the UCL Institute of Education.

UCL Centre for Longitudinal Studies is core funded by:

Economic and Social Research Council

This document is available in alternative formats. Please email the Centre for Longitudinal Studies at clsdata@ucl.ac.uk.

Contents

Abstract **6**

1. Introduction **7**

2. The problem of causal inference with observational data **10**

2.1 Correlation ≠ Causation: What next? 10

2.2 Threats to (internal) validity 12

2.3 Why cohort studies? 35

3. Three approaches to causal inference **36**

3.1 Model-based approaches 36

3.2 Design-based approaches 37

3.3 Elaborate theories 39

4. Attributes of CLS's cohort studies that support causal inference **43**

4.1 Breadth and Depth 43

4.2 Extensive Data Linkages 44

4.3 Longitudinal Assessment 44

4.4 Extended Timespan 45

4.5 Large Sample Sizes 45

4.6 Ability to Harmonise Data 46

5. The methods of causal inference **48**

5.1 Control via Regression Adjustment 50

5.2 Matching 57

5.3 Inverse Probability of Treatment Weighting 62

5.4 Missing Data Methods 65

5.5 (Longitudinal) Fixed Effects 71

5.6 Difference in Differences 76

5.7 Study-Initiated Experiments 84

5.8 Natural Experiments 89

5.9 Instrumental Variables 95

5.10 Regression Discontinuity Designs 105

5.11 Target Trial Emulation 113

5.12 G-methods for Time-Varying Exposures 118

5.13 Causal Mediation 127

5.14 Family Designs 142

5.15 Negative Control Outcomes and Exposures 147

5.16 Cross-Context Analysis 154

5.17 Biases of Known Direction 159

5.18 Differential Effects 162

5.19 Reasons for Effects 165

5.20 Quantitative Bias Analysis 168

6. Discussion 181

7. Statements 184

7.1 Declaration of interest 184

7.2 Funding 184

7.3 Acknowledgements 184

7.4 Author contributions 184

7.5 Data availability 184

8. Appendix 185

8.1 Correlation ≠ Causation: A Short Empirical Review 185

8.2 Why Does Correlation (Often) Not Equal Causation? 188

9. References 191

Abstract

Correlations between variables in cohort and other observational studies do not necessarily indicate causal effects and may instead reflect bias due to confounding, sample selection processes, or measurement issues. Thankfully, multiple statistical methods, frameworks, and research designs have been developed that – when used judiciously – can aid making causal inferences with observational data. This is especially the case with CLS's cohorts, which have several properties (e.g., longitudinal assessment and rich and varied data collection) that that can make the assumptions of these methods more plausible or provide ample opportunities for each method's use. In this guide, we provide an overview of the possibilities for causal inference in CLS's cohorts. Specifically, we: (1) provide background on the problem of causal inference in observational data, (2) detail the attributes of CLS's cohorts that can support causal inference, and (3) outline an extensive set of methods and research designs that – assumptions permitting – can be used to estimate causal effects in CLS's cohort data, illustrating each with existing applications and highlighting opportunities for further work.

Keywords: causal inference; observational studies; elaborate theories; cohort studies; quasi-experiments; research design; quasi-experimental devices

1. Introduction

The Centre for Longitudinal Studies (CLS) manages several world-leading cohort studies, including the 1958 National Child Development Study (NCDS), the 1970 British Cohort Study (BCS70), Next Steps (formerly known as the Longitudinal Study of Young People in England), and the Millennium Cohort Study (MCS). These studies have offered researchers a unique window into social change in Britain over the past 68 years (Ferri et al., 2003; Pearson, 2016) – for instance, in the realms of physical and mental health, education, social mobility, and the labour market (Anders & Dorsett, 2017; Erikson & Goldthorpe, 2010; Gondek, Bann, Patalay, et al., 2021; Johnson et al., 2015). Beyond documenting these trends, the astonishing richness of the cohorts – each contains several thousands of variables from many thousands of participants – has enabled researchers to answer important causal questions across a range of research domains – from the effects of breastfeeding on child development (Fitzsimons & Vera-Hernández, 2022) to the impact of laws promoting flexible working practices on workers' wellbeing (Avendano & Panico, 2018).

Making plausible causal inferences with observational data is not straightforward. Correlations between study variables may not reflect causal effects due to the variables sharing common causes (i.e., confounding), biases in the sample of individuals available for analysis (i.e., selection bias), and difficulties obtaining accurate measurements (i.e., information bias). As reviewed in the Appendix, there are many examples in which findings from observational studies have failed to be confirmed in randomised controlled trials (RCTs).

Nevertheless, many successes have been achieved; most famously, the discovery of tobacco smoking's deleterious impact on health, much of the evidence for which came from observational studies (Cornfield et al., 1959). Decades of methodological development within econometrics, epidemiology, statistics and computer science has formalised the conditions under which causal inferences can be made with observational data and identified several methods that can – when applied appropriately – help researchers make valid causal inferences. Further, increasingly adopted frameworks and cognitive tooling – notably, the Potential Outcomes

Framework and Causal Directed Acyclic Graphs (DAGs) – have made it far easier to understand and communicate sources of bias in observational studies and to determine appropriate analyses for estimating causal effects with the data at hand.

The aim of this document is to review efforts towards making causal inferences using CLS's cohorts and to highlight what, to us, appear to be promising opportunities for causal inference that warrant further exploration with our data. To this end, we provide background on the problem of causal inference using cohort data (Section 2), outline three overarching philosophies for making causal inferences (Section 3), describe the features of CLS's data that can (circumstances permitting) enable causal effect estimation (Section 4), and review specific methods especially suited to our data, including describing existing applications of these methods and discussing further opportunities for the methods' use (Section 5). For each method, we provide key references for readers to further explore. We close by arguing that evidence from cohort studies should be considered within an overall strategy of 'triangulation' (Section 6), combining results from multiple, complementary research designs and taking 'varied views of a single object' (Rosenbaum, 2023) – submitting a claim's assorted premises and predictions to empirical testing, in order to increase the robustness of causal inferences.

How to read this guide

This guide assumes familiarity with regression. Otherwise, it is relatively introductory: the focus is on intuition – equations are used only where necessary or enlightening and mathematical complexity is kept to a minimum. Sections are relatively standalone and include clickable cross-references to other sections where necessary. Readers new to causal inference should begin with Section 2.

The subsections in Section 5 can be read selectively. Sections 5.12-5.13 address causal effect estimation for time-varying exposures and for mediation. Sections 5.15-5.20 discuss sensitivity to bias and ways to investigate alternative explanations for causal effects. These sections are best read alongside Section 3.3 on elaborate theories. For readers with a background in causal inference, Section 3.3 and Sections 5.15-5.20 may be most novel as they describe an arguably underappreciated perspective on the task of causal inference.

Note this is a work in progress, which we are releasing now, to enable users to benefit. Any feedback – much appreciated – can be sent to the CLS communications team (ioe.clscomms@ucl.ac.uk) or raised as an issue on the Causal Inference Review GitHub page.

2. The problem of causal inference with observational data

In this section, we provide background on the problem of making causal inferences using cohort and other observational data. We begin by noting that causal questions are central to much quantitative research, regardless of the truth of mantras such as 'correlation does not equal causation'. We then provide a theoretical overview of the three main causes of bias (threats to internal validity) in observational studies: confounding bias, selection bias, and information bias. The latter two of these typically receive less attention among applied researchers, though they are no less capable of leading to incorrect inferences. We use the Potential Outcomes Framework and Causal DAGs as tools for describing these biases and conclude with a discussion of the necessity of observational data for scientific progress, notwithstanding potential problems of bias.

For interested readers, in the Appendix (Section 8), we demonstrate the problem of making causal inferences using cohort and other observational data empirically, reviewing literature that compares results from observational studies and RCTs and shows discrepancies between the two – in many cases because of important biases present in observational studies (though not always, it must be noted). In the Appendix, we also provide general theoretical perspectives on why correlation may not equal causation in observational studies; though the precise causes of bias may be analysis-specific, there are reasons to expect widespread confounding in practice and broad classes of confounders that are likely to be important in varied situations.

2.1 Correlation ≠ Causation: What next?

At root, many of the questions in the health and social sciences are causal in nature. For instance, what impact do recessions have on lifetime wages? What effect does parental university education have on the likelihood of attending university oneself? What consequence does low birthweight have for life expectancy? Repeating 'correlation does not equal causation' does not get us very far, either as scientists

interested in learning about the world or as citizens hoping to put policy on an evidenced basis. Euphemisms do not help, either – neutral-sounding terms like "association", "links", and "risk factors" do not satisfy the need to answer causal questions and can instead divert attention from the causal thinking that is necessary (Hernán, 2018b).

The first step in any research programme, then, is to recognise one's aims and which of the three "tasks" of data science – description, prediction and counterfactual prediction (i.e., causal inference; Hernán et al., 2019) – these aims fit into. Unlike descriptive studies, which are broadly concerned with quantifying and characterising features (e.g., outcomes or events) in a population of interest, or prediction studies, which aim to classify cases or forecast future events from patterns in the data, causal inference is instead concerned with asking how one variable influences another or what would happen if we were to intervene on the world (e.g., how would a person's education change if they were adopted by highly educated, rather than less educated, parents; Sacerdote, 2002).

After recognising one's aims, the second step is crafting a good research question; research can fail prematurely, even with technically proficient analyses, if the research question is not well crafted. A good causal research question requires thinking carefully about, and defining, its key features, namely: the exposure ('treatment') and comparator, the outcome, the population, and – of course – the particular effect of interest. These should be clear and specific enough that the question can be translated into a statistical 'estimand' (Lundberg et al., 2021) – i.e., a quantity one wishes to estimate, representing a key parameter in a causal system (e.g., the difference in wages at age 33y if someone born in the UK entered adulthood during a recession compared with a healthy economy). One way of doing this is to specify the research design as a 'target trial': one benefit of RCTs is that they encourage the study designer to specify well-defined causal questions by requiring explicit inclusion and exclusion criteria, a specific exposure (and comparator), and so on (Hernán et al., 2025). This can be an iterative process, requiring an awareness of what is feasible with available data. Target trials will be discussed further in Section 5.11.

To give an example: suppose you are interested in studying the effect of planned Caesarean section on the risk of asthma in childhood. The exposure and comparator would be two levels of a binary variable (1 = Caesarean section, 0 = vaginal delivery). The outcome would need to be defined at a specific time point or in a time window, decided based on subject matter expertise and data availability. For instance, it could be asthma diagnosis between ages 5 and 12 years. The population would be the set of individuals one wishes to make inferences to, determined by the intersection of research interest and data availability: e.g., it may be children in the UK born between 2000 and 2002, the target population of the Millennium Cohort Study. Finally, one would define the specific effect ('estimand') of interest – e.g., the population average causal ('treatment') effect. Taken together then, the developed causal question might read: "What is the average treatment effect of being born by planned Caesarean section instead of vaginally on the risk of asthma diagnosis at age 5-12 years, among children in the UK born between 2000 and 2002?"

Carrying out this process allows us to more easily translate the question into a well-defined causal estimand, to design an appropriate study, to determine the most appropriate analytical strategy required to answer the question (Lundberg et al., 2021), and – crucially – to interpret the correct interpretation of the resulting estimate and the population it relates to. Naturally, there exist many different procedures for specifying a good causal question; this section is only intended to serve as a starting point for consideration. For instance, for a continuous exposure, it is advisable to define specific levels of the exposure to compare (e.g., "60 vs 30 minutes of vigorous physical activity per day"), and, additionally, one might consider including specific effect modifiers of interest. See Section 5.11 on Target Trial Emulation and Section 2.2.1, which introduces treatment variation irrelevance, for further discussion of elements worth considering when specifying a suitable causal question.

2.2 Threats to (internal) validity

Though observational studies have generated multiple cases where associations have not produced reliable estimates of causal effects (see Section 8.1, Appendix, for a narrative review), there are numerous examples of observational studies that have. Why

is this the case? The ability to estimate an unbiased causal effect in a study sample is referred to as 'internal validity'. Several disciplines have converged upon three major threats to internal validity that can generate bias in (observational) data, sometimes using their own terms or the same terms in different ways. In this section, we introduce these threats using two frameworks: the Potential Outcomes Framework and Causal DAGs.

2.2.1 What is a causal effect? Counterfactuals and the potential outcomes framework

We first need to define a causal effect. This is a subject of substantial philosophical discussion (Mumford & Anjum, 2013), but in practical terms, statisticians, social scientists and epidemiologists have adopted the counterfactual definition: for an individual, an exposure X has a causal effect on an outcome Y if that individual's Y would be different under different levels of X. More formally, for an exposure with two levels:

$$Y_i^{X=1} \neq Y_i^{X=0} \text{ for some individual } i$$

where $Y_i^{X=x}$ is the level of *Y* for individual *i* had they received *X* = *x*.

We will refer to a person as treated (or exposed) if they in fact received the exposure and untreated (or unexposed) if they in fact did not. In practice, we only observe one state of the world and so we can only see one of these *potential outcomes* (i.e., the *factual* rather than *counterfactual* outcome) for any given individual. For treated individuals, this is their outcome having received the treatment ($Y_i^{X=1}$). For untreated individuals, this is their outcome having not received the treatment ($Y_i^{X=0}$).

Given we only observe one potential outcome for each individual, individual causal effects cannot be observed – this is the "fundamental problem of causal inference" (Holland, 1986). To estimate causal effects, we instead rely on population quantities – comparisons between groups – such as the difference in the mean of *Y* according to levels of *X*. A key causal parameter is the Average Treatment Effect (ATE) – i.e., the average causal effect across a population:

$$ATE = E(Y^{X=1}) - E(Y^{X=0})$$

That is, it is the difference in the average level of *Y* had everyone been treated vs. had everyone been untreated.

Estimating the ATE is not straightforward as we generally cannot just compare those who, in fact, received the treatment with those who, in fact, did not, as these potential outcomes may not be identical in treated and untreated groups – i.e., where $E(Y^{X=1}) \neq E(Y^{X=1} | X = 1)$ or $E(Y^{X=0}) \neq E(Y^{X=0} | X = 0)$. That is, the mean outcome that would be observed had the untreated group in fact received treatment may not be the same as the mean outcome that was observed among the treated group and/or the mean outcome that would be observed had the treated group in fact <u>not</u> received the treatment may not be the same as the mean outcome that was observed among the untreated group.

A simple comparison of means would recover the ATE if potential outcomes are *independent* of actual treatment assignment (i.e., where $E(Y^{X=1}) = E(Y^{X=1} | X = 1)$ and $E(Y^{X=0}) = E(Y^{X=0} | X = 0)$). The aim of causal inference then is to identify situations where independence (also known as *exchangeability* or *ignorability*) is met, which may include conditioning (achieving *conditional exchangeability*) – for instance, by statistical adjustment or stratifying according to factors related to potential outcomes (e.g., sex). However, (conditional) independence can only ever be an assumption because, ultimately, we do not observe all potential outcomes for any individual.

Beyond conditional exchangeability, the standard confounder-adjustment approach requires three further assumptions (together known as the '*identifiability conditions*') for population-average causal effects to be identified.[1] The first, *positivity*, states that, for the population of interest, there is a non-zero (conditional) probability of receiving each level of the treatment across all relevant confounder strata – otherwise, it would not be

[1] Some other methods, such as difference-in-differences, instrumental variables, and the regression discontinuity design, use different identifying arguments, though each includes a version of conditional exchangeability. These assumptions are stated in Section 5 when relevant. Note, these assumptions also apply when targeting other estimands, including the Average Treatment Effect on the Treated (ATT) and the Average Treatment Effect on the Untreated (ATU).

possible to estimate causal effects using (conditional) population averages as observations may not exist for all comparisons.[2] Another assumption is that of *no interference,* meaning the potential outcomes of one individual are unrelated to the treatment status of other individuals. This is most commonly at risk of violation when studying the effect of vaccines or behavioural interventions.

Finally, interventions must be *consistent* with respect to an outcome – that is, they must be sufficiently well-defined so that for each individual there is one potential outcome corresponding to each exposure level, i.e., $Y_i = Y_i^{X=1}$ when X = 1 (Hernán & Robins, 2023). This may seem self-evidently true, but it may be violated when there are multiple versions of treatment (for this reason consistency is also known as *treatment variation irrelevance*). An example of a question for which consistency fails is "does obesity shorten life?" (Hernán, 2016) – since there are multiple 'versions' of attaining a level of obesity, each with different effects on longevity, there is no single potential outcome corresponding to levels of the intervention. This type of intervention is sometimes referred to as a "fat-handed intervention" (Bailey et al., 2024) that is generally only achieved as part of an envelope of actions that may have independent effects on outcomes. $Y_i^{X=1}$ could mean "obesity due to quitting smoking" or "obesity due to calorific restriction", both of which have different implications for life expectancy and thus potential outcomes. The narrower question "does 10% weight loss via calorific restriction between ages 40 and 50 years affect risk of death?" is more likely to satisfy consistency (Hernán, 2016).[3]

---

[2] Some methods, such as the Sharp Regression Discontinuity Design (Section 5.10), do not rely upon the positivity assumption as they instead use statistical models or other assumptions to extrapolate counterfactual estimates.

[3] The consistency assumption can be highly restrictive given treatments can vary along many different dimensions (e.g., the angle at which a vaccine is administered; Hernán, 2016). In practice, it may not be possible to determine which aspects of different versions of a treatment are irrelevant for (potential) outcomes, and thus whether consistency is met. Domain-specific expertise is necessary to make these judgements, but is therefore limited by present knowledge (Hernán & Robins, 2023). Given it may be impossible to completely specify a consistent treatment, observed outcomes could be considered a weighted average of the different versions of treatment captured in

2.2.2 An introduction to causal DAGs

Besides the Potential Outcomes Framework, another useful development in the causal inference literature has been causal DAGs. DAGs are non-parametric visual representations of causal relationships that have proved exceptionally helpful for understanding and communicating biases in data and for helping researchers design studies that obtain valid estimates of causal effects. In this section, we give a brief introduction to DAGs as we use them extensively throughout this document to explain the intuition behind individual causal inference methods. This introduction is relatively self-contained, but for further background, see one of the many tutorials on DAGs that are now available (Digitale et al., 2022; Hernán & Robins, 2023; McElreath, 2020; Pearl, 2022; Pearl & Mackenzie, 2018; Williams et al., 2018). We particularly recommend Miguel Hernán's (2018a) free online course, *Causal Diagrams: Draw Your Assumptions Before Your Conclusions*; several of the DAGs shown here follow its examples.

In DAGs, variables are represented by nodes and causal relationships by directed arrows (arcs). A graph is a DAG if it fulfils two criteria:

1. It is directed – arcs between variables have directions that indicate the direction of causality.
2. It is acyclic – it is not possible to follow arrows forward and arrive at a variable already reached.

Figure 2.1: A simple DAG.

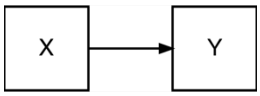


Figure 2.1 shows a very simple DAG containing two variables, X and Y. This DAG encodes three beliefs: (1) that X has a causal effect on Y, as indicated by the presence of an arrow going from X to Y; (2) that Y does not have a causal effect on X, as indicated

a definition ("10% weight loss via calorific restriction"). Bias can arise where there are (possibly) subtle differences in distribution of these treatments between treatment and control groups, even where conditional exchangeability has been achieved. Note treatment variation is one reason results from one population may not generalise to another (Hernán & Robins, 2023).

by the absence of an arrow going from Y to X (the graph is acyclic); and (3) that there are no variables which cause X *and* Y, as indicated by the absence of nodes with arrows going to X and Y – on the convention, discussed below, that DAGs are drawn 'complete'.[4]

A (frankly miraculous) property of DAGs is that beyond being conceptual models, DAGs obey a simple set of rules that entail empirical predictions: provided the DAG is complete – that is, the common causes of any two variables in the DAG also appear in the DAG – "open" paths (defined below) between variables permit statistical association, while their absence implies statistical independence. Implied independencies are testable and so some assumptions encoded in the DAG can be tested and refined by applying the DAG to real-world data.[5] The DAG in Figure 2.1 predicts that the variables X and Y should be correlated due to X having a causal effect on Y. Whether X and Y are correlated is testable, though a correlation could also arise because the DAG is mis-specified relative to the data it is applied to, either in principle (e.g., because Y causes X or X and Y share a common cause) or in practice (e.g., due to sample biases or measurement problems in the data).[6]

[4] Note it is possible to include in a DAG factors that have causal effects on one another. This is handled by treating variables as time-specific to respect temporality – e.g., $X_t \rightarrow Y_{t+1} \rightarrow X_{t+2}$.

[5] DAGs were initially developed by Judea Pearl, a computer scientist interested in artificial intelligence. As DAGs follow rules, a property is that they can be programmed and manipulated algorithmically. The software DAGitty (https://www.dagitty.net/) can be used to draw DAGs and can also generate the set of testable predictions (i.e., [conditional] independencies) implied by a DAG. Note, (conditional) independencies, which represent the **absence** of a path, represent stronger assumptions than dependencies, which could result from any number of paths (possibly of different directions, too).

[6] The distinction between principle and practice is not clear cut because, as we shall see, DAGs can be drawn to represent data issues and DAGs are most useful when drawn for the data they are applied to.

Figure 2.2: A more complex DAG.

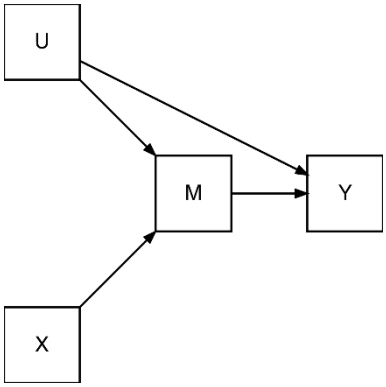


Figure 2.2 shows a more complex DAG. There are now four variables (X, U, M, and Y) and six directed paths representing causal effects, some of which pass through intermediate variables: X → M, X → M → Y, M → Y, U → M, U → M → Y, and U → Y. M has a causal effect on Y that is not mediated by any other variable in the DAG (M → Y),[7] X has a causal effect on Y that is fully mediated by M (X → M → Y), and U has a causal effect on Y that is only partly mediated by M (paths U → M → Y and U → Y). Each of these paths is said to be a 'causal' (or 'directed') path as each begins by following the arrows from the first variable onwards in the direction of causality.

Paths between variables do not necessarily proceed in the direction of causality, and association can travel along paths regardless of their direction. For instance, in Figure 2.2, M will also be associated with Y via a 'backdoor' path through U (M ← U → Y); such a path is said to be backdoor because it exits the back of the initial variable (i.e., it begins by following an arrow *against* the direction of causality). The path is non-causal in the sense that it does not consist solely of *descendants* (consequences) of the initial variable M, as indicated by the presence of arrows pointing in the opposite direction of causality (M ← U). This path therefore does not reflect a causal effect of M upon Y, but it nevertheless would lead to an observed association between M and Y that is introduced

[7] This is not to say the effect of M on Y is unmediated by any other variable (e.g., a biological mechanism operating at the cellular level). Recall, a DAG is complete if common causes of any two variables in the DAG also appear in the DAG. Mediators therefore only need to be included in a DAG if they cause, or are caused by, two variables in the DAG, though they may be optionally specified if of analytic interest.

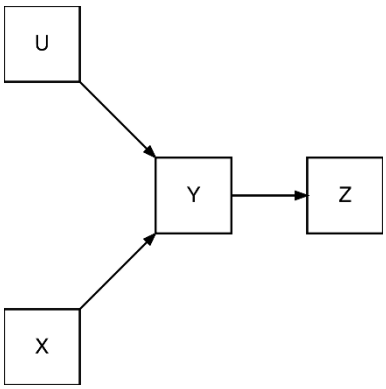
by their common cause (i.e., confounder), U. Controlling for this common cause would remove the association transmitted through the backdoor path.

More formally, in DAGs, paths between variables are either 'open' or 'closed' (*blocked*). Variables that are connected by an open path should be associated,[8] while variables that are connected only by closed paths should not. Whether a path is open or closed depends on the variables being *conditioned* upon; conditioning can be explicit (e.g., through regression adjustment or stratification) or implicit (e.g., by analysing a dataset restricted to one or more levels of a relevant covariate, such as only those with response at each wave). Whether a path is open or closed can be determined by applying the four rules of d-Separation (Hernán, 2018a). These rules are shown below with accompanying DAGs. In these and other DAGs, we denote a variable that has been conditioned upon with a red outline and grey fill.

d-Separation Rules

1. If there are no variables being conditioned upon, a path is closed if and only if two arrowheads on the path *collide* at some variable on the path.
    - Below, the path X → Y ← U collides at Y (Y is referred to as a *collider*), and so, without conditioning upon Y, the path between X and U is closed.

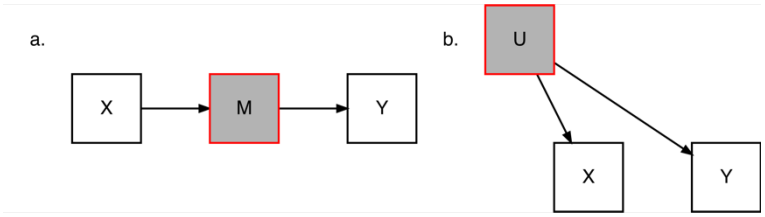
Figure 2.3: d-Separation Rule #1. The path X → Y ← U is closed because the collider Y has not been conditioned upon.



[8] An exception is when paths counterbalance. High winds can blow boats off course, but a good captain handling the tiller may keep the boat's bearing (Cunningham, 2021). Note, as DAGs are non-parametric, statistical relationships may be of a form that simple (e.g., Pearson's) correlations cannot detect.

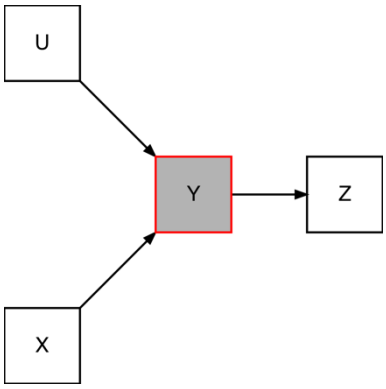
2. Any path that contains a non-collider that has been conditioned upon is blocked.
    - Below, the paths (a) X → M → Y and (b) X ← U → Y are blocked by conditioning upon the non-colliders M and U, respectively.

Figure 2.4: d-Separation Rule #2. (a) The path X → M → Y is closed because M has been conditioned upon. (b) The path X ← U → Y is closed because U has been conditioned upon.



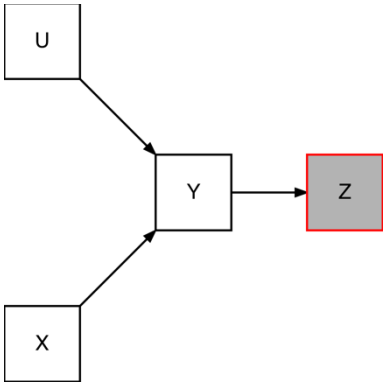
3. A collider that has been conditioned upon does not block a path.
    - Below, the path X → Y ← U is open because the collider Y has been conditioned upon.

Figure 2.5: d-Separation Rule #3. The path X → Y ← U is open because the collider Y has been conditioned upon.



4. A collider that has a descendant that has been conditioned upon does not block a path.
    - Below, the path X → Y ← U is open because Z, a descendant (consequence) of the collider Y, has been conditioned upon.

Figure 2.6: d-Separation Rule #4. The path X → Y ← U is open because a descendant of the collider Y has been conditioned upon.



Applying these rules to Figure 2.2, the path M ← U → Y is open but can be closed by conditioning on U. This path represents the belief that U is a confounder for the relationship between M and Y; we discuss confounding further below. X is also connected to Y via M and U (X → M ← U → Y). M is caused by both X and U and so is a collider on this path (the paths from X and U *collide* at M). The path X → M ← U → Y is therefore closed because neither M nor its descendant, Y, has been conditioned upon. An open collider path generates 'collider bias', a form of selection bias, examples of which abound and which can also affect RCTs. We discuss selection bias further below.

The benefit of drawing the DAG in Figure 2.2 is threefold. First, it is a very compact and straightforward way of representing our beliefs about the causal relationships pertaining to a particular research question. The DAG succinctly communicates these beliefs to other researchers. Second, as noted, DAGs imply a set of empirical predictions that can enable us to test our beliefs. The DAG in Figure 2.2 predicts, among other things, that (a) M should be associated with Y, either unconditionally or having conditioned upon U; (b) X should be associated with M and Y, but not with Y when conditioning upon M and U; (c) U should be associated with Y, either unconditionally or conditioning on M; and (d) X should be independent of U unless conditioning on M and/or Y. If we have data with measurements of X, M, Y and U, we should be able to test these predictions, i.e., evaluate the DAG-data consistency (Textor et al., 2016).

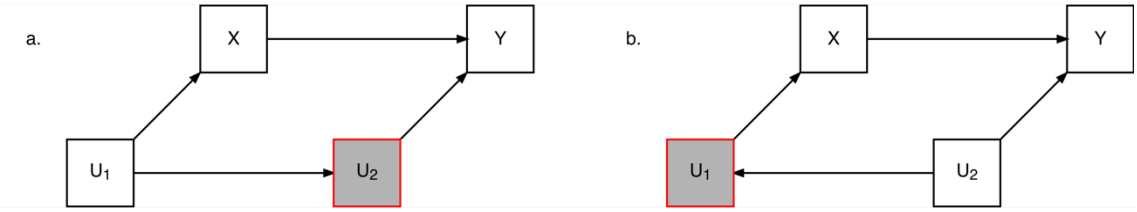
Third, accepting the assumptions encoded in Figure 2.2, we can use the DAG to understand the biases that may apply in an existing study or, better still, plan a new study that obtains an unbiased causal effect estimate even with observational data. For

instance, to obtain an estimate of the causal effect of M on Y, we can condition upon U or use X as an instrument for M (more on this in Section 5.9). This is the power of DAGs.

2.2.3 Confounding Bias

The three main sources of bias in observational studies can each be represented with DAGs. We have already seen an example of confounding in Figure 2.2 – U was a confounder of the relationship between M and Y (M ← U → Y).

Figure 2.7: DAGs representing the structural definition of confounding bias.

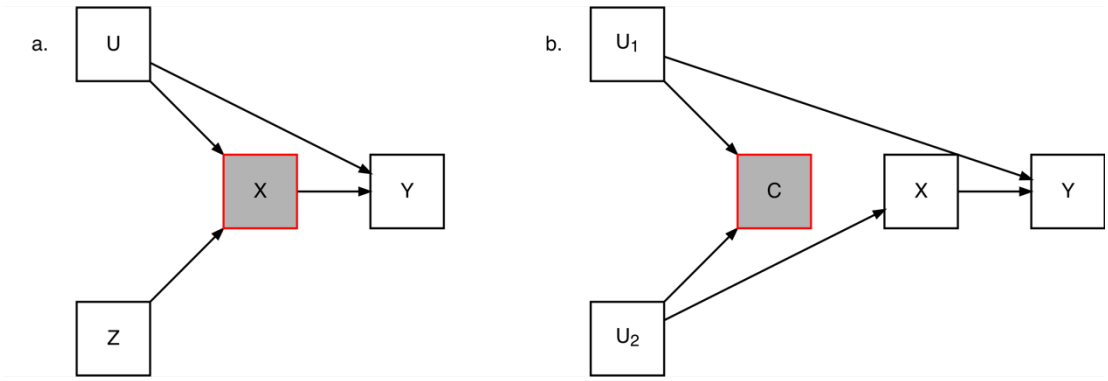


DAGs provide a distinct perspective on confounding. Specifically, the focus is on the bias itself, which is determined by the causal structure, rather than the role individual variables may have, which is the more usual focus (Hernán, 2018a).[9] Though confounders (common causes; Tennant et al., 2021) of exposure and outcome may exist, whether there is confounding bias depends on the variables that have been conditioned upon. For instance, in Figure 2.7a, U_1 is a confounder of the relationship between X and Y ($X \leftarrow U_1 \rightarrow U_2 \rightarrow Y$). Yet, there is no confounding bias because U_2, which is not a common cause of X and Y but blocks the backdoor path between them (d-Separation Rule #2), *is* conditioned upon. A similar observation can be made for the DAG in Figure 2.7b. [10]

[9] For more on variable roles within DAGs, see Tennant et al. (2021).

[10] In Footnote 7, we discussed how DAGs can be high level and do not necessarily have to include all mediators of a causal effect from one variable to another within them (for instance, a DAG representing the effect of education on health might not include a node for income). Given this, we should note that confounders of the relationship between an exposure and a mediator also confound associations between exposures and outcomes (Cinelli, Forney, et al., 2024).

Figure 2.8: DAGs showing issues with data-driven definitions of confounders that do not take causal structure into account. (a) Z meets the data-driven definition even though it does not confound the association between X and Y and, in fact, could be used as an instrumental variable for X and controlling for it, in the absence of control for U, would actually increase bias. (b) C meets the data-driven definition even though controlling for it creates bias that otherwise would not be there.

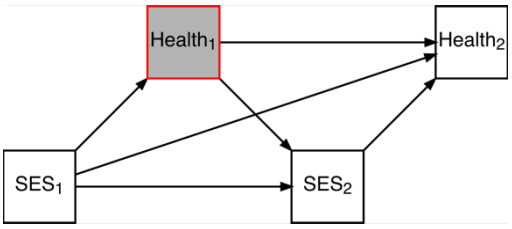


Compare the structural definition with a lay, partly data-driven (and ultimately incorrect) definition of a confounder: a variable that is associated with X and Y (conditional on X) and is not on the causal path between X and Y. Figure 2.8a shows a variable Z that would meet this definition (conditioning on X opens the path Z → X ← U → Y) despite there being no confounding bias from Z (i.e., no path X ← Z → Y): no adjustment for Z is needed, and omitting Z does not by itself bias the estimate of the effect of X on Y. Figure 2.8b shows a variable C that would also meet the data-driven definition and yet controlling for C actually creates bias because it is a collider on the backdoor path between X and Y (this bias is referred to as M-bias; W. Liu et al., 2012). Conditioning upon U_1 and/or U_2 would close this (now opened) path but conditioning on either is unnecessary unless C is conditioned upon.

We can draw two lessons from the DAGs in Figure 2.7 and Figure 2.8. First, as a confounder may only be one step on a longer causal path, it is possible to have multiple options to close biasing backdoor paths. Some may not be "common causes" of exposure and outcome – e.g., according to the DAGs in Figure 2.7, U_1 could be conditioned upon, or U_2 could, or both. Second, theory and domain-specific knowledge are required to make appropriate choices with regard to adjustment sets. "Causal Salad" approaches (McElreath, 2020), where a host of potentially relevant variables are

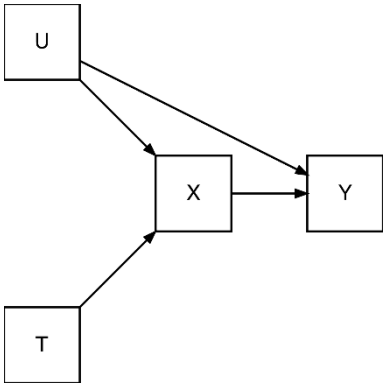
thrown haphazardly into a regression model (e.g., C in Figure 2.8b), can make an analysis more biased than if no adjustments were made at all.

Figure 2.9: DAG representing intermediate confounding.



In Figure 2.7 and Figure 2.8, the exposure was a single point intervention, but many research questions relate to estimating joint effects of multiple exposures or the effect of an exposure varying over time – for instance, questions related to testing models in life course epidemiology (Ben-Shlomo & Kuh, 2002). Where the causal effect of one exposure is mediated by a confounder of others, simple regression adjustment will not be able to identify the joint causal effect of the exposures combined, a problem known as *intermediate* confounding (De Stavola et al., 2015) or *time-dependent* or *time-varying* confounding in the longitudinal setting, specifically (Hernán & Robins, 2023; VanderWeele et al., 2016). Figure 2.9 shows a DAG with an example demonstrating the issue: the effect of low socioeconomic status in both childhood and adolescence upon health (Sheikh et al., 2016). When attempting to identify the joint effect of SES_1 *and* SES_2 upon $Health_2$ (e.g., SES_1, SES_2 = Low, Low vs. High, High), it would be necessary to condition on $Health_1$ to block the (confounding) backdoor path from SES_2 to $Health_2$ via $Health_1$ ($SES_2 \leftarrow Health_1 \rightarrow Health_2$). However, this also blocks part of the causal effect of SES_1 upon $Health_2$ ($SES_1 \rightarrow Health_1 \rightarrow Health_2$). Relatively complex statistical methods beyond simple regression adjustment are necessary in this setting and are described in Sections 5.12 and 5.13.

Figure 2.10: The Randomised Controlled Trial Design.

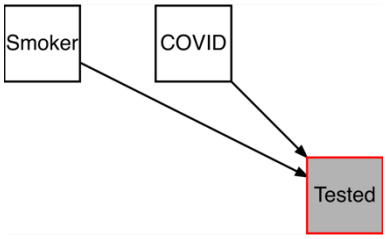


Before moving on to discuss the second threat to internal validity, selection bias, we can briefly show why RCTs are so powerful. Figure 2.10 represents the RCT design within a DAG (note the similarity to Figure 2.2). In a standard RCT, individuals are randomised to treatment or control groups (denoted by T). This randomisation provides a source of variation in the exposure X that is not caused by any other causes of Y, including any confounders (U) of the relationship of X and Y that may exist in the population; that is, treatment is allocated independently of potential outcomes for Y. Given that the effect of T upon Y should only operate through its effect upon X, T can be used as an 'exposure' variable in its own right (as in an "Intention to Treat" analysis) or as an instrumental variable for X. The extraordinary feature of RCTs is that the researcher need not even know what confounders U are – differences are not systematic and chance imbalances contribute to sampling variability.

2.2.4 Selection Bias

Figure 2.8b showed an example of M-bias. M-bias is, in fact, a form of selection bias as it arises from conditioning upon a collider. While most attention is usually placed on confounding as a source of bias in observational studies, selection bias can have greater consequences and be sufficient to generate estimates in the entirely opposite direction to the true causal effect. For example, during the COVID-19 pandemic, there was evidence that suggested smoking was protective against infection. However, as Tattan-Birch and colleagues (2021) note, both smoking and COVID-19 infection make people cough more, which is a reason for getting tested. Among people tested for COVID-19, smoking may be negatively correlated with infection because smokers are more likely to be tested *despite* not having COVID-19 (Figure 2.11).

Figure 2.11: Selection bias in the study of the effect of smoking on COVID-19 where being tested for COVID-19 determines eligibility into the analytic sample. There will be an association between the two even in the absence of a causal effect due to (implicitly) conditioning on a collider.



Since being tested for COVID-19 is a potential consequence of both smoking and COVID-19, it can be conceptualised as a collider variable. As the analytic sample includes only tested individuals, this collider has been implicitly conditioned upon through the study design using restriction, instead of explicit analytical choice (e.g., by regression adjustment, as in the case of M-bias in Figure 2.8b). But regardless of whether conditioning is implicit or explicit, the consequences are the same: selection bias. This can – and should – be represented in a DAG to fully capture biasing processes (Lu et al., 2024). Note, the phrases 'selection' and 'selection bias' are not used consistently across the academic literature, which can cause confusion. Notes 2.1 below offers some clarifications.

Notes 2.1: Notes on selection bias terminology.

- Economists frequently use the phrase 'selection' (i.e., 'selection on unobservables'), but this is synonymous with confounding – e.g., 'selection into unemployment based on human capital'.
- 'Selection bias' is often used to refer to issues of 'transportability' of an estimate to another population (i.e., bias in *external* rather than *internal* validity). This bias occurs when effects are heterogeneous and there are differences in the distribution of effect modifiers in the analytical sample relative to the target population. Lu et al. (2022, 2024) refer to this type of bias as 'Type 2' selection bias.
- Selection bias as we have used it in the paragraph above is alternatively referred to as 'collider stratification bias' and can be separated into 'collider restriction bias' and 'collider adjustment bias' (Lu et al., 2022, 2024).

- Collider restriction bias arises from selecting data points according to levels of a collider (e.g., only studying individuals with a COVID-19 test)

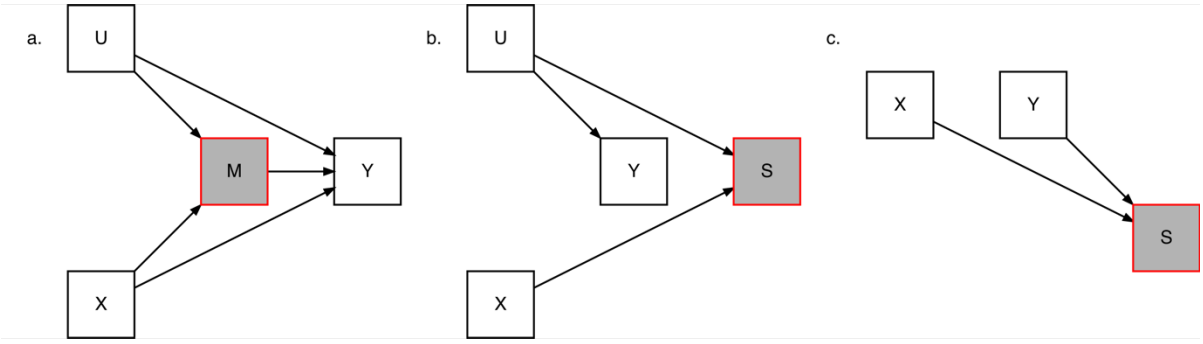
Collider adjustment bias arises from including colliders in a regression model (e.g., as in the M-bias example in Figure 2.8b).

- Though collider adjustment bias is often considered a form of 'selection bias', Lu et al. (2024) argue against this as it does not involve selecting only some data points for analysis, instead referring to it as a form of 'overadjustment' bias.[11] However, we use 'selection bias' to capture both forms of collider bias, given this remains in predominant use.

Selection biases are particularly important to consider in cohort studies due to the possibility of sweep and item non-response. While the initial sample in each of CLS's cohorts was designed (after applying design weights) to be nationally representative, there has been non-random attrition from each study over time (Katsoulis et al., 2024; Mostafa et al., 2021; Silverwood, Calderwood, et al., 2024). The length and depth of the survey questionnaires also mean that some items have less than complete response. Accordingly, analysing only complete cases can mean restricting the sample and implicitly conditioning on colliders, similar to Figure 2.11. In Section 5, we discuss methods to address these problems, but if unaccounted for, non-response can lead to selection bias. Selection bias also has further salience in mediation analyses, which are commonly applied to cohort data because of their longitudinal nature. These analyses are often used when a researcher is interested in decomposing a total causal effect into direct (unmediated) and indirect (mediated) effects, e.g., as in a Baron and Kenny (1986) type analysis.

[11] Other examples of overadjustment bias are conditioning on instrumental variables (see Section 5.9), mediators (where one wishes to estimate a total, rather than indirect, effect), or consequences (descendants) of an outcome (Cinelli, Forney, et al., 2024; Lu et al., 2021; Schisterman et al., 2009).

Figure 2.12: Bias from conditioning upon a collider variable.



These issues (differential non-response and adjusting for a mediator) are represented graphically in Figure 2.12. Figure 2.12a shows a typical mediation analysis where a mediator is explicitly conditioned upon in an attempt to obtain an estimate of a direct causal effect (X → Y). Controlling for M opens the path from X to Y via M and U (X → M ← U → Y). A researcher wanting to obtain an unbiased estimate of the direct (or indirect) effect of X upon Y will not do so using this analysis unless they also condition on U. Note though, here U stands in for *all* variables independently affecting M and Y, so controlling for U may be a difficult task indeed. The size of the bias is determined by the strength of the path X → M ← U → Y – i.e., the size of the effect X and U have on M and the size of the effect U has on Y.[12]

Figure 2.12b-c represent two situations where selection into the analytical sample is related to exposure and outcome. Because we can only analyse the data that are available, in both cases, a spurious association is introduced between the two. We have already noted that this may arise due to issues with the data collection process – for instance, if we have unit, sweep or item non-response – but it can also arise for substantive reasons: we only observe people who are alive and we may observe a variable only for some individuals because it is only defined for a certain group (e.g., wages are only defined for the employed). This can have important ramifications. For instance, at older ages smoking is associated with a *lower* risk of dementia, but this is likely due to the onset of dementia occurring late in life after many heavy smokers have

[12] This provides a straightforward solution to the so-called "obesity paradox" of obese individuals displaying greater survival rates from cardiovascular disease (CVD; Banack & Kaufman, 2014). As CVD is a consequence of obesity, comparing survival rates among obese and non-obese individuals with CVD compares obese people against those who must have developed CVD for other reasons, some of which may have far worse prognoses, even if obesity itself is deleterious for health.

died (Hernán, 2018a); surviving smokers are relatively robust compared to the larger proportion of surviving non-smokers but not *because* they smoke.

Before moving on to discuss the third threat to internal validity, information bias, we note that selection bias can also affect RCTs for the same reason as observational studies. Selection bias would, for instance, arise where there is attrition from the trial that is related to both treatment and the outcome. Selection bias can also arise in mediation analyses as conditioning upon a consequence of an exposure leads to incomparable groups. To give an example: imagine an RCT in which participants were randomised to a work placement scheme (treatment) or no intervention (control). We may be interested in the long-term effect of work placement on wages, but wages are only observed for those with a job. If the work placement scheme increases employment, then the employed individuals in the treatment group will include not only those who would have found a job anyway, but also those who are only employed because of the scheme. By contrast, the employed individuals in the control group will only include those able to find a job without the scheme. If the additional employed individuals in the treatment group tend to earn lower wages, a naïve comparison of wages among those employed will underestimate the effect of the scheme on wages. See Table 2.1.

Table 2.1: Hypothetical example of selection bias in an analysis of an RCT for a job placement scheme

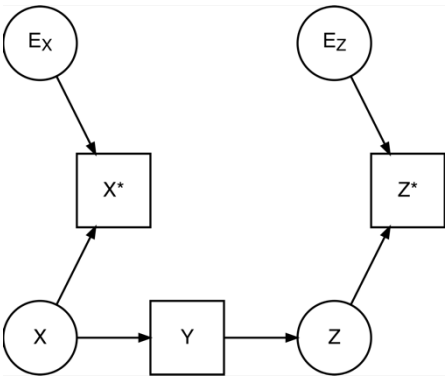
Employability	Treatment group (work placement)	Control group (no work placement)	Difference in weekly wages
High	£600 p.w.*	£550 p.w.**	+ £50 p.w.
Low	£400 p.w.*	£350 p.w. (latent: not observed)	+ £50 p.w. (not observable)
Mean wage among those with a job (observed)	£500 p.w.	£550 p.w.	- £50 p.w.
Mean wage among everyone (true)	£500 p.w.	£450 p.w.	+ £50 p.w.

* Observed: everyone in the treatment group has a job, including those employed only because of the placement.

** Observed: in the control group, only people with high employability find a job.

2.2.5 Information Bias

Figure 2.13: Non-differential and independent measurement error in two variables, X* and Z*.



The third threat to internal validity is information bias. Information bias arises when measured variables are not perfect representations of the phenomena they are

designed to capture – i.e., when variables are measured with error. In DAGs, measurement error is represented by distinguishing the measured variable (X*) from the latent, unobserved variable (X) and drawing the measured variable as a descendant of the latent variable and an error node (E_X). Nodes for latent variables and unobserved errors are displayed as circles or ellipses. See Figure 2.13 for an example.

Errors can be classified along two dimensions: they can be *dependent* or *independent* and *differential* or *non-differential*. An error is dependent (as opposed to independent) if it is correlated with the measurement error of another variable. An error is differential (as opposed to non-differential) if its value is related to the true value of a variable (e.g., error in the measured outcome depends on exposure). Figure 2.13 displays a situation where there are independent and non-differential errors in two study variables, X* and Z*; classical measurement error is a type of independent, non-differential measurement error.

Whether independent, non-differential error leads to a biased estimate of a causal effect depends on the analysis. In a simple linear regression with classical (additive) measurement error in the exposure ($X^* = X + E_X$), the coefficient on the exposure will be biased towards the null. Intuitively, this is because each unit increase in X* can be either caused by X, which has an effect on Y, or E_X, which does not. This phenomenon is known as regression dilution (Hutcheon et al., 2010). On the other hand, if it is the outcome that is measured with classical measurement error ($Y^* = Y + E_Y$), there is no bias, but statistical precision is reduced.

Figure 2.14: Combinations of differential and dependent measurement error in two variables, X* and Y*.

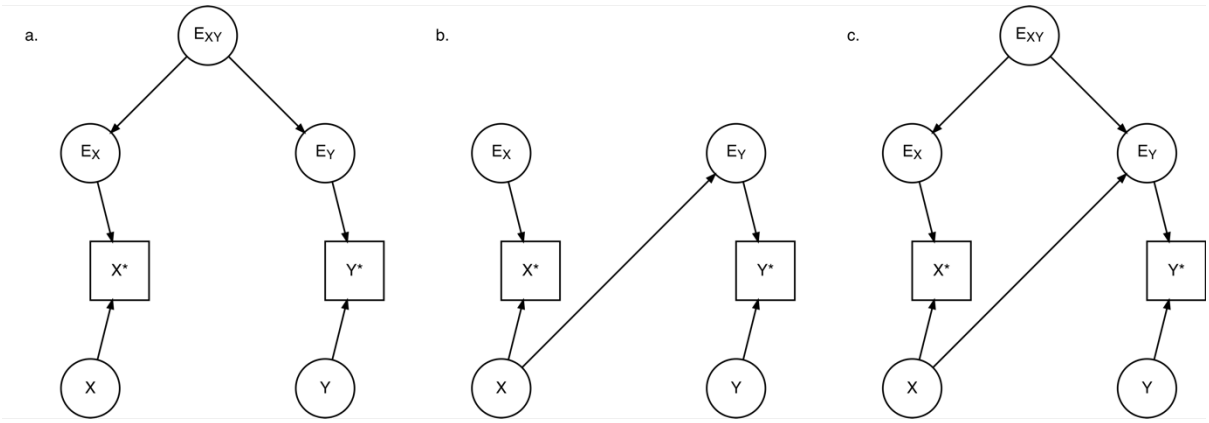
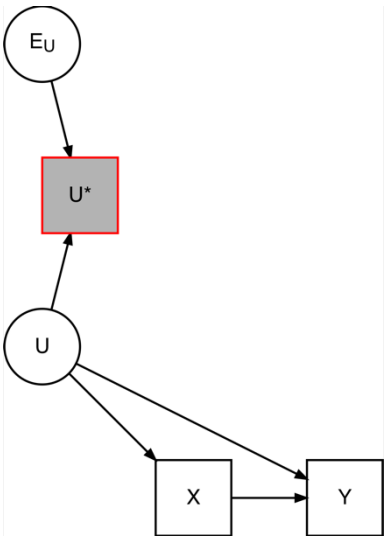


Figure 2.14 completes the three other combinations of dependent/independent and differential/non-differential error. In each case, X does not have a causal effect on Y but X* and Y* are associated due to measurement error. Figure 2.14a displays a situation where there are dependent, non-differential errors in the two measured variables, X* and Y*. X* and Y* are associated via the backdoor path $X^* \leftarrow E_X \leftarrow E_{XY} \rightarrow E_Y \rightarrow Y^*$, even though there is no causal effect of the latent variable X on Y. Figure 2.14b displays a situation where there are independent, differential errors in two measured variables X* and Y* (Y* is differential with respect to X). X* and Y* are associated via the path $X^* \leftarrow X \rightarrow E_Y \rightarrow Y^*$, even though there is no causal effect of X on Y. Figure 2.14c displays a situation where there are dependent, differential errors in two measured variables X* and Y* (Y* is differential with respect to X). X* and Y* are associated via the two biasing paths in Figure 2.14a-b even though there is no causal effect of X on Y: $X^* \leftarrow X \rightarrow E_Y \rightarrow Y^*$ and $X^* \leftarrow E_X \leftarrow E_{XY} \rightarrow E_Y \rightarrow Y^*$.

Figure 2.15: Measurement error in a confounding factor.

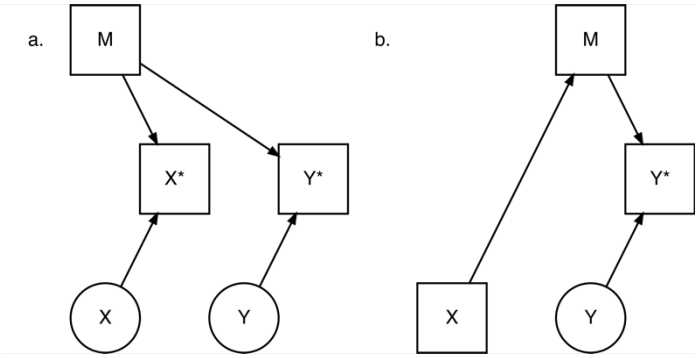


Information bias is not only a problem where exposure and outcome variables are subject to measurement error but also when there is measurement error in variables used to close biasing paths. Figure 2.15 displays a common situation where the confounder U is measured with some error. Since U* is not a perfect measure of U, conditioning on U* will not close the backdoor path $X \leftarrow U \rightarrow Y$ completely, leaving residual confounding. Thus, the effect of X on Y is still confounded, though the size of the resulting bias depends on the relative size of the error in U* and the strength of the

confounding pathway. The more accurate the measurement, the less bias; the stronger the confounding, the greater the bias.

A potential source of information bias that users of CLS's data should be particularly aware of is *survey mode effects*. Each of CLS's cohort studies has at various points adopted mixed-mode designs, with cohort members responding via, e.g., web, video, telephone or face-to-face interview. Mixed-mode designs mean survey modes differ between individuals *within a sweep* or within individuals *across sweeps*. Mixed-mode designs can increase response rates while also lowering the cost of data collection. A drawback, though, is that responses to survey items can differ systematically between modes, in particular due to differences in interviewer presence/absence and in how items are presented (Tomova, Silverwood, & Wright, 2026a). For instance, a participant may be less willing to answer sensitive questions accurately in non-anonymous modes. Differences between modes due to how items are measured are termed mode effects.

Figure 2.16: Bias from mixed-mode designs. (a) Bias due to shared mode effects in two measured variables, X* and Y*. (b) Bias due to a combination of mode effect (M → Y*) and mode selection (X → M). Note, mode is an observable source of measurement error.

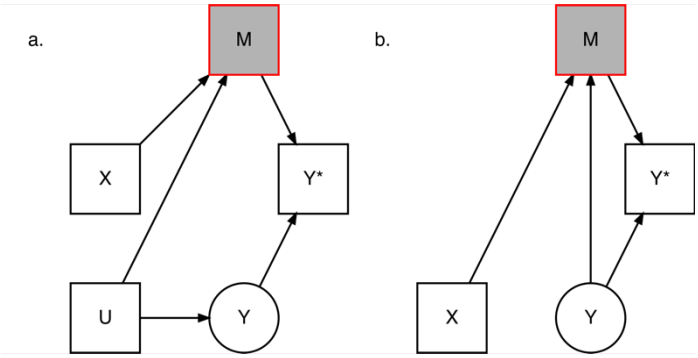


If unaccounted for, mode effects can cause bias because they induce dependent measurement errors; two variables subject to mode effects will be associated where mode is shared (Figure 2.16a; X* ← M → Y*) or correlated. Mode effects can also cause bias because they are differential (Figure 2.16b; X → M → Y*); in mixed-mode designs, participants are typically not randomised to mode, and the characteristics of participants in each mode can differ due to active participant choice or as a direct or indirect consequence of the fieldwork design. For instance, people without a telephone number may only be offered face-to-face interview, while in the sequential mixed-mode

design where one mode is only offered after non-response to previously offered modes, groups who are slower to respond will only appear in modes offered last. Systematic differences between modes due to *who* is being measured are referred to as *mode selection*.

In the DAGs in Figure 2.16, bias can be removed by conditioning upon mode (e.g., by including mode as an indicator variable in a regression model). However, conditioning upon mode can generate collider bias where there are unaccounted-for factors that also determine selection into mode (X → M ← U → Y, Figure 2.17a), a Gordian Knot-like problem when the unobserved factors are the latent variables of interest themselves (X → M ← Y, Figure 2.17b).

Figure 2.17: Bias from controlling for mode in the presence of mode selection.



CLS has recently released guidance for users of its cohort data on handling mode effects (L. Wright, Ploubidis, et al., 2024). This guidance includes a description of the mixed-mode elements in each cohort, DAGs to conceptualise a wider set of situations in which mode effects can bias analysis, and discussion of methods for accounting for mode effects (see also Tomova, Silverwood, Tennant, et al., 2026; Tomova, Silverwood, & Wright, 2026b).

2.2.6 Researcher degrees of freedom

In the previous three sections we have seen how bias can arise from analytical decisions and the data-generating process, but bias can also arise when researchers (or editors or publishers) choose which results to publish. Many analyses admit multiple defensible specifications (e.g., how to treat missing data, how to operationalise key variables), which may give slightly different results, some more favourable to the

researcher's aims or prejudices, some less. If the analyst spends their "researcher degrees of freedom" (Simmons et al., 2011) on getting the answer they want, the absence of confounding, selection, and information biases is no guarantee of obtaining an unbiased estimate. This problem can be lessened through pre-registration of analysis plans, including using registered report publication models (Crüwell et al., 2019; Weston et al., 2019). Methods that involve running and presenting the "universe" of defensible model specifications, such as Specification Curve Analysis (Simonsohn et al., 2020) or Multiverse Analysis (Steegen et al., 2016), can also help. See Orben & Przybylski (2019) for an example of specification curve analysis using MCS data.

2.3 Why cohort studies?

Given the empirical and theoretical expectation of the potential for bias in cohort and other observational studies (see Section 8 [Appendix] for a review), readers may wonder why cohort studies should be used at all. Besides highlighting the utility of cohort studies for purely descriptive analyses, a typical defence would note that there is little other option – RCTs are expensive, potentially unethical, and in many cases impractical or even impossible. Large cohort studies can answer many different research questions at relatively low cost. Because they already exist, they can give answers quickly. And because they record behaviours that occur anyway (e.g., problematic drug use), they avoid the ethical issues that can arise in RCTs.

But this is a negative case. More positively, there are situations that are recorded in observational data where confounding and other forms of bias are unlikely to exist or to materially impact results, including "natural" experiments. There are also frameworks and methods that have been developed which, when used in appropriate situations, can help researchers draw sensible causal inferences from observational data. Methods also exist for detecting and correcting for bias *post hoc*, and in many important cases the size or direction of bias will not be sufficient to substantively affect results. Compelling causal claims can therefore still be made with observational data even when bias is present and known. In Section 4, we further discuss the attributes of cohort data that support causal inference, and in Section 5, we describe methods that exploit these attributes.

3. Three approaches to causal inference

Building on Section 2, we now outline three approaches for attempting to make causal inferences with cohort data: model-based approaches, design-based approaches and *elaborate theories*. These approaches are not mutually exclusive and the boundaries between them are not sharp. They do not, for instance, map onto a set of non-overlapping methods – a particular study may use a combination of approaches, as the assumptions of one may only be correct when using another, or alternatively could use multiple approaches to examine whether approaches with different biases converge on the same answer. For example, an instrumental variable (IV) study may also use model-based statistical control for known confounders to improve the plausibility of the IV assumptions (Brumpton et al., 2020). Regardless, delineating and describing the three approaches helps clarify the virtues and limitations of each, and provides a useful organising schema for the specific causal inference methods detailed in Section 5.

3.1 Model-based approaches

The model-based approach is the most frequently used in quantitative research. In this approach, the researcher develops a statistical model – or uses other related algorithms, such as matching – to account for the influence of factors that bias the relationship between exposure and outcome in the available data. The model developed reflects the researcher's beliefs about the underlying data-generating process: accounting for biasing influences is *post hoc*. The purpose of the model is to remove the influence of biasing factors in the exposure or outcome – to "attempt to compare people who are not directly comparable" (Rosenbaum, 2020). The model-based approach is often referred to as the non-equivalent groups design (Reichardt, 2019), which highlights the fundamental problem.

The most popular model-based method is multivariable regression, where variables other than the exposure are added as covariates and their influence on the outcome modelled. Stratification – where associations are calculated within sub-strata of the data (e.g., males and females, separately) and then pooled – can also be considered a model-based approach, though it does not involve modelling the influence of the

stratifying variables directly; within strata, exposure assignment is assumed to be as-if random (potentially after adjustment for additional variables, as described above).

Model-based approaches can be more principled or less. With correct knowledge of the data-generating process (e.g., how individuals are assigned to the exposure), accurate measurement of all relevant factors, and an appropriate understanding of how to incorporate the data into a model (e.g., with the correct functional form and sensible choice of adjustment set), the model-based approach would recover a causal effect, sampling variation notwithstanding. However, this is not how the model-based approach is typically used. Perhaps more frequent are 'Causal Salad' approaches (McElreath, 2020) where an array of covariates is haphazardly thrown into a statistical model, with only a vague sense of how these will ameliorate bias as the researcher has not systematically considered the causal network within which these covariates sit; as shown in Section 2.2.3, a researcher can unwittingly increase bias by conditioning on certain variables.

Though researchers using model-based approaches may note (usually in a Limitations section) the possibility of bias remaining, the ambition is to have obtained a causal estimate. The caveat is regularly vague or noncommittal and the researcher may write the rest of the paper as if a causal estimate had been obtained, though it had not. Because the model-based approach relies upon recapitulating the data-generating process, and in many circumstances this is unknown, it is easy for analyses to obtain biased estimates of causal effects and for researchers to fool themselves into believing this is not the case.

3.2 Design-based approaches

The design-based approach differs from the model-based approach in that the researcher uses knowledge about treatment assignment to identify *a priori* circumstances where treatment assignment should be as-if random. Data are analysed to exploit this 'exogenous' (originating outside the system) variation to obtain a causal effect. Examples of this approach include studies that use natural experiments or quasi-experimental approaches such as regression discontinuity or instrumental

variable designs, where for some set of individuals, exposure is proposed to be unrelated to factors that confound the relationship of interest more generally.

As with the RCT, randomness of the assignment mechanism can guarantee exchangeability, even without detailed understanding of what potential confounders may be. For instance, who wins a prize among lottery players is a fully random process (that is, unless the game is rigged), which is why lottery wins have been used to study the causal effect of wealth on health (Cesarini et al., 2016). However, not all design-based approaches exploit straightforwardly random processes. Assignment may be for fully legible, predictable reasons, but still be exploitable in a design-based approach if it can be argued that assignment is not related to (potential) outcomes, except through its effect on exposure. This is the basis of many instrumental variable studies. For instance, in a classic paper in political economy, Acemoglu et al. (2001) use early European settler mortality rates as an instrument for whether a colonising country set up extractive institutions in a colonised country in order to assess the effect of institutional quality on long-term economic development. Their argument is that countries where tropical disease made mortality high would be less attractive for individuals from the colonising nation to settle in, thereby incentivising the development of institutions oriented to short-term resource extraction. Identifying this variation and arguing convincingly that it meets the conditions for a valid instrumental variable (Section 5.9) requires deep domain-specific knowledge.

Additionally, some design-based methods are not able to account for all sources of bias but instead use knowledge about assignment to account for distinct classes of confounder. For instance, sibling designs account for shared familial factors by comparing outcomes among siblings discordant on an exposure, without knowing exactly what these shared familial factors are (Donovan & Susser, 2011): factors that are shared do not contribute to the outcome variance within sibling sets and so cannot confound the association between exposure and outcome, though factors that aren't shared still have the potential to confound associations. The sibling design and other related methods are examples of *isolation*, where a smaller set of comparisons is exploited that is expected to be less biased than the wider set of comparisons that could be made (Zubizarreta et al., 2018).

3.3 Elaborate theories

The model- and design-based approaches are concerned with causal effect estimation. The Elaborate Theories approach follows a different tack and focuses more broadly on strengthening causal inference (which, according to Shadish & Cook, 1999, is 'more a matter of logic than statistics'). Instead of being occupied with identifying random variation, it takes theories seriously and asks what specific predictions a theory makes about the world. Instances where these predictions differ from those of competing theories are where causal inferences can be made. Two variables, X and Y, being associated because X *causes* Y is a theory, but X and Y being associated because they are confounded by U is a theory, too. Both theories may entail a larger set of (more specific) predictions – a *pattern* of anticipated empirical observations (Trochim, 1985, 1989) – that may not fully overlap; the places of non-overlap are where useful information can be gleaned.

A historical example may make this more concrete. In the mid-19th century, London was beset by outbreaks of cholera; one such outbreak in August 1854 centred on Soho's Broad Street pump (Adhikari et al., 2024; Coleman, 2019). The prevailing theory of the time was that cholera was transmitted via *miasmas* – bad airs. A prediction of this theory was that individuals living close together would succumb to the illness as they breathed the same air. The physician John Snow had another theory, however – that cholera was caused by contaminated water. This theory was supported by several observations that were inconsistent with miasma causing the disease, including: no deaths among workers at the nearby Lion Brewery, who did not take water from the Broad Street pump (instead drinking the beer they produced); and the case of Mrs Eley, a Hampstead resident who died of cholera, who had water from the Broad Street pump delivered to her house as she liked the taste, an unfortunate preference, given a cesspit containing sewage from cholera victims had been leaking into the well the pump drew from. At one level, both the miasma and contaminated water theories made the same prediction – clustering of cases in certain areas – but at a more specific level, the predictions varied, and this allowed John Snow to provide compelling evidence, even with purely observational data, of how cholera was caused.

Exceptions to theories like the inebriated employees of Lion Brewery are rarely so vivid, but the Elaborate Theory approach can still prove insightful when arguments are drier. For instance, Elaborate Theory-style reasoning has been used to appraise explanations of the well-established association between socioeconomic position (SEP) and health. One explanation is that the associations reflect confounding by cognitive ability (Gottfredson, 2004), rather than a causal effect of SEP as more commonly claimed (Link & Phelan, 1995; Marmot, 2015, 2016). One corollary of the former theory is that correlations between SEP and health should be stronger for measures of SEP that are more closely related to cognitive ability (education > socioeconomic class > income) as confounding would be stronger, a prediction not necessarily entailed by *causal* explanations of the association (e.g., associations driven by the causal effect of differences in financial resources). The empirical observation that associations are not stronger for measures of SEP more closely related to cognitive ability is therefore evidence against cognitive ability being sufficient to explain the social gradient in health (Singh-Manoux et al., 2005) and refocuses attention on other possible confounders.

The Elaborate Theories approach had its first expression in Sir Ronald Fisher's gnomic reply to a question from William Cochran on how to move from correlation to causation in observational data ('Make your theories elaborate'; quoted in Cochran, 1965). Cochran concurred with this assessment ("this multi-phasic attack is one of the most potent weapons in observational studies") and the approach has since been staunchly and eloquently advocated by, among others, the statistician Paul Rosenbaum, including in his superlative book *Observation and Experiment* (Rosenbaum, 2019), which we highly recommend. See also Rosenbaum (2023), Reichardt (2019), Shadish & Cook (1999), Shadish et al. (2002), and Trochim (1985, 1989).[13]

How does one make a theory elaborate in practice? Several methods exploit the logic of Elaborate Theories and are particularly suited to cohort data. Generally, these exploit

[13] The Elaborate Theories approach has alternatively been referred to as "pattern matching" (Reichardt, 2019; Trochim, 1985, 1989). Expressed by Shadish et al. (2002; quoted in Reichardt, 2019): "The more complex pattern that is successfully predicted, the less likely it is that alternative explanations could generate the same pattern, and so the more likely it is that the treatment had a real effect".

situations where an important hypothesised mechanism of action (either for the main causal effect of interest or for competing explanations) is expected to be absent or else stronger or weaker than otherwise. Above, confounding was expected to act more strongly for some measures of SEP than others. The following examples give additional strategies.

- **Variation in active ingredient:** Observational studies attempting to ascertain whether cannabis has a causal impact on schizophrenia risk may exploit the expectation that correlations between cannabis use and schizophrenia should be stronger where concentrations of THC (the supposed schizophrenia-inducing compound) are greater (Gage et al., 2016).
- **Variation in confounding structures:** Observational studies that want to ascertain the effect of breastfeeding on child development may compare associations in settings where the social patterning of breastfeeding (a major hypothesised confounder) is entirely opposite (Brion et al., 2011).
- **Specificity of association:** Observational studies attempting to ascertain the carcinogenic effect of tobacco use could determine whether the profile of specific cancers is consistent with how the tobacco is used – for instance, an excess of lung cancers among cigarette smokers (who inhale smoke) and mouth cancers among cigar smokers (who do not) (Cornfield et al., 1959).
- **Timing of effects:** The legalisation of abortion may reduce aggregate crime rates by disproportionately leading to terminations of pregnancies that would have yielded children at higher risk of going on to commit crimes (Donohue & Levitt, 2001). This effect should only be observed as the cohorts affected by such legalisation grow to an age where they are more likely to commit crime (i.e., adolescence and early adulthood). Similarly, flu vaccination should only have effects on mortality at times of the year when influenza is prevalent (L. A. Jackson et al., 2006).
- **Absence of hypothesised mechanism:**
    - *In a subgroup*: Observational studies that want to ascertain the effect of alcohol on cardiovascular disease may surmise that genes related to alcohol use (such as the ALDH2 gene that determines the comfort with which one

can metabolise alcohol) should also be related to cardiovascular disease but only among populations where alcohol is used (L. Chen et al., 2008).

- *In a historical period*: Observational studies that seek to understand the effect of socioeconomic position on a health outcome and posit access to health information and technologies as a key mediator may examine changes in inequalities in health before and after relevant medical knowledge is discovered or publicised (Clouston et al., 2016).
- *In a geographic area*: Observational studies attempting to ascertain whether policies to reduce teenage pregnancies were effective may investigate whether reductions in teenage pregnancy rates were confined to countries in which policies were implemented (Baxter et al., 2021).

A concept related to Elaborate Theories is the *quasi-experiment device*. Studies using quasi-experimental devices utilise additional data to disambiguate an observation that would otherwise be ambiguous – e.g., to judge whether an association reflects a causal effect of interest or that of a competing theory (Rosenbaum, 2019). For instance, to assess whether the English Teenage Pregnancy Strategy achieved its aims, a quasi-experimental device would be to use comparator data from Scotland and Wales (Baxter et al., 2021).

4. Attributes of CLS's cohort studies that support causal inference

Our thesis in this document is that the richness of the data in CLS's cohort studies can support researchers in making causal inferences. In this section, we highlight several features of these data that are especially useful – specifically, the breadth and depth of the data, the ability to perform data linkage, the longitudinal nature of the studies, the extended timespan they collectively comprise, their large sample sizes, and their capacity for harmonisation with other cohort studies.

4.1 Breadth and Depth

Collectively, CLS's cohort studies include several tens of thousands of variables available under then standard "Safeguard data" End User Licence on the UK Data Service. These variables are largely derived from survey questionnaires – the ability to link with other data increases this number dramatically. Combined, these data offer profound insight into participants, their lives and the milieus in which they are situated. Data collections span birth (or, in the case of Next Steps, adolescence) to the current day. Within a survey sweep, data are collected on a range of participant characteristics – for instance, batteries of questions on mental and physical health, psychosocial development, socioeconomic circumstances and the home environment. This includes frequently analysed factors, such as employment status, but also relatively niche information, such as participants' use of suncream, which was asked at the Age 44/45 year Biomedical Sweep of the NCDS. (In Section 5, we will argue that seemingly unimportant information like this can be used to support causal inference – for instance, as falsification tests.) The range of data available also makes assumptions of no unobserved confounding (in standard regression) or missingness at random (in multiple imputation) more plausible (Mostafa et al., 2021). Extensive data are also available on cohort members' parents, partners, children and siblings. This can enable several causally informed research designs, such as sibling comparison and parental comparison designs and within-family Mendelian Randomisation.

4.2 Extensive Data Linkages

Data linkage increases the size and variety of information available, mechanistically increasing the possibilities for causal inference (e.g., increasing sample representativeness; Rajah et al., 2023). Some data linkages represent opportunities that are specific to the type of data being linked. Some of this data linkage may be qualitative and contextual; for instance, linkages that are possible from knowing significant dates (e.g., interview dates or dates of birth). Fitzsimons & Vera-Hernández (2022) use information on whether MCS participants were born at the weekend as an instrument for maternal breastfeeding (due to lower staffing on maternity wards, new mothers are less likely to be offered breastfeeding support on weekends). Other examples of this contextual linkage include studies of the effects of news and current events on mental health and other outcomes using interview date as a determinant of exposure to the information (Muñoz et al., 2020).

CLS's cohorts are increasingly linked to administrative and geographical data and each cohort has also now been genotyped, allowing for genetic data to be linked with survey responses. Administrative data and geographic linkages provide many opportunities for utilising quasi-experiments – events or episodes in which exposures are allocated as-if randomly. Examples include studies exploiting arbitrary thresholds that are used to decide eligibility for programmes or interventions (e.g., thresholds for determining low birth weight, grading schemes in education that determine a pass or a fail, or geographic boundaries that determine the government policies one is subject to). Genetic data enable a specific form of instrumental variable analysis – Mendelian Randomisation – that exploits random assortment of parental genes during meiosis.

4.3 Longitudinal Assessment

The longitudinal nature of the studies means that the same or similar questions are asked across time. For instance, in the MCS, there are questions on several dimensions of mental health asked across childhood to late adolescence and now into early adulthood. Longitudinal data have four properties that aid causal inference. First, prospective data collection and repetition can enable more accurate measurement of

study variables. Second, longitudinal assessment allows researchers to specify the correct temporal ordering (causes should precede consequences) with variables measured at different sweeps. Third, repeated measurement allows researchers to control for prior exposures and outcomes. Controlling for prior exposure allows for the 'treatment' to be more specifically defined (e.g., unemployment at age 33y among those who have not yet been unemployed), and controlling for prior exposure and/or outcomes can absorb some of the confounding bias that may otherwise exist as well as account for reverse causality processes (VanderWeele et al., 2020). Fourth, one can use within-person variation, rather than rely solely on between-person differences. Using within-person variation enables researchers to, e.g., account for time-invariant confounding factors by design, rather than through statistical control (Section 5.5).

4.4 Extended Timespan

Mainstage data collections in CLS's cohorts collectively span the 68 years from 1958 to the present day. This period can be extended when combined with other cohort studies, such as the National Survey of Health and Development (NSHD; the 1946 British Birth Cohort Study; Kuh et al., 2011; Wadsworth et al., 2006). Several profound societal-level changes have occurred over this period, including the obesity epidemic, the expansion of further and higher education, and the greater participation of women in the labour market (Anders & Dorsett, 2017; Erikson & Goldthorpe, 2010; Gondek, Bann, Patalay, et al., 2021; Johnson et al., 2015). As will be expanded upon in Section 5, these societal-level changes offer several opportunities for causal inference. Briefly: confounding structures for some exposure-outcome relationships may have changed across time or causal effects may be confined to some periods or may vary across time in predictable ways (e.g., stronger in some periods than others). Elucidating each of these possibilities is one way in which theories can be made *elaborate* and causal arguments can be developed.

4.5 Large Sample Sizes

The large sample size of CLS's cohorts is important in its own right for precise statistical estimation – some causal inference methods, such as instrumental variables, are

power hungry – but it also means that important subgroups can be identified. For instance, where an exposure is confined to a local geographic area (e.g., the 1962 London smog), a large overall sample size can mean sufficient numbers appear in the exposed group. There may also be circumstances where a research design is more plausible within a subgroup; for instance, Rosenbaum (2020) argues that a frequently used instrumental variable, time to nearest hospital, can be made more plausible by focusing on individuals outside large cities who have less choice over the hospital they use. Further, some subgroups possess particular properties that make them useful for analyses. This includes groups that are frequently used in family-based designs. Each of CLS's cohorts contains a number of twins and data are sometimes collected from other (older or younger) siblings, too; for instance, MCS contains two sweeps of information from an older sibling. NCDS also contains a number of adopted children – again, useful in particular family-based study designs. Some of these subgroups may be small, but as the next section notes, CLS's cohorts may be harmonised and analysed alongside other cohort data, so the subgroups within CLS's cohorts can contribute to a larger analytic sample.

4.6 Ability to Harmonise Data

Harmonising data with other cohorts can expand the time span of studies or the geographic area covered. This again enables studies that test elaborate theories: confounding structures may differ across space as well as time, and causal effects may be present, weaker or stronger in some areas than others. For instance, Brion et al. (2011) use UK and Brazilian cohorts to assess the effect of breastfeeding on development – the (confounding) association between SEP and breastfeeding is strong in the UK, but not in Brazil. Various efforts have been made to collate and harmonise data across different cohorts.

CLOSER (2023) has released several projects harmonising data from CLS's cohorts on, among other things, socioeconomic status (see also Centre for Longitudinal Studies, 2026). Also useful are the Atlas of Longitudinal Datasets (2024) and the Catalogue of Mental Health Measures (2026), whose names are self-explanatory, and Harmony

(2023), an online tool that uses natural language processing to identify survey items that are similar semantically across different studies (including non-CLS studies).

5. The methods of causal inference

In this section, we list individual methods that can, when assumptions are met, either directly help estimate causal effects with CLS's cohort data, or strengthen any causal inferences made. Generally, we focus on methods that account for confounding, though methods to account for measurement error or selection bias are also discussed. This section is something of a gallimaufry: some of these methods relate to statistical *estimators* (e.g., multivariable regression; Section 5.1), others to identification *strategies* (e.g., instrumental variables; Section 5.9), and others to *falsification tests* (e.g., negative controls; Section 5.15). As with the 'Three Approaches to Causal Inference' described in Section 3, the individual methods are not mutually exclusive, and a single study may use a mix of them (e.g., multiple imputation to address missing data, with statistical control to meet the assumptions behind instrumental variables analysis).

For each method, we provide three pieces of information: Method, Applications, and Opportunities. *Method* provides a brief overview of the method, its strengths and limitations, with references for further reading. The focus is on intuition, rather than the statistical and algorithmic nuts and bolts. *Applications* gives examples of papers using the method with CLS data. We focus on particularly well-executed studies or examples that demonstrate issues that can arise. For other examples, please search the CLS Bibliography website. *Opportunities* discusses general research ideas that could utilise the method in question with CLS data. Readers interested in learning more about these methods are invited to read textbooks by Huntington-Klein (2022, 2026), Cunningham (2021), Fox et al. (2021), Rosenbaum (2019, 2020), Hernán and Robins (2023), and van Buuren (2018). Some of the examples and explanatory material in this section are derived from these textbooks.

Subsections (links are clickable)

5.1 Control via Regression Adjustment 50

5.2 Matching 57

5.3 Inverse Probability of Treatment Weighting 62

5.4 Missing Data Methods 65

5.5 (Longitudinal) Fixed Effects 71

5.6 Difference in Differences 76

5.7 Study-Initiated Experiments 84

5.8 Natural Experiments 89

5.9 Instrumental Variables 95

5.10 Regression Discontinuity Designs 105

5.11 Target Trial Emulation 113

5.12 G-methods for Time-Varying Exposures 118

5.13 Causal Mediation 127

5.14 Family Designs 142

5.15 Negative Control Outcomes and Exposures 147

5.16 Cross-Context Analysis 154

5.17 Biases of Known Direction 159

5.18 Differential Effects 162

5.19 Reasons for Effects 165

5.20 Quantitative Bias Analysis 168

5.1 Control via Regression Adjustment

5.1.1 Method

The most common method for accounting for confounding bias is to take a model-based approach and controlling for confounders or other nuisance factors via regression adjustment – for example, by adding covariates to an ordinary least squares (OLS) regression model. The data-generating process is modelled mathematically, with elements of the model (e.g., regression coefficients) learned from the data itself (Rosenbaum & Rubin, 1983). Statistical adjustment models the effect of relevant differences between treatment and comparison groups to block backdoor or other nuisance paths between exposure and outcomes, isolating variation in the exposure and outcome unrelated to these (observed) confounders.

This description is not just poetically true but reflects the underlying mechanics. The Frisch-Waugh-Lovell theorem demonstrates that the coefficient obtained by multivariable regression of an outcome upon an exposure plus covariates is equivalent to (1) obtaining residuals from regressions of (a) the exposure on covariates and (b) the outcome on covariates and then (2) regressing the residuals from (b) upon (a). In other words, statistical control winnows, and then utilises, residual variation. Readers can confirm this in R or Stata using the code below.

```
// Stata Code
sysuse auto, clear
regress price mpg length foreign // full model
regress mpg length foreign
predict mpg_resid, residuals
regress price length foreign
predict price_resid, residuals
regress price_resid mpg_resid

# R Code
data(iris)
lm(Sepal.Length ~ Sepal.Width + Petal.Length + Species, iris) # full
model
width_mod <- lm(Sepal.Width ~ Petal.Length + Species, iris)
length_mod <- lm(Sepal.Length ~ Petal.Length + Species, iris)
lm(residuals(length_mod) ~ residuals(width_mod))
```

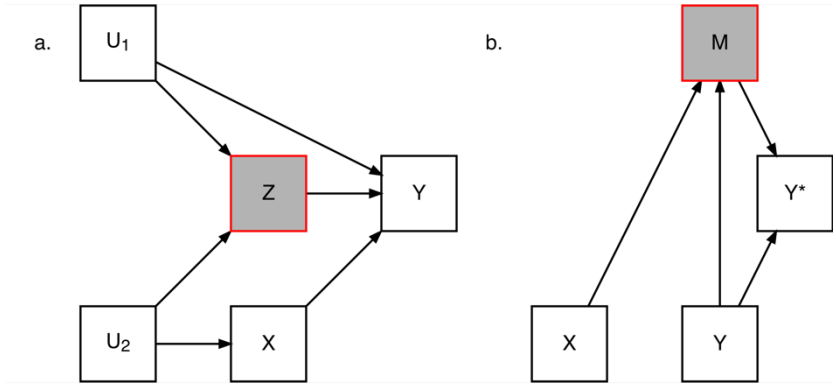
Statistical control is an example of a 'non-equivalent groups design': the analyst starts with groups expected to differ systematically but attempts to model the effect of these differences on the outcome in order to remove them. Nick Huntington-Klein (2022, p. 469) evocatively likens methods like this to creating a statue by taking a large block of marble and chipping away until the desired form is revealed underneath. Where successful – i.e., using a suitably accurate model, possibly reflecting an appropriate DAG – a valid causal estimate can be obtained. However, validity cannot be ascertained directly. Unlike a sculptor, the data analyst cannot say whether the finished product looks quite right. They do not know if they are left with Michelangelo's *David* or Emanuel Santos' bust of Cristiano Ronaldo.

Not including a sufficient adjustment set in the statistical model leaves open backdoor or other biasing paths between exposure and outcome variables and will generate a biased estimate of the causal effect. The size of the bias will equal the (aggregate) strength of the biasing paths. For confounding paths, it will reflect the strength of the associations between the confounders and the exposure and between the confounders and the outcome. Adding too many covariates – the Causal Salad 'approach' (McElreath, 2020) – can also bias estimates where covariates serve as colliders or mediators of the effect of the exposure upon the outcome (the latter where an estimate of the total causal effect is sought). The statistical adjustment approach will also obtain a biased estimate if confounders are not modelled appropriately or measured accurately. Simple models with variables incorporated as linear terms without interactions or other non-linearities (e.g., absence of required polynomial terms) may not reveal suitable residual variation.

Theory may not be sufficiently strong to clarify the correct functional form *a priori*, which limits standard regression algorithms like OLS, where the functional form requires explicit parameterisation. On the other hand, machine-learning and deep-learning algorithms, such as random forests and neural networks, are capable of modelling non-linearities without their express inclusion in a model. The Frisch-Waugh-Lovell theorem has motivated the development of other methods, notably the double machine learning method (Chernozhukov et al., 2018), where machine-learning algorithms are used to obtain residuals for exposures and outcomes, increasing the

plausibility that data-generating processes have been modelled correctly. Still, ever more complicated algorithms cannot use data that aren't there; bias via backdoor paths (and incorrectly blocked causal paths or conditioned colliders) is still a fundamental issue regardless of the method used for statistical control.

Figure 5.1: DAGs showing situations where controlling for a variable may increase or decrease bias. In (a), conditioning on Z closes the path $X \leftarrow U_2 \rightarrow Z \rightarrow Y$ but opens the path $X \leftarrow U_2 \rightarrow Z \leftarrow U_1 \rightarrow Y$ (Source: Cinelli, Forney, et al., 2024). (b) represents a situation where there are mode effects and mode selection (Section 2.2.5). Conditioning on M (mode) closes the path $X \rightarrow M \rightarrow Y^*$ representing differential measurement but opens the path $X \rightarrow M \leftarrow Y \rightarrow Y^*$. Tomova et al. (2026b) use simulations to explore the conditions in which this would increase bias.



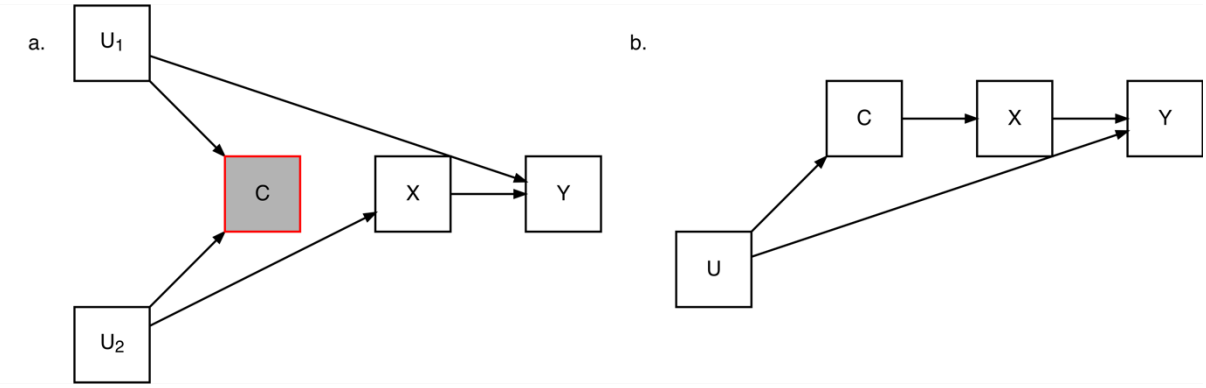
Though non-parametric in nature, DAGs can help identify control variables that are sufficient to identify causal effects: the software DAGitty (2025) can be used to programmatically identify minimal sufficient adjustment sets or biasing pathways, while Cinelli et al. (2024) provide a "crash course" with example DAGs showing where adjustment would reduce, increase or leave bias unchanged.[14] There are also situations where controlling for a variable could go either way; Figure 5.1 shows two DAGs where control closes one path but opens another due to the variable operating as a collider ('damned if you do, damned if you don't'; Cinelli, Forney, et al., 2024; Tomova, Silverwood, Tennant, et al., 2026; Woolf et al., 2023). Whether bias would be reduced is an empirical question and depends on the relative strength of the paths in question; judgements about this benefit from domain expertise. Regardless, adjustment sets will

[14] Control for some variables does not reduce bias, but by reducing residual variation, increases statistical power and the precision of estimates.

only be valid in so far as the assumptions encoded in the DAG (e.g., the set of variables, causal arrows, and their direction) are valid, too. Though some assumptions in a DAG may be testable empirically (i.e., implied [conditional] independencies), ultimately not all can be – e.g., there is no statistical way to identify the correct direction of an arrow between two adjacent variables, measured at the same moment in time.

VanderWeele (2019) offers principles for selecting confounders when knowledge of underlying causal structures is incomplete or uncertain. Specifically, he recommends, with two caveats, conditioning upon any pre-exposure covariate that is a cause of the exposure or the outcome or both. This 'disjunctive cause criterion' avoids issues that can arise if simply controlling for any pre-treatment variable (i.e., M-bias) or insisting confounders must be common causes of both exposure and outcome (which can miss variables that are valid controls); see Figure 5.2 below.

Figure 5.2: DAGs showing bias that can arise when (a) controlling for any pre-treatment variable or (b) controlling for a variable only if it is available and is a common cause of both exposure and outcome. In (a) and (b) variables U are not observed. In (a), conditioning upon C opens a biasing path from X to Y via C ($X \leftarrow U_2 \rightarrow C \leftarrow U_1 \rightarrow Y$; i.e., M-bias). In (b), not conditioning upon C keeps the biasing path $X \leftarrow C \leftarrow U \rightarrow Y$ open, though it could be closed.



VanderWeele's first caveat to this rule is that analysts should not control for any instrumental variable (Section 5.9) – that is, a variable which is only related to the outcome via its effect on the exposure – as this can generate 'Z-bias' where the bias from an unconditioned common cause of exposure and outcome is amplified by adjustment for an instrument (Ding et al., 2017). This is because controlling for the instrument removes a source of variation in the exposure that was unbiased, meaning a

greater proportion of the variation that is left will represent bias or measurement error.[15] VanderWeele's second caveat is that analysts should also consider conditioning on 'proxy confounders' – variables that explain sufficient variation in true confounders, such that they can be used instead of them, for example when the confounders themselves are not available or are poorly measured. These proxy confounders can be thought of as imperfect measurements of the true confounders. However, a drawback of including these is that there may also exist an unobserved common cause of the proxy confounder and the outcome, which could then create M-bias if the proxy confounder is conditioned upon. This would be the case if C were considered a proxy confounder for U_2 in Figure 5.2a above.

VanderWeele (2019) also discusses issues that can arise when covariates and exposures are measured simultaneously as it is possible that covariates are in fact mediators rather than confounders – for instance, conditioning upon cognitive ability when trying to estimate the causal effect of educational attainment on outcomes as this is a consequence of education, as well as a cause (Ritchie & Tucker-Drob, 2018; Strenze, 2007). A key benefit of CLS's cohorts is that multiple sweeps of data are available which include repeated measurement, so it is usually possible to find measurements for confounders from prior to the measurement of the exposure (as well as to find measurements of the exposure prior to the outcome), thereby respecting temporality. Additionally, it may also be possible to control for lagged measurements of exposures and/or outcomes (VanderWeele et al., 2020), which can again reduce bias, particularly where time-invariant confounders exist or there are confounders that have effects on exposure and outcome at multiple time points. Note, controlling for earlier exposures more specifically links the study estimate (exposure to treatment at a specific time point) to a specific intervention and to the randomised controlled trial design; we discuss this further in Section 5.11. However, one issue is that the time between adjacent sweeps can be quite long (e.g., a nine-year gap between the Age 33

[15] This is also known as 'bias amplification'. Bias amplification can occur even if the confounder is a weak cause of the outcome because, again, once conditioned upon, the residual variation in the exposure may be more confounded than before. Similar logic also applies to sibship studies and other fixed effects designs as discussed in Sections 5.5 and 5.14.

year and 42 year NCDS sweeps), so control may be imperfect when this strategy is used.

Less abstractly, VanderWeele et al. (2020) provide a list of variables they suggest should be considered for inclusion as covariates in regression models, given these represent confounders of many different associations in the health and social sciences.[16] Ideally these should be measured prior to exposure and outcome, or otherwise with consideration of the possibility that they mediate rather than confound associations (see Table 5.1). Most or all of these variables are available in each of CLS's cohorts.

Table 5.1: List of confounders to include in regression models recommended by VanderWeele et al. (2020).

Domain	Covariate
Demographic	Race (Ethnicity) Age Gender Marital Status
Economic, Social and Political	Income Education Employment Social integration Neighbourhood Religious service attendance Political affiliation
Health	Self-rated health Number of health conditions Exercise Smoking Alcohol consumption Depression
Psychological	Happiness Loneliness Parental warmth Purpose/Meaning Big five personality

[16] In Section 8.2 (Appendix), we also offer some general theoretical perspectives on confounding processes.

5.1.2 Applications

There are thousands of papers using CLS data that have used the statistical adjustment approach. A particularly rigorous example of statistical control is Purba et al. (2025), who use data from MCS to assess the effect of time spent on social media on cigarette and e-cigarette use. The authors based their (pre-registered) analysis upon a DAG developed from subject knowledge, existing evidence, and input from a policy advisory group, from which they identified a minimally sufficient adjustment set for their analysis. A further virtue, though not related to reducing bias via regression adjustment, is that the authors also elected to use a more reliable time-diary-based measure of social media use rather than other available measures, reducing measurement error.

5.1.3 Opportunities

The regression adjustment approach is straightforward to implement and particularly suited to CLS's cohort data given the amount and variety of data available on participants, including their individual characteristics and the characteristics of their backgrounds, families and the localities in which they live. Control can be made for many common and important confounders, such as sex, ethnicity, psychological traits, and lifetime socioeconomic position. While not all confounders may be available, nor all confounders known, the ability to incorporate additional plausible confounders may still improve upon existing work; bias may not be fully eliminated but at least reduced. The statistical adjustment approach may also be beneficial when treated as a data point within a broader literature – where remaining bias is unrelated to biases hampering other studies, particularly those using different study designs, consistency of results across studies ("triangulation") makes a causal interpretation more plausible (Lawlor et al., 2016). Additionally, as will be discussed further in Section 5.20 on Quantitative Bias Analysis, information on important confounders can be used to calibrate expectations about whether remaining confounders are sufficient to explain observed associations.

5.2 Matching

5.2.1 Method

Section 2.2.3 showed that confounding arises where treated and untreated individuals differ on background characteristics that also influence the outcome variable, with the consequence that a naïve comparison of outcomes would mix the effect of treatment with the effect of these pre-existing differences. Matching – a non-equivalent groups design – addresses confounding by identifying sets of treated and untreated individuals who are similar on observed confounders. If these capture all the relevant differences between the groups, the only systematic difference within matched sets is treatment itself, and any difference in outcomes can be attributed to it. Like regression adjustment (Section 5.1), this relies on confounders being observed and measured well – matching on observed confounders does not remove systematic differences in unobserved confounders.

There are multiple matching algorithms, which differ in how 'similar' is defined. Many attempt to identify, for each treated individual, untreated individuals who are as similar as possible on confounders (or vice versa), with outcomes then compared within matched sets. *Exact Matching* identifies individuals who have the same values for a set of characteristics. However, as the number of possible combinations of these characteristics explodes exponentially as more are added (the *curse of dimensionality*), exact matching is only feasible for a small number of characteristics with few unique values or, alternatively, with sample sizes that are very large (e.g., as in whole population administrative data). *Coarsened Exact Matching* makes matching more feasible by coarsening the variables used to match upon – e.g., matching within age groups, rather than by integer age. But again, the curse of dimensionality applies – with just four variables, each taking four values, there are 256 (4^4) possible strata in the data. Matching can instead be approximate and utilise statistical measures of 'distance' that aggregate difference on many dimensions simultaneously. For instance, individuals may be matched upon Mahalanobis distance, a multivariate analogue of Cohen's *d*. Though matching may not be exact, differences between treatment and control are made smaller than they otherwise would be.

Another alternative metric is the propensity score (Rosenbaum & Rubin, 1983). In propensity score matching (PSM), individuals in treatment and comparison groups are matched upon their probability of having been treated, with this probability typically estimated (e.g., using a regression model) based on a set of predictors of treatment. Importantly, this probability is a single scalar value (ranging 0-1) making matching more straightforward. However, as there may be numerous profiles of covariates that yield the same probability of being treated, pairs of individuals matched upon propensity scores do not necessarily share similar values on underlying covariates – balance is instead a property of the overall matched sample rather than of each matched pair. Yet, because covariates can still differ within propensity-matched pairs, pruning poorly matched pairs beyond the point at which the groups are approximately balanced can *increase* imbalance (King & Nielsen, 2019), though the extent to which this matters in practice has been questioned (Ripollone et al., 2018). Matching on the propensity score can also reduce statistical efficiency compared to more direct matching methods. (Note, the issue of imbalance applies to using the propensity score to identify matches, rather than for inverse probability weighting, which is discussed in Section 5.3.)

Matching does not have to be 1:1 – several matches can be obtained for each 'treated' individual and vice versa. Matches can also be made with *replacement*, so that one individual may be matched to multiple others. The advantage of matching with replacement is that it can reduce bias as the best match is used in each case. However, it also decreases statistical power as the effective sample size is reduced – in other words, there is a bias-variance trade-off (Huntington-Klein, 2022). Conversely, using several matches per individual increases statistical power but also increases bias as more data than just the very best match are used.[17] Bias can be reduced by using a *caliper* – i.e., setting a maximum dissimilarity between matches – and by using kernel weighting, where matches are weighted more highly the better they are. An alternative is entropy balancing (Hainmueller, 2012), which instead of identifying similar individuals identifies weights to create matched treatment and control groups that are overall

[17] That said, many-to-one or one-to-many matching can also guard against overfitting.

balanced for specified moments of the distributions of the matching variables (e.g., same means, variances and skewness for covariates).

As with statistical adjustment, the validity of causal estimates obtained via matching depends on whether, within matched groups, there is no remaining confounding. Unlike statistical adjustment, matching can be thought of as a data preprocessing step rather than an estimation procedure, *per se* (Ho et al., 2007) – matching is a way to identify an appropriate analytical sample before substantive analyses are undertaken. Matching can therefore be combined with regression – e.g., by estimating a regression model within a group of matched individuals. This two-step approach has the advantage that it can be "doubly robust" (Stuart, 2010) – robust to the failures of one method, if the other is valid. For instance, a valid causal estimate may be obtained with a mis-specified regression model, if matching was sufficient for achieving non-confoundedness, or vice versa. Note, where untreated individuals are matched to each treated individual, matching yields the Average Treatment Effect on the Treated (ATT). This may not be equal to the ATE if there is effect heterogeneity.[18]

Matching has several advantages versus regression adjustment. First, matching itself does not require the outcome to be modelled as the method focuses on approximating the data-generating process for *treatment* specifically. This is useful where it is difficult to model the outcome (e.g., where the outcome is rare) or where there are multiple outcome variables – i.e., when performing an outcome-wide epidemiological analysis (VanderWeele, 2017; VanderWeele et al., 2020). One drawback, however, is that including instruments in the treatment model can generate Z-bias (Section 5.1) if unobserved confounders remain.

A second advantage of matching is that it can ensure comparisons between treatment and control groups are made within regions of 'common support' (overlap in covariate

[18] Where treated individuals are matched to each untreated individual, the procedure will yield the Average Treatment Effect on the Untreated (ATU), which again may differ from the ATE. Combining the two (weighted) can produce the ATE. Which estimand – ATE, ATT, ATU – is relevant and most useful depends on the research application. For a policy to be expanded across the population, the ATU may be of most interest. For a policy to be rolled back, it may be the ATT.

distributions), whereas regression approaches, which interpolate between data points, do not. Some combinations of characteristics may not appear in both treated and control groups (i.e., positivity may be violated) – this is at least explicit in matching but hidden when performing regression. While interpolation may be reasonable in a given case, its validity cannot be ascertained with the data at hand. Interpolation is also more plausible if matching has been used as a data preprocessing step. Relatedly, the assumptions of propensity score matching are partly testable as balance can be assessed within strata of the propensity score, though of course this does not tell the analyst anything about confounders that are unobserved.

Matching does have several disadvantages, however. First, if matches cannot be obtained for all observations, some units may need to be discarded, potentially yielding a valid causal estimate, but within a sample with no clear interpretation or wider applicability. Second, matching creates challenges where exposures are continuous rather than binary, though extensions, such as kernel matching, have been developed.

5.2.2 Applications

Persson (2014) uses data from BCS70 to assess whether the strong association between higher education and political behaviour (e.g., voting in general elections) reported in prior studies is causal. Unlike data used in previous studies, the BCS70 allowed the author to match upon important confounders measured prospectively from childhood, such as parental behaviour and cognitive ability at ages 5 and 10 years. The author used genetic matching – an algorithm that identifies weights for a generalised Mahalanobis distance matrix minimising imbalance between treatment and control groups (Diamond & Sekhon, 2013) – observing that the procedure performs better than coarsened exact matching and propensity score matching in the BCS70 sample (due to larger retained samples and better balance, respectively). He finds little evidence of association after matching is performed. In support of the matching, the author reports multiple tests of balance, including a comparison of the imbalance remaining against that expected under random assignment and placebo tests using pre-treatment variables that higher education could not have caused. A caveat is that only 569 cohort

members were treated and confidence intervals are not reported – the analysis indicates no detectable effect rather than the absence of one.

In two papers, Jerrim & Sims (2019, 2020) use data from the MCS to assess the causal effect of grammar (i.e., selective) school attendance upon an array of non-cognitive factors, such as wellbeing and educational aspirations. As with Persson (2014), the authors exploit the rich life course information on confounding variables (e.g., prior measures of cognitive ability and wellbeing), though do not observe entry test results, the main basis upon which entry is determined. In both papers, the authors match upon propensity scores, but use two separate groups from which to draw untreated observations. In Jerrim & Sims (2019), the authors use geographic variation across Local Education Authority (LEA) areas, comparing those attending grammar schools with matched cohort members living in LEAs with no selective schools at all. In Jerrim & Sims (2020), the authors compare individuals who attended grammar school with those who applied but ultimately did not attend.[19] They find little evidence of association with non-cognitive outcomes overall, except some evidence of an effect on educational aspirations (Jerrim & Sims, 2019). One limitation of their papers is a lack of discussion of what unobserved confounders may be present, especially for the analysis using grammar school applicants for the control group. A second is precision: the 2019 estimates draw on 4,785 cohort members, whereas the 2020 analysis matches roughly 650 in England and 500 in Northern Ireland. Several estimates are large enough to be substantively meaningful but have wide confidence intervals.

5.2.3 Opportunities

As with statistical adjustment (Section 5.1), there are ample opportunities in CLS's cohorts to apply matching methods. An additional advantageous property of CLS's cohorts is their large sample size – matching can lead to a large loss of observations if few or no comparison observations can be found for a treated observation. Though not a design that matches on *treatment*, CLS's cohorts can also provide general population

[19] The authors note that individuals usually apply for grammar school after they have received entry test results, so applicants, regardless of whether they ultimately attend or not, are likely to have passed the entry test.

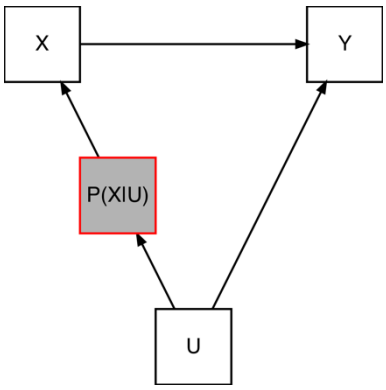
controls in case-control designs – for instance, the MCS provided a control group as part of the National Evaluation of Sure Start (Joshi & Fitzsimons, 2016) and the NCDS is part of the Wellcome Trust Case Control Consortium, which was used in many of the first molecular genetic studies of health and disease (Wellcome Trust Case Control Consortium, 2007).

5.3 Inverse Probability of Treatment Weighting

5.3.1 Method

Section 5.2 discussed issues with using propensity scores to match treatment and control groups. Regardless of this problem, when used as a covariate or to weight observations, propensity scores can recover causal effects: the propensity score theorem (Rosenbaum & Rubin, 1983) shows that if treatment is independent of potential outcomes conditional on covariates, U, then treatment will also be independent of potential outcomes conditional on the propensity score (the probability of treatment conditional on the same covariates). Intuitively, where a binary exposure is influenced non-deterministically, confounders can be thought to influence the *probability* of an exposure occurring, rather than the exposure itself. Conditioning upon the propensity score thus closes a backdoor path. See Figure 5.3 for a DAG.

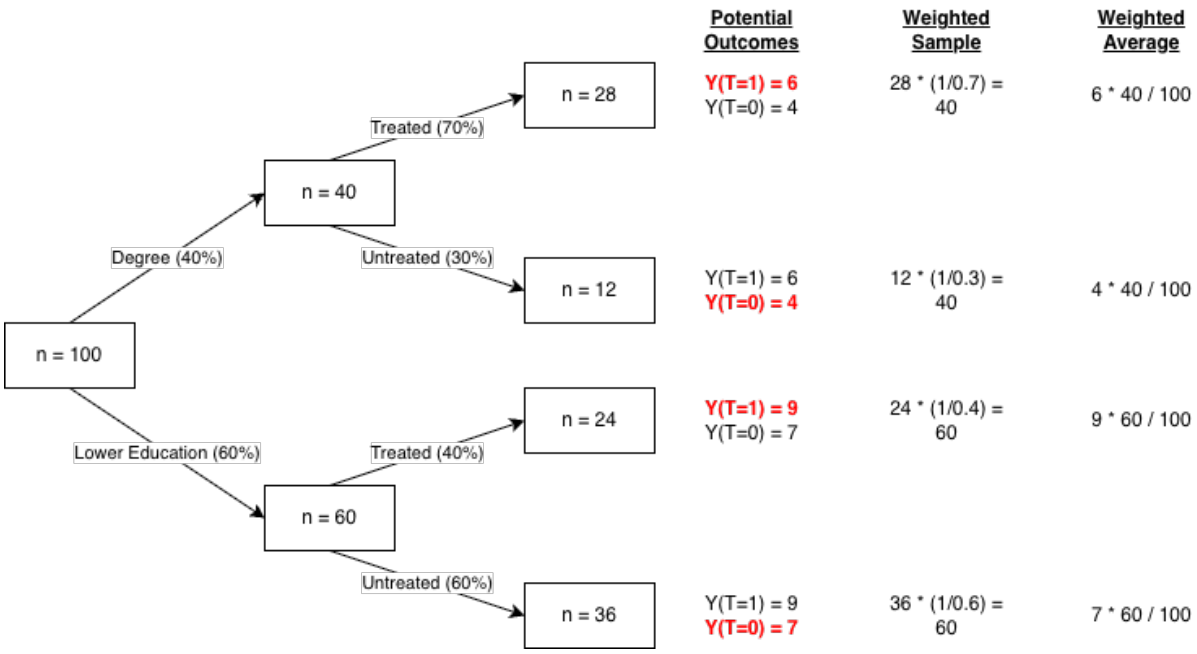
Figure 5.3: DAG representing conditioning on the propensity score (adapted from Cunningham, 2021).



The typical way of using a propensity score is as a weight in Inverse Probability of Treatment Weighting (IPTW). In IPTW, treated observations are weighted by the inverse of their probability of treatment and untreated units are weighted by the inverse of their probability of non-treatment (i.e., one minus the propensity score). The upshot is that,

when implemented successfully, data are reweighted to create a pseudo-population in which treatment assignment is independent of the measured confounders, thereby mimicking a feature of randomisation. The schematic in Figure 5.4 displays the procedure where there is a single confounder, education, that is associated with both the probability of treatment and potential outcomes.

Figure 5.4: A representation of the IPW method with a single confounder, education level.



In Figure 5.4, a naïve comparison of treated and untreated groups would yield a downwards-biased estimate due to the disproportionate number of treated individuals who are degree-educated and who have lower potential outcomes than those with less education – i.e., the naïve comparison is 1.13 = (6*28 + 9*24)/52 – (4*12 + 7*36)/48, instead of the correct causal effect, which is 2 (compare the potential outcomes Y[T = 1] and Y[T = 0] in each group). Using IPTW recovers the true causal effect as, when weighted, both treated and untreated groups contribute the initial sample size of the groups they were drawn from (i.e., n = 40 for those with a degree, n = 60 for those without): 2 = (6*28*1/0.7 + 9*24*1/0.4)/100 - (4*12*1/0.3 + 7*36*1/0.6)/100. This example generalises to scenarios where the propensity score is based on more than one covariate.

Inverse probability weighting (IPW) can also be used to account for bias from non-response, where responding units are weighted by the inverse of their probability of response, so that they represent similar non-responding units. We discuss IPW for missing data further in Section 5.4. Note, as with other 'non-equivalent groups design' methods, IPTW will provide unbiased estimates only in so far as control variables are sufficient to close backdoor confounding paths. Another issue that can arise with IPTW is where there are extreme weights, most typically where some treated units have very low probabilities of treatment and/or control units have very high probabilities of treatment. In this case, trimming is often used, either by winsorising (capping) particularly large weights or by dropping observations with extreme weights or propensity scores. As with matching, discarding observations or winsorising weights can, in some cases, shift the population for whom the effect is being estimated, by focusing on those within the region of common support. Typically, this is an unintended side effect of trimming, in which case the resulting estimate may no longer answer the original research question and may be challenging to interpret. That said, the effect in the overlap region can sometimes answer policy-relevant questions and therefore may still be of interest (e.g., if one does not expect treatment effect heterogeneity).

5.3.2 Applications

Madia et al. (2022) use data from Next Steps to estimate the causal effect of school exclusion on labour market outcomes. They use several analytical approaches to address different sources of confounding, including school fixed effects (Section 5.5), matching (Section 5.2) and IPTW, with IPTW used to address statistical power issues with matching given the small number of treated individuals (i.e., persons permanently excluded from schools; N = 86). IPTW estimates show sizeable effects upon several outcomes, including higher risks of being not in employment, education, or training (NEET) or unemployed, specifically. Estimates are similar in size to those obtained using propensity score matching, but are more statistically precise as no observations are discarded.

The authors condition on multiple factors, exploiting Next Steps' rich data collection on individual characteristics and the family environment, and apply non-response weights

alongside the IPTW weights to account for attrition. However, they are unable to control for prior educational attainment or detailed measures of cognitive and non-cognitive skills, which plausibly confound associations. The authors pay attention to the *consistency* assumption, comparing individuals who have been excluded with those who have never been suspended or excluded (though, an alternative, policy-relevant option could have been to compare excluded individuals with those only suspended). Finally, the authors use quantitative bias analysis (QBA) methods (Mantel-Haenszel and Rosenbaum bounds) to assess sensitivity to unobserved confounding. Two omissions are that no diagnostics for the weights are reported and the plausibility of the estimates given the QBA results is not discussed.

5.3.3 Opportunities

Similar to matching and statistical adjustment, there are ample opportunities in CLS's cohorts to apply IPTW methods given the number of plausible confounders that can be included in IPTW models. Additionally, the longitudinal nature of the cohorts enables the creation of plausible non-response weights, as discussed in the next section.

5.4 Missing Data Methods

5.4.1 Method

In Section 2.2.4, we discussed how large-scale intensive and longitudinal data collections, characteristic of CLS's cohorts, can contain a sizeable amount of missing data through unit, sweep, and item non-response. Where non-response is not random, missingness can lead to selection bias, impacting internal (and even external) validity (Type 1 and Type 2 selection bias; Lu et al., 2024). For instance, if individuals with poorer mental health are more likely to drop out from a cohort, associations between factors that cause mental ill-health and measurements of mental health may be attenuated in available (complete case) data.

There are two main methods for addressing missingness: inverse probability weighting (IPW) and multiple imputation (MI). IPW in this setting involves two steps. First, the probability of response is calculated or estimated, typically with a regression model using data available on participant characteristics for both respondents and non-

respondents (in longitudinal data, this would be data collected prior to an individual attriting). Second, the inverse of this probability is taken and used to weight (non-missing) observations in substantive analyses. By taking the inverse, individuals with lower probability of response are given greater weight – in effect, they are used to represent more individuals whose data are missing. For instance, an individual with only a 10% predicted probability of response will be weighted ten times as highly as someone certain to respond as the former is used to reflect both themselves and nine similar individuals with the same (10%) probability of response who did not respond.

MI is typically performed using chained equations: in this procedure, missing data are imputed using a series of regression models where the statistical relationships between observed values are used to predict values that are missing. For instance, in longitudinal data where some participants drop out, BMI collected during childhood could be used to predict adulthood BMI by applying the association detected in individuals for whom BMI was observed on both occasions. These imputed values are then used to predict and impute missing data in other variables. As predictions are unlikely to be perfect, multiple imputed datasets are created to reflect imputation uncertainty. Substantive analyses are then performed in each imputed dataset and results pooled accounting for the variation both within and between imputations – typically, pooling is done by applying 'Rubin's Rules' (Rubin, 1987), though bootstrapping is also possible (Bartlett & Hughes, 2020; Schomaker & Heumann, 2018). Note, MI (and IPW for non-response) are not methods for estimating causal effects in themselves; rather, they address a source of bias in downwards analyses.

In MI, imputation and substantive models should be 'congenial' (van Buuren, 2018); that is, the imputation model should be no more restrictive than the substantive model, accommodating every variable and functional form the substantive model contains – i.e., outcome, exposure, covariates, and interactions or nonlinear terms. Otherwise,

excluded terms are implicitly set to zero in the imputation models and are not directly reflected in imputed values, attenuating associations in downstream analyses.[20]

MI and IPW rely on assumptions about the mechanism by which data became missing. These mechanisms can be classified into Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). To illustrate each, we take the example of using partially observed data to obtain an unbiased estimate of average (mean) income:

- MCAR: Missingness is unrelated to observed or unobserved values.
  - Example: Income data for a random subset of participants are lost due to hard disk failure.
- MAR: Missingness is unrelated to unobserved values, conditional on observed data.
  - Example: Participants in jobs that involve frequent travel are more likely to miss a survey interview, and these jobs may pay more on average. However, among those in such jobs, who is away – and so whose data are missing – is unrelated to income.
- MNAR: Missingness is related to unobserved values, even conditional on observed data.
  - Example I: Participants whose income derives from illicit sources (and who may have higher income, on average) decline to answer questions on income, so whether income is reported is related to income itself, even among participants who otherwise look alike on everything observed (e.g., education level).
  - Example II: Due to social desirability concerns, people with low income refuse to answer questions on income.

Note, in MNAR Example I, missingness was due to an unobserved variable, rather than income itself, whereas in Example II, it is income itself that is the cause of missingness.

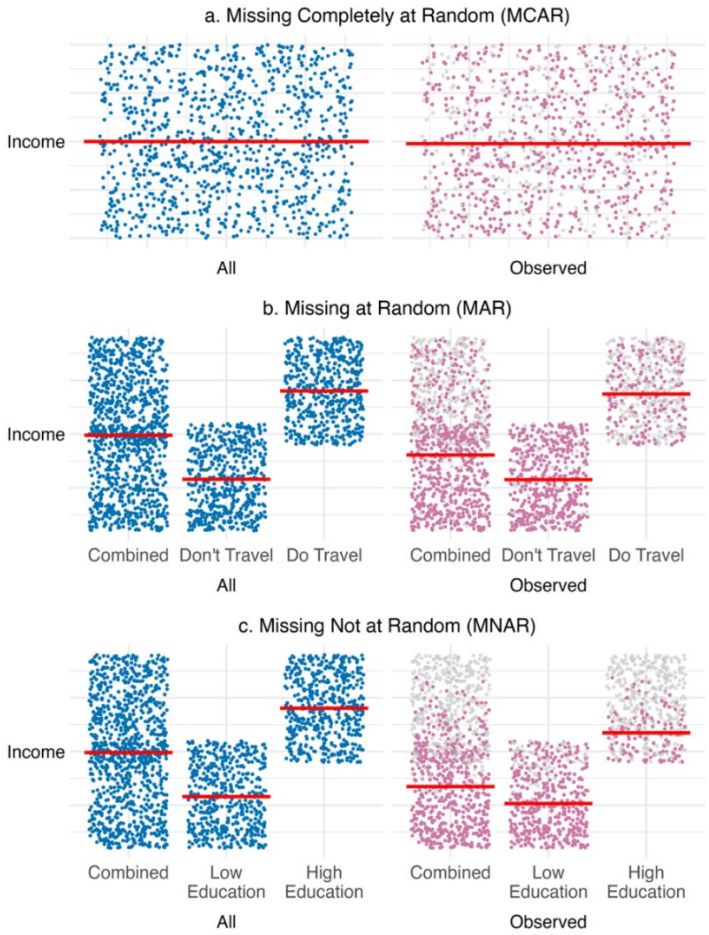
[20] Congeniality is sometimes instead referred to as compatibility, though compatibility has another definition in MI that is distinct (van Buuren, 2018).

However, both are still MNAR processes, though if we had data on whether a participant derived income from illicit sources, the former example would be MAR.

MCAR, MAR and MNAR processes all reduce statistical power because each leads to a loss of information. However, they have different consequences for biasMCAR processes do not generate bias. MAR processes bias complete-case analyses, unless the variables related to missingness are conditioned upon in the substantive model or MI or IPW is used with imputation or response models including those variables appropriately. MNAR processes also generate bias, and neither standard MI nor IPW can address this. However, this is partly a question of degree: in longitudinal data, the plausibility of the MAR assumption can be increased using lagged measurements and other rich information from over the life course (e.g., data on early-life cognitive ability, socioeconomic background, prior measures of income, etc.; Mostafa et al., 2021; Silverwood, Narayanan, et al., 2024); the extent of violation of the MAR assumption can at least be reduced, if not fully eliminated. Administrative linkages – which can continue even if the participant stops responding to the main survey – can also help here (Rajah et al., 2023). In Section 5.20, we describe a variant of MI – Pattern Mixture Modelling – that can account for MNAR processes.

Figure 5.5 below show MCAR, MAR and MNAR processes for a simulated variable, income. Left-hand plots show all data points; right-hand plots show only the observed points, with missing points in light grey. Red lines show mean income, the target quantity we use the observed data to estimate.

Figure 5.5: Simulations displaying MCAR, MAR and MNAR processes in a variable, income. In row (a), data are MCAR. Missingness is unrelated to income, so mean income among the observed data is an unbiased estimate of the true mean. In row (b), data are MAR. People whose jobs involve travel earn more and are more likely to have missing income, so mean income among the observed data is too low. Within each travel group the observed mean is unbiased, so conditioning on travel, or using it to impute missing incomes, removes the bias. In row (c), data are MNAR.People with higher incomes are less likely to report them, so observed mean income is too low. Conditioning on a predictor of income such as education reduces the bias but does not remove it.



An advantage of MI over IPW for handling missingness is that, by retaining partially observed cases rather than discarding them, MI uses more of the available data and can therefore yield more precise estimates. Nonetheless, higher missingness entails lower statistical power and potentially greater sensitivity to modelling choices in either method. In IPW, low probabilities of response can generate large weights for some observations, which inflates the variance of estimates and increases dependence on the response model being correctly specified. In MI, greater missingness increases both

the contribution of imputed values to estimates in substantive models and variation between imputed datasets as imputation models are less precisely estimated. Predictive accuracy can be increased by including 'auxiliary' variables in imputation or response models – additional variables that are predictive of missing values, but which do not appear in substantive models.

MI and IPW are not mutually exclusive and can be used alongside each other. For instance, it can be useful to use IPW to account for sweep non-response, and within a responding sample, use MI to address item missingness (see, for example, L. Wright, Davies, et al., 2024). It is also possible to use MI to impute missing values in variables to be used to create IPW weights. There are methods for combining IPW and MI in ways that are 'doubly robust'; see Seaman & Vansteelandt (2018) for more details. IPW can also be performed combining weights for selection (i.e., response) and for treatment (IPTW, Section 5.3).

Note, the potential outcomes framework conceptualises causal inference as an inherent missing data problem (Parra et al., 2021); counterfactuals are not observed and are therefore unknown. Some newer methods approach estimating causal effects as a missing data issue (e.g., Matrix Completion; Athey et al., 2021); given these methods are more advanced and niche we do not discuss these further here.

5.4.2 Applications

Wright et al. (2024) use a combination of IPW and MI to estimate the association between a polygenic score (a summary measure of genetic predisposition to a trait) and observed BMI over the life course in four different cohorts (including the NCDS, BCS70 and MCS). As genetic data were only collected in later sweeps of each cohort (44y, 46y, and 14y for the NCDS, BCS70 and MCS, respectively), there was a high amount of missing data, including for BMI at later ages. The authors therefore used IPW to weight observations in the genotyped sample. MI was then used to impute remaining item missingness; earlier observations of BMI were used to impute later values of BMI and several auxiliary variables predictive of missingness (Mostafa et al., 2021) were also included in imputation models. Code used in the study is available here: https://osf.io/nxusz.

5.4.3 Opportunities

There are many opportunities to apply IPW and MI in CLS data. The longitudinal nature of data collection means there is generally a lot of information available with which to create non-response weights or to incorporate into imputation models. This includes earlier measurements of variables that may predict non-response (e.g., measures of mental health earlier in life [including in adulthood]). Several papers show that this information can increase the plausibility of the MAR assumption, as evidenced by the ability to recover initial characteristics in imputed data (Katsoulis et al., 2024; Mostafa et al., 2021). These papers also give useful information on which factors predict non-response that can be incorporated into analysts' own models. CLS has additionally released an accessible and extensive user guide on Handling Missing Data, which readers may find helpful (Silverwood, Narayanan, et al., 2024).

5.5 (Longitudinal) Fixed Effects

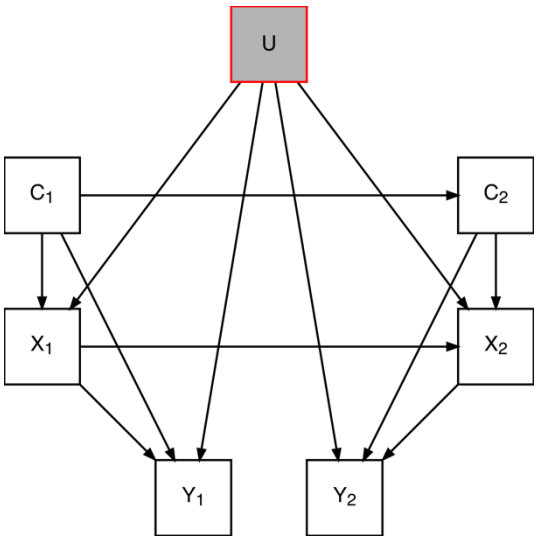
5.5.1 Method

Each of CLS's cohorts benefits from the tracking of the same individuals over the life course. Section 4.3 described four advantages of longitudinal data for estimating causal effects: more accurate measurement, ability to establish temporal ordering, availability of baseline and lagged measures of exposures and outcomes, and the ability to use within- rather than between-person variation. This section focuses on methods that exploit within-unit variation.

Some confounders vary over time; others are time-invariant. Income, sex and country of birth may confound an association of interest, but income varies throughout life, while sex and birth country are fixed attributes of an individual. Utilising within-person variation fixes time-invariant confounders, accounting for their contribution to outcomes *by design* (Huntington-Klein, 2022); see Figure 5.6 for a DAG. This is the idea behind the longitudinal *fixed effects* method. The approach can be understood as a form of matching (Section 5.2), in which each observation is matched not to an observation from a similar individual but to other observations from the *same*

individual, so that time-invariant characteristics are balanced exactly, whether or not they have been measured.

Figure 5.6: DAG representing the fixed effects method with longitudinal data. X and Y are measured at two time points and their contemporaneous association is confounded by a time-variant factor, $C_{t=x}$, and a time-invariant factor, U. Using within-variation effectively controls for the person, implicitly conditioning on all time-invariant factors having constant effects over time.



The equations below show the typical fixed effects set-up when used with longitudinal data. *i* and *t* index individuals and time periods, respectively. The outcome, Y_{it}, is related to an exposure of interest (D_{it}), time-variant confounders (X_{it}), and time-invariant confounders that can be aggregated into a single person-specific *fixed effect*, u_i. Person-specific means for each variable are indicated with the 'bar' accent (e.g., $\overline{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$).

$$Y_{it} = \beta_0 + \beta_1 \cdot D_{it} + \beta_K \cdot X_{it} + u_i + \varepsilon_{it} \qquad [\varepsilon_{it} \sim N(0,\ \sigma^2)]$$

$$\overline{Y_i} = \beta_0 + \beta_1 \cdot \overline{D_i} + \beta_K \cdot \overline{X_i} + \overline{u_i} \qquad [\overline{u_i} = u_i; \overline{\varepsilon_i} = 0]$$

$$Y_{it} - \overline{Y_i} = \beta_1 \cdot (D_{it} - \overline{D_i}) + \beta_K \cdot (X_{it} - \overline{X_i}) + \varepsilon_{it} \qquad [\beta_0 - \beta_0 = 0; \overline{\mu_i} - u_i = 0]$$

In the usual implementation (performed automatically by statistical software), time-invariant confounding is accounted for by calculating person-specific means for each variable and subtracting these from observed values (the 'within' transformation). As time-invariant confounders are (by definition) fixed, the average is the same as the

individual data points. Subtracting person-specific means therefore has the effect of implicitly removing the influence of these time-invariant factors (i.e., as $u_i - \overline{u_i} = 0$).[21]

The advantage of the fixed effects method is that time-invariant confounders are accounted for without needing to measure them or even know what they are. It is not without limitations, however. Importantly, fixed effects do not account for time-varying confounding or reverse causality, which may be important sources of bias in a specific application. Fixed effects also do not account for the time-varying effects of time-invariant factors – fixed characteristics like sex and birth country may not differ across life, but their importance can vary considerably over development. For instance, sex could have larger effects on outcomes post-puberty.[22] This is less of an issue where the time span used is small, but CLS's cohorts span decades, typically with several years between survey sweeps. The large gaps can also mean few observations per individual, so person-specific averages can be overfitted.

Because it uses only within-person variation, the fixed effects method also cannot be used to assess the (time-invariant) effects of time-invariant variables, which may be of substantive interest. Additionally, only individuals whose exposure does in fact vary contribute to the estimated outcome-exposure association. Where treatment effects are heterogeneous, this can raise interpretational issues. Specifically, fixed effects estimates the change in an outcome accompanying a change in exposure between measurement occasions, averaged over individuals whose exposure changes during follow-up, and weighted towards those whose exposure varies most. For instance, fixed effects has been frequently used to examine the effect of unemployment on health (see, e.g., Schmitz, 2011) but this implicitly utilises data from people whose employment status changes (frequently), rather than those with unbroken (un)employment episodes. Moreover, where classical measurement error is an issue,

[21] This is known as the within-estimator. It is also possible to account for u_i by adding person-specific fixed effects (i.e., person dummies) to the regression model.
[22] Within-person differencing does, however, account for the average effect of these time-invariant confounders.

focusing on within-variation makes the problem worse as a larger proportion of within-variation compared with between-variation will reflect noise rather than signal.

There are several other methods that address some of these limitations, which readers can delve into further by following the references. Random effects models capture person-specific time-invariant effects as random variables, but assume the random effects are independent of covariates, which may be unrealistic in practice (and is in fact a reason to use the fixed effects model in the first place). However, specific parameterisations of the random effects model (e.g., the within-between [REWB] model; Bell et al., 2019; Bell & Jones, 2015) can get round this limitation. Methods also exist to account for reverse causality and for serial correlation in the outcome variable (note, the absence of an arrow between $Y_{t=1}$ and $Y_{t=2}$ in Figure 5.6) – specifically, dynamic panel data methods (Wooldridge, 2010) and the random intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015) and its extensions (Mulder & Hamaker, 2021; Zyphur et al., 2019, 2020).

Longitudinal data are a type of hierarchical data – observations are nested within individuals. The fixed effects method can be applied to other types of hierarchical data, too. For instance, Next Steps recruited multiple students from participating schools – using variation within schools allows one to account for school-level confounders (see, e.g., Madia et al., 2022). Another important type of hierarchical data used frequently in the causal inference literature is sibships (Sjölander et al., 2022). Using sibling fixed effects controls for family-level confounders by design (discussed further in Section 5.14). Note, fixing effects at one level also fixes characteristics at higher levels the units are nested within – e.g., looking within schools also controls for Local Education Authority; looking within neighbourhoods also controls for town, region, and so on (Abdellaoui et al., 2022).

To best make use of fixed effects, researchers should consider at which level confounding is expected to occur. To give an (admittedly niche) example, Cinelli & Hazlett (2020), in their analysis of the causal effect on political beliefs of being a victim of violence during the Darfur war, note that once an army had reached a village, violence was likely largely indiscriminate and thus using data from multiple individuals within a village could achieve unconfoundedness. Similarly, Evans (1986) studies the

effect of seatbelts on car crash fatalities using data from crashes where one occupant was belted and the other unbelted. This design controls for many pertinent shared factors (e.g., the speed the car was travelling at the time of crash, the road conditions, etc.), producing estimates that are likely less biased than comparing belted and unbelted individuals overall (Rosenbaum, 2019). The fixed effect design is an expression of what Zubizarreta et al. (2018, p. 16) term 'isolation': "Within a large but biased data system, isolation means constructing or finding a smaller set of comparisons that are likely to be much less biased". Nevertheless, this is situation specific, and it is also possible that within-unit variation is more confounded, not less, than the between-person variation, producing problems akin to Z-bias (discussed in Section 5.1).

5.5.2 Applications

Scott (2022) uses BCS70 data to estimate the causal effect of university education upon political attitudes measured on multiple occasions between ages 16y and 42y. Using two-way fixed effects (Section 5.6) and REWB estimators, he shows substantial within-person change in attitudes after graduation (compared to prior to enrolment). A key strength of the study is an explicit statement on the estimand and its causal aim.

Fitzsimons & Pongiglione (2019) use MCS to assess the effect of maternal employment on child BMI, showing children have higher BMI in periods in which their mothers are employed. Interestingly, no such within-person change is observed for paternal employment, though as the authors note, in the MCS sample, fathers' employment is more stable – thus a different sample contributes to the estimated causal effect. Speyer et al. (2022) use RI-CLPMs to assess the effect of parenting behaviours on children's socioemotional development (as assessed by the Strengths and Difficulties Questionnaire) across early to mid-childhood (3y, 5y, and 7y), allowing for reverse causality. They show that harsh parenting practices (e.g., hitting) are followed by worse SDQ scores, and vice versa. The authors also perform Quantitative Bias Analysis to assess sensitivity to unobserved confounding.

5.5.3 Opportunities

Given the extent of repeated measurement in CLS's cohorts, there are potentially many opportunities to use fixed effects methodology with the data. As discussed, a key issue, however, is the gap in age between data collections: time-invariant confounders may not have time-invariant effects over these periods. This can also be the case even when data collection is more frequent if this occurs over periods of substantial developmental change (e.g., MCS has five data collections to age 11y, but this spans early to mid-childhood and even the onset of puberty). Further, where one is using (e.g.) RI-CLPM and similar models to assess lagged effects, the gap in age may be implausible for the causal effect under study. For instance, one study that used MCS data to assess the effect of social media upon mental health relied upon 3-year gaps between data collections (Kelly et al., 2022). This would not capture shorter-lived effects.

5.6 Difference in Differences

5.6.1 Method

In Section 3.3, we introduced some of the evidence that John Snow marshalled to demonstrate that cholera was transmitted via faecal-infected water rather than miasmas. This was not the only evidence he brought to bear. A further piece of evidence was his "Grand Experiment", comparing cholera cases in households supplied by either of two London water companies, the Southwark-Vauxhall Company and the Lambeth Waterworks Company. Both companies drew water from the River Thames, but in 1852, Lambeth moved its intake pipes upstream of the main location where sewage was discharged (Cunningham, 2021). Cholera cases in households supplied by the two companies in two years, before and after Lambeth relocated its pipes, are shown in Columns 2-3 of Table 5.2.

Table 5.2: John Snow's 'Grand Experiment' comparing cholera incidence (per 100,000) in households supplied by two water companies across two years (Cunningham, 2021). The before–after change in Lambeth (−66), the 1854 difference between companies (−128) and the difference in differences (−78) are the three comparisons discussed in the text.

Water Company	Cholera Incidence			Decomposition		
	1849	1854	Diff.	t_1	t_2	Diff.
Southwark-Vauxhall	**135**	**147**	+12	u_{SV}	$u_{SV} + t_{SV}$	t_{SV}
Lambeth	**85**	**19**	-66	u_L	$u_L + d_L + t_L$	$d_L + t_L$
Diff.	-50	-128	-78	$u_L - u_{SV}$	$u_L + d_L + t_L - u_{SV} - t_{SV}$	$d_L + t_L - t_{SV}$

Multiple comparisons are possible with these data to try to determine the causal effect of Lambeth Waterworks Company's decision to move the pipe up the River Thames (and serendipitously reduce customers' exposure to cholera-infected water; Huntington-Klein, 2022). One can treat Southwark-Vauxhall as a control group and use the data cross-sectionally, comparing rates in 1854 between the two companies (19 - 147 = -128). Alternatively, one can make a simple before-and-after (pre-post) comparison, looking at incidence in Lambeth before and after it moved the pipes (19 - 85 = -66). However, these rely on a number of assumptions, which the decomposition in Columns 5-7 of Table 5.2 clarifies. In those columns, the observed incidence is decomposed into a water company fixed effect (u_{WC}) which contributes regardless of time period, a water-company time effect (trend; t_{WC}) which contributes to the change between 1849 and 1854, regardless of whether the water company's pipes were moved, and a treatment effect (d_{WC}) which affects outcomes if 'treated' – i.e., the water pipes have been moved. d_L – the treatment effect of moving the pipe for Lambeth-supplied households – is the quantity (causal effect) of interest.

Cross-sectionally, the observed difference between Lambeth and Southwark-Vauxhall in 1854 is a function of the treatment effect, but also the time effects and fixed effects ($d_L + u_L + t_L - u_{SV} - t_{SV}$; Column 6, Row 3). Intertemporally, the observed difference is a function of the treatment effect and the time effect – i.e., it is confounded by time differences that would have occurred anyway ($d_L + t_L$; Column 7, Row 2). A DAG representing Snow's Grand Experiment is displayed in Figure 5.7.

Figure 5.7: DAG representing John Snow's Grand Experiment (adapted from Huntington-Klein, 2022).

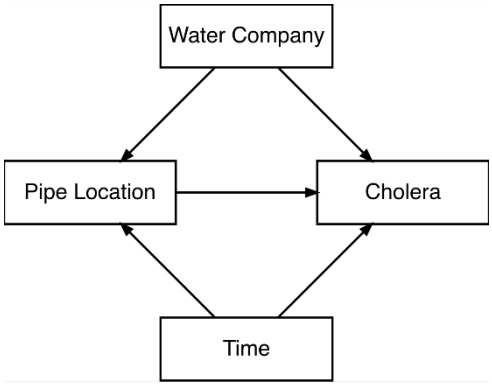


Table 5.2 shows that a further comparison is possible: the difference in differences over time between treated (Lambeth) and untreated (Southwark-Vauxhall) units (-66 - 12 = -78). The decomposition shows that this is a function of the treatment effect and the difference in the time trends in Lambeth and Southwark-Vauxhall ($d_L + t_L - t_{SV}$). This is the (2x2) Difference-in-Differences (DiD) estimator.[23] It is unbiased where the time-trends in treated and untreated units are the same (i.e., $t_L - t_{SV} = 0$); that is, the counterfactual time-trend is the same regardless of treatment assignment. Here, that means Lambeth's cholera incidence in 1854 would have been 12 units higher than its 1849 level – 97 per 100,000 – had it not moved the pipe upriver. This is known as the 'parallel trends assumption'.[24]

The quantity DiD recovers is the average treatment effect on the treated (ATT) – the effect among units that actually received treatment over the periods observed – rather than the average treatment effect (ATE) for the population as a whole (Section 2.2.1). The purpose of the DiD estimate is to use within-variation in treated and control groups to condition on unit and time, simultaneously (i.e., blocking Pipe Location ← Time →

[23] Several authors have used Snow's Grand Experiment to introduce Difference-in-Differences or placed it in a DiD framework. See, for example, Caniglia & Murray (2020), Coleman (2019), Cunningham (2021), and Huntington-Klein (2022).

[24] Even where the assumption is violated, the DiD estimator would be less biased than the simple fixed effects one (Section 5.5) where the difference in time trends ($t_L - t_{SV}$) is smaller than the time-trend in the treated unit itself (t_L).

Cholera and Pipe Location ← Water Company → Cholera in Figure 5.7). The basic 2x2 DiD empirical set-up is:

$$Y_{it} = \beta_0 + \beta_1 \cdot Treat_i + \beta_2 \cdot Post_t + \beta_3 \cdot Treat_i \cdot Post_t + \varepsilon_{it}$$

where $Treat_i$ is an indicator for whether the observation is treated at any point and $Post_t$ is an indicator for whether the observation falls in the period after the treated unit was treated. The coefficient β_3, which is the unique difference post-treatment in the treated group, represents the causal effect of treatment (subject to the parallel trends assumption). In practice, a regression model may also include control variables that make the parallel trends assumption more plausible or could be paired with matching to identify plausible comparator observations.[25]

As the parallel trends assumption relates to counterfactual outcomes (how the outcome would have changed over time in the treated group, absent treatment), it cannot be verified (Cunningham, 2021). However, in settings where more pre-treatment data are available, it is possible to perform two types of placebo test to at least explore the plausibility of the assumption. First, one can assess empirically whether trends were parallel in this pre-treatment period (e.g., one might assess whether changes across 1844-1849 were the same in Lambeth and Southwark-Vauxhall). Second, one can use placebo treatment dates to assess whether non-zero 'treatment' effects are observed (e.g., one could use data from 1839 and 1844 and pretend that the pipe was moved in 1844).

Failing either test does not prove that the parallel trends assumption is violated (the assumption concerns the post-treatment period), but it should give some cause for concern. However, one should also consider the degree to which the assumption would need to be violated to overturn results (i.e., taking a quantitative bias analysis approach; Rambachan & Roth, 2023). Further, including extra untreated comparator groups similarly violating the parallel trends assumption can sometimes still identify causal

[25] Readers may also explore the Synthetic Control method, which weights untreated observations to create a synthetic control data unit such that pre-treatment trends in the outcome for treated and synthetic control are similar. See Cunningham (2021) for further detail and Vagni & Breen (2021) for an application in cohort data.

effects. For instance, to estimate the effect of government provision of emergency birth control on teen pregnancy rates using local authority (LA) level data, Girma & Paton (2011) use pregnancy rates among older women as extra controls (difference-in-difference-in-differences) on the assumption that these share similar within-LA time trends to younger women.[26] See Cunningham (2021) for a decomposition of the 'triple-difference' design.

Parallel trends can be violated for several reasons, in addition to treated and untreated groups evolving separately. One reason is changes in composition in the groups over time (e.g., if Southwark-supplied households moved to Lambeth-supplied houses). A specific issue in cohort data is where there is non-random attrition. Another issue is anticipatory effects, a type of reverse causality, where the promise of treatment in future affects pre-treatment outcomes (e.g., fear of future job loss could affect present wellbeing). Finally, the assumption of parallel trends is dependent on the scale of the outcome; it may, for instance, hold on the natural (linear) scale but not on the log (percentage) scale, or vice versa.

In cohort data, more than two time periods may be available. In this case, the set of time and unit fixed effects can be added to the regression model – including both time and unit fixed effects in a model is referred to as the two-way fixed effects (TWFE) estimator, and is an extension of the standard fixed effects estimator seen in Section 5.5. Having data from more than two time periods can also enable one to assess treatment effect dynamics – i.e., how the differences between treated and untreated groups develop over time – for example, by adding further interaction terms between group and time dummies or modelling time trends in each group parametrically (e.g., with linear terms). This is useful where causal effects depend on length of exposure (e.g., long-term unemployment vs. short-term unemployment). Note, when using repeated data from the same observational units (i.e., cohort members), care must be

[26] The triple-difference design has overlap with the negative control outcome and negative control population designs discussed in Sections 5.15 and 5.16.

taken to ensure standard errors reflect potential interdependencies (e.g., by using clustered standard errors).

In many settings, including in cohort data, the 'event' may occur at different times for different people (e.g., due to a policy being rolled out to different areas over time or people experiencing an event [e.g., unemployment] at different ages). This creates issues where treatment effects vary over time. To see this, imagine we have three groups: never treated, early treated, and late treated. With these data, there are four 2x2 DiD comparisons, which would be captured by a standard DiD estimator: early treated vs. never treated; late treated vs. never treated; early treated vs. late treated in the early post-treatment period; late treated vs. early treated in the late post-treatment period. Table 5.3 decomposes the DiD estimator for the late treated vs. early treated comparison, which causes the issue. Now the difference in the differences is $d_L + t_L - t_E - \Delta d_E$ (Column 4, Row 3; Δd_E refers to the change in the treatment effect between pre- and post-late period for the early treated group). In other words, to represent an unbiased estimate of the treatment effect on the late treated (d_L), it requires the parallel trends assumption as before ($t_L - t_E = 0$), but also a constant treatment effect over time (i.e., $\Delta d_E = 0$).

Table 5.3: Decomposition of the 2x2 DiD estimator comparing late treated vs. early treated groups when the treatment effect evolves over time.

Group	Decomposition		
	Pre_{Late}	$\text{Post}_{\text{Late}}$	Diff.
Early Treated	$u_E + d_E$	$u_E + d_E + \Delta d_E + t_E$	$\Delta d_E + t_E$
Late Treated	u_L	$u_L + d_L + t_L$	$d_L + t_L$
		Difference in Differences:	$d_L + t_L - t_E - \Delta d_E$

The TWFE estimator uses already-treated units as controls for later-treated units, generating bias. Estimators have been developed to avoid this – e.g., only using comparisons with never-treated or not-yet-treated groups. Goodman-Bacon (2021) shows that a DiD estimator that combines data from groups treated at different times is

a weighted average of the separate 2x2 DiD comparisons that can be constructed. Weights depend on the group sizes but also on variation in treatment status, so are larger for groups treated near the middle of the observation window. This also means weights (and resulting estimates) can differ by changing the length of the observation window. The details of this are technical and beyond the scope of this document; see Cunningham (2021), Huntington-Klein (2022), or Goodman-Bacon (2021) for further details.

5.6.2 Applications

Li & Avendano (2023) use MCS data to assess the employment and mental health impacts of the Lone Parent Obligation (LPO), a welfare-to-work scheme that placed job-search requirements on lone parents receiving benefits, that was introduced during MCS cohort members' childhoods. The authors use a DiD design, comparing children in lone-mother and two-parent households before and after the rollout of the policy to MCS-aged households in 2012. They find that, relative to the change in outcomes in the two-parent group, the introduction of the LPO was associated with greater maternal employment but also greater maternal and cohort member psychological difficulties in lone-parent households. The authors also find some evidence in support of the parallel trends assumption when looking at pre-treatment trends. However, their analysis assumes that parallel trends would have held post-LPO, which in this case would be during the cohort members' adolescence – it is possible there are unique difficulties at this age in lone-parent households, which would violate parallel trends. Further, there were policy changes other than the LPO over the period which may have disproportionately affected lone parents (e.g., the benefit cap and the move to Universal Credit).

An additional limitation with the study was that the treatment variable – lone-parent household – is in fact time-varying and could be biased by changing composition of this group, especially if the LPO influences partnership decisions. However, they observed similar results when restricting to households whose partnership status did not change. An additional analysis that was not explored by the authors but could have been informative would be to compare lone-parent households according to whether they

were or were not in receipt of income support pre-LPO (and thus may vary in their exposure to the policy).

Avendano & Panico (2018) also use the Millennium Cohort Study for policy evaluation. Specifically, they estimate the wellbeing and health effects of a 2003 legislative change that encouraged, though did not mandate, employers to implement flexible working practices (e.g., part-time and school-term contracts). Using a DiD design, they compare a group likely unaffected by the law (working mothers whose employers in Sweep 1 – prior to the law's implementation – already offered flexible working) with a group who could be affected (working mothers whose employers did not offer flexible working in Sweep 1). They find no consistent evidence of an effect of the policy on mothers' health and wellbeing in Sweeps 2-4 (2003-08). As the pre-treatment period comprised only Sweep 1, the authors were not able to assess differences in pre-treatment trends.[27]

Note, the estimand targeted in this study is the causal effect of the law upon mothers employed (at baseline) by firms yet to offer flexible working practices. While relevant for policy, such an effect would presumably mainly operate among the subset of women whose employers begin to offer flexible working practices. The estimand is therefore not the effect of the flexible working practices per se. Moreover, mothers expected to benefit most from flexible working practices could plausibly have chosen employers already offering these prior to the law's enactment. The estimand therefore does not represent the effect of flexible working practices on those who choose such firms either. Both of these factors may explain the null results and have ramifications for the interpretation of the study.

5.6.3 Opportunities

Given the length and breadth of CLS's cohorts, there are potentially many opportunities to use DiD designs with the data, both for policy evaluation and to assess the effect of events happening within cohort members' (or their families') lives (unemployment, divorce, etc.). This is also helped by the large, diverse samples that the cohorts

[27] A possible alternative (or addition) in this case, which the authors did not avail themselves of, would be to use negative control 'placebo' outcomes (see Section 5.15).

comprise as this can make it possible to identify groups plausibly affected and unaffected by an exposure (e.g., as with lone- vs. two-parent households in Li & Avendano, 2023).

The cross-sectional variation used in the Li & Avendano (2023) and Avendano & Panico (2018) analyses described above exploited differences in whether households were targeted by the policy. Other opportunities can arise from geographic variation. For instance, Hawkins et al. (2011) use MCS data to compare smoking rates in Scotland and England before and after the introduction of anti-smoking legislation in Scotland.[28] DiD designs can also arise from temporal variation in birth dates, though this is limited in CLS cohorts due to the recruitment criteria for members of some cohorts (e.g., NCDS cohort members were born in a single week of March 1958), though this limitation does not apply to cohort members' families (e.g., parents). Temporal variation can also appear due to variation in interview dates (see Section 5.8).

5.7 Study-Initiated Experiments

5.7.1 Method

Several of CLS's cohorts have embedded randomised experiments to test the effect of different survey methodologies on participation rates or measurement. We discuss these experiments here as we believe they are currently underutilised. That said, the material is somewhat out of keeping with the rest of this document as it is not focused on making causal inferences with observational data. This section can therefore be safely skipped if it is not of interest. Note, experiments have their own methodological considerations which interested readers will need to explore if opting to analyse the data.

Many of the experiments in CLS's cohorts have only been implemented in small pilot studies and used to inform ongoing data collection, but some have been implemented among most cohort members. Almost all have appeared in recent sweeps as the

[28] Rather than a DiD setup, Hawkins et al. (2011) condition upon lagged dependent variables. See Ding & Li (2019) for discussion of the correspondence between the DiD and lagged dependent variable estimators.

studies have become more complex and as survey technologies have developed. Below is a brief overview of experiments in CLS's cohorts that we are aware of. Further details can be found in academic publications (see next subsection) or sweep-specific technical reports.

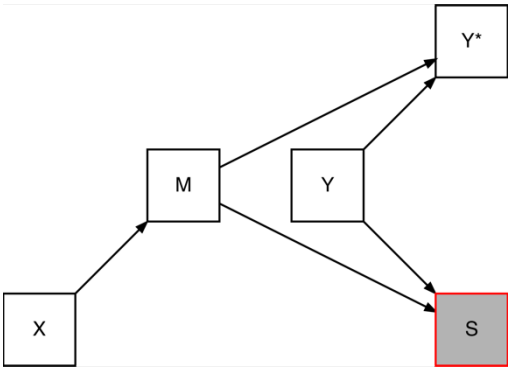
Table 5.4: Experiments in CLS's Cohort Studies.

Experiment	Description
NCDS Age 55y (Sweep 9) Mixed-Mode Survey Experiment	Cohort members with a known UK telephone number (n = 10,558) were randomised to telephone-only (n = 1,499) or a web-then-telephone sequential mixed-mode design (n = 9,059).
BCS70 Age 42y (Sweep 9) Targeted Interviewer Experiment	A set of cohort members who had refused prior surveys (n = 1,009) were randomly allocated to 'high performer' interviewers known to achieve high response rates or to other interviewers.
Next Steps Age 25y (Sweep 8) Early-Bird Web Incentive Experiment	Participants in a soft launch sample (n = 2,215) were given £10 to complete the survey which, for half the sample, increased to £20 if the survey was completed within three weeks of the invitation and via web questionnaire.
Next Steps Age 25y (Sweep 8) Progress Bar Experiment	Participants responding via web were randomised to see a progress bar or not while completing the questionnaire, though an error in coding meant no participants early in fieldwork saw the progress bar. The experiment was originally intended to be run only during the soft launch but continued when there was no clear evidence of effects on completion rates; such results-driven decisions can increase the risk of false positives (Simmons et al., 2011).
Next Steps Age 32y (Sweep 9) Differential Incentive Experiment	Participants in a soft launch sample (n = 3,206) were given a tapered incentive, which for some participants was dependent on their response at the Age 25y sweep. In one trial arm, Age 25y non-respondents were offered £35 to complete the Age 32y survey (£25 after the first three weeks) while respondents were offered £25 (£15 after the first three weeks). In the other arm, a single tapered incentive was used (£30, dropping to £20 after the first three weeks). Offering a greater incentive to non-respondents did not boost participation rates, so the single tapered incentive was used post-soft launch.
Next Steps Age 32y (Sweep 9) Interview Length Experiment	Non-respondents from a soft launch sample (n = 1,337) were randomised to be invited to complete either a short or long version of a mop-up web questionnaire. As the short version was more successful at boosting response, it was used for (post-soft launch) non-responding cohort members.
MCS Age 9m (Sweep 1) Health Records Experiment	In the sweep dress rehearsal, parents of cohort members were randomised to be asked to give consent for either (a) the interviewer to copy information from the cohort member's Personal Child Health Record (Red Book) or (b) routinely collected health data to be linked to survey responses.

Experiment	Description
MCS Age 7y (Sweep 4) Refusal Experiment	Refusals at this sweep were randomised to one of four groups in a 2x2 design where (a) cohort members' families were sent or not sent a leaflet encouraging a response and (b) all refusing families were reissued or only those judged to be soft or circumstantial refusals were reissued.
MCS Ages 7y-11y (Sweeps 4-5) Between-Sweep Mailout Experiment	In 2010, cohort families (n = 15,653) were sent letters requesting contact information. Families were randomised to receive letters that used formal or informal and simpler language, the latter also being signed by the cohort maintenance officer, rather than the study Principal Investigator.
MCS Age 11y (Sweep 5) Interviewer Pay Experiment	For cohort families in the first wave of fieldwork, interviewers were randomised to receive a bonus fee for successful interview or not.
MCS Age 23y (Sweep 8) Mixed-Mode Survey Experiment	Cohort members in the soft launch sample were randomised to face-to-face interview-only or a web-then-face-to-face sequential mixed-mode design.
MCS COVID-19 Wave 3 Survey Incentive Experiment	Cohort members were randomised to receive a £10 incentive for successful completion of the survey or no financial incentive whatsoever.

An important feature of these experiments is that the interventions implemented could have effects on both response rates and measurement. For instance, offering a web questionnaire in a sequential mixed-mode design can incentivise more cohort members to respond (Goodman et al., 2022), but it may also affect data quality if, for instance, lack of interviewer presence in the web mode leads to more hurried questionnaire completion. Given that measurements are only taken from cohort members who choose to participate, simple comparisons of observations across experimental arms may be biased due to implicitly conditioning on participation, a collider variable (Figure 5.8).

Figure 5.8: DAG representing an RCT of a survey mode experiment (X → M) where mode (M) influences measurement (M → Y*) and response rates (X → S). The mode measurement effect (M → Y*) cannot be estimated without bias using the observed data due to measurements only being observed for individuals who participated (M ← X → S ← Y → Y*). See Section 2.2.5 for further discussion of mode effects.



The differential non-response this implies should be accounted for in statistical analysis. Potentially suitable methods include multiple imputation (Section 5.4) and/or quantitative bias analysis (Section 5.20). Bias from differential non-response may be detected by examining the distribution of baseline covariates in each experimental arm or using *negative controls* (variables unlikely to be impacted by the experimental treatment; Section 5.15) as alternative outcome variables. In practice, absolute differences in response rates may be small, limiting the size of the bias arising from selection.

5.7.2 Applications

Several studies have used data from experiments in Table 5.4. Goodman et al. (2022) used data from the NCDS Age 55y mixed-mode experiment in which eligible cohort members were randomised to telephone-only or sequential mixed-mode (web-then-telephone) survey designs, in order to estimate effects of survey mode (telephone interview vs. web questionnaire) upon item and sweep non-response rates and measured responses (e.g., on socioeconomic status, health behaviours, and physical and mental health). Sweep response rates were higher for cohort members in the sequential mixed-mode arm, and there were differences across arms in reported values for several socially sensitive characteristics such as mental health and alcohol consumption (the web survey was entirely anonymous, unlike the telephone survey). Given the (small) difference in response rates by arm, missing values for non-

respondents were multiply imputed under the Missing at Random assumption. As 'treated' participants were randomised to the web-then-telephone sequential mixed-mode design, rather than to web specifically, treatment group was used as an instrumental variable (Section 5.9) to obtain a 'complier average causal effect' (CACE) – here, the average causal effect of web participation (relative to telephone response) among those who 'complied', i.e., responded via web instead of telephone when offered. Among other things, this analysis assumes that assignment to mixed-mode itself did not influence measured responses among those who instead responded via telephone.

Calderwood (2014) examined the effectiveness of a redesigned letter sent to MCS participants between waves to minimise loss to follow-up. She found that cohort families sent the redesigned letter were more likely to return the completed forms early, including among non-English-speaking families and families with parents who had no qualifications, subpopulations for whom effects were expected to be greatest. However, longer-term effects on form returns were more muted. Booth et al. (2024) investigated the effect of a monetary incentive given to MCS participants to complete a COVID-19 survey – the incentive increased response rates and improved data quality. As data quality could only be ascertained from respondents, statistical adjustment was used to account for selection resulting from differences in response rates by treatment arm.

5.7.3 Opportunities

Existing studies that have exploited experiments embedded within the cohorts have focused on short-term outcomes – e.g., response rates or measurement at the sweep in which the experiment took place. As further sweeps become available it will be possible to examine long-term outcomes – for instance on attrition from the cohorts or the measurement of change across time (e.g., ageing and other trajectories). It is possible that incentives to participate have only short-term effects on retention if marginal participants drop out at subsequent sweeps. Alternatively, the offering of incentives in just one sweep could reduce long-term response, if the removal of financial incentives reduces intrinsic motivation (cf. Gneezy & Rustichini, 2000). Existing studies assessing effects of survey design on measurement have also focused on univariate descriptive

statistics, but this is only one type of analysis performed with cohort data. Bias in sample means and proportions may not translate into material levels of bias in, e.g., associational analyses, which could be tested in the experimental data.

Studies that have investigated the effects of survey design on measurement and data quality have used statistical control (Booth et al., 2024) or multiple imputation (Goodman et al., 2022) to account for differential non-response. Quantitative Bias Analysis (Section 5.20) could instead be used to estimate effects under various plausible models of selection, though differences in response rates across treatment arms are typically small and selection effects may not be large enough to make substantive differences to results.

It is sometimes the case that variables capturing treatment arm within the experiment are not available in the standard datasets on the UK Data Service and require bespoke data access instead. It is also sometimes the case that data on the UK Data Service are only available for participants at a given sweep, though non-respondents may have been part of the experiment, too. As noted, non-response can lead to bias in analyses of experiments, so data from non-respondents may be required. This is the case for the NCDS Age 55y mixed-mode experiment, at least. Data enquiries can be sent to the CLS Data team.

5.8 Natural Experiments

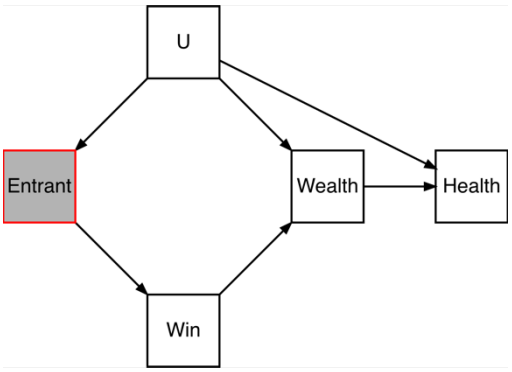
5.8.1 Method

In RCTs, researchers intentionally use a procedure to assign people to treatment groups at random. In expectation, the outcome of the procedure is unrelated to any factor – known or unknown – that affects later outcomes, other than the treatment it determines (and its consequences). In so far as the randomisation procedure can be said to affect outcomes, it is via the intervention that membership of a particular treatment group entails. For example, in sufficiently large RCTs that use the results of a fair coin toss to determine assignment, the results of coin tosses should not be correlated with participants' sex, age or other characteristics, other than through chance, but could still be related to outcomes through the treatment group the coin tosses determine. This

bears repeating: the coin toss can have a causal effect on outcomes, but it is only through its effect on treatment group.

Analogous processes sometimes occur without a researcher's input. Most clearly, this occurs where randomisation is adopted for reasons other than sating academic interest – for instance, to allocate a resource. During the Vietnam War, the US Government used a lottery to draft soldiers. Besides being a politically palatable method of conscription, the draft has also been a valuable tool for understanding the effects of military service on various life outcomes (see, e.g., Conley & Heerwig, 2012). Games of pure chance, such as roulette and national lotteries, provide players with an endorphin rush, but also help researchers study the effects of wealth shocks on, e.g., health (Cesarini et al., 2016). Note, in each of these, randomisation occurs only among people who are entered into the lottery. In the Vietnam draft, this was healthy males of conscription age. In gambling lotteries, it is individuals who have bought a ticket (Figure 5.9). Comparing draftees or prize winners with everybody else, entrant or not, would not recover a causal effect.

Figure 5.9: DAG representing a study using lottery wins to examine the effect of wealth on health. In observational data, the association between wealth and health may be biased by a host of confounders U (e.g., education level). Lottery wins represent a source of random variation in wealth, though only among lottery players – the decision to enter a lottery may be influenced by factors that also influence health (e.g., education level), but which player wins is a random process (unless rigged). More accurately, the total number of lottery tickets bought determines the expected win amount, so an ideal lottery study would condition upon number of tickets bought.



The Vietnam draft and national lotteries are examples of *natural experiments* and have the same properties that justify RCTs – assignment (a) influences exposure, (b) only has

an effect on outcomes via its effect on exposure, (c) is not determined by other causes of the outcome, and (d) individuals (or other relevant observational units) subject to the natural experiment have a non-zero chance of both exposure and control. As with RCTs, these properties arise because of the explicit use of a randomisation procedure. But it is these properties that are of value, not randomisation itself.

There are many other cases where random processes are used to allocate resources. Sometimes this is an intentional feature of the process, as with lotteries. Other times it is a byproduct of other concerns. In social housing, individuals at the top of a waitlist may be offered whatever accommodation becomes available first (Silva et al., 2025). In health, being born on a weekday rather than the weekend influences staff availability and thus the breastfeeding advice a new mother and her baby receive (Fitzsimons & Vera-Hernández, 2022). Processes like these are not random in the sense that exposure occurred for no predictable reason; housing becomes available when someone leaves a property and hospitals lower costs by reducing staff levels on weekends. Rather, the reasons for exposure are unrelated to potential outcomes: expectant mothers have little control over whether they go into labour on a Friday or a Saturday, and people on a housing waitlist do not control which houses become vacant as they move to the top of the list. Assignment mechanisms like these are said to be "as-if random" and have the properties that justify RCTs – assignment only has an effect on outcomes via the effect on exposure and assignment is not determined by other causes of the outcome.

Many natural experiments appear in survey data due to as-if random assignment to interview dates. The Unexpected Event During Survey Design (UESD; Muñoz et al., 2020) exploits large-scale events, such as major sporting competitions (e.g., the Olympics or the World Cup), political incidents, natural disasters, and terrorist atrocities occurring at some point during survey fieldwork and which bifurcate the sample into those surveyed before the event and those surveyed during or after. On the assumption that interview timing is unrelated to potential outcomes, comparing groups surveyed before and after the event can be used to determine the event's causal effects, such as its effect on measures of psychological wellbeing, behaviour and political and social attitudes.

Assessing whether the assumption of as-if random interview timing is plausible in a given application requires detailed knowledge of the survey being utilised, its design and how it was conducted. The assumption may be violated if, for example, participants with the lowest chance of response are allocated to be interviewed first, so that further contact attempts could be made before the end of fieldwork, if necessary. In many surveys, participants are allocated to fieldwork 'waves' based upon shared characteristics – random assignment to interview date *within* these groups may be plausible, but between groups, not. The assumption of random timing can also be violated if the event itself affects survey design (e.g., delaying fieldwork) or a participant's likelihood of responding. Another way the assumption can be violated is if there are pre-existing time trends that would have produced different average outcomes before or after the date of the event had it not actually occurred (e.g., if there is seasonality in recorded wellbeing). Further, the assumption can be violated if there are other events occurring simultaneously or there are collateral events, in which case the effect of the event itself cannot be identified or generalised to similar events that do not produce the same sequelae. Events like these are examples of 'fat-handed interventions' (Bailey et al., 2024) that involve a bundle of possible causes of an outcome, making it difficult to determine which aspects are operative.

Another usual assumption of the UESD is that the participants surveyed after the event were in fact exposed to the event (i.e., aware of it). There are several ways of assessing this, including indirectly: in the absence of a specific survey item on the event, analysts can use data sources such as Google Trends, contemporaneous polling, or news headline aggregators to assess the likelihood a person knew about the event. If fewer than 100% of those surveyed after the event were exposed to it, the estimate will be attenuated roughly in proportion to the share who were exposed. (This is equivalent to an *Intention to Treat* analysis, discussed further in Section 5.9.) In practice, differences in awareness can be beneficial as an elaborate theory would predict weaker or non-existent associations in groups less likely to be exposed – for instance, among those who do not read a newspaper or own a television (see James, 2024, for an example in this vein).

The UESD design does not map onto a single statistical method, though it has clear similarities with the Sharp Regression Discontinuity Design (discussed further in Section 5.10). UESD analyses can proceed cross-sectionally, with responses compared in groups recorded before and after the event (potentially in subgroups for whom the assumption of random timing is more plausible, as discussed above). The longitudinal nature of CLS cohorts can also be exploited – for instance, by performing fixed effects analysis to account for time-invariant confounders (Section 5.5) or matching upon prior measurements of the outcome (Section 5.2) to strengthen the plausibility that interview timing is independent of potential outcomes. Note, assumptions about random event timing can also be tested in part by (a) assessing covariate balance either side of the event and (b) assessing trends in the outcome variable in the period prior to the event occurring (i.e., to examine whether there were pre-existing time trends).

There are other natural experiments contained in CLS's cohort data, including those related to the timing of birth (e.g., influencing relative age in school year and the minimum age of school leaving), the rollout of public policies, and the nature of how systems distribute resources. There likely exist others that are as yet unidentified as uncovering these requires detailed knowledge of relevant policy areas (e.g., awareness about how social housing is allocated). Several natural experiments are analysed in Instrumental Variables or Regression Discontinuity Design (RDD) frameworks, for reasons that will be discussed in Sections 5.9 and 5.10. It is worth noting, though, that some events may at first glance appear random but in fact are not. For instance, one may expect dying by murder or from accidents to reflect happenstance, but smokers have a higher chance of the former (Davey Smith et al., 1992) and women using HRT have lower odds of the latter (Petitti et al., 1986).

5.8.2 Applications

The UESD is relatively underutilised in CLS's cohorts. Two studies have used data from Next Steps to examine the effect of the 7th July 2005 London Bombings, which were followed by rises in reported hate crimes committed against Muslims (Astorga-Rojas, 2023; Hole & Ratcliffe, 2020). Hole & Ratcliffe (2020) show a specific decline in happiness (though not an increase in depression) among Muslim girls in the period after

7th July, as well as an increase in their expectation of discrimination in the labour market. Supporting a causal interpretation, the authors also show the effect is specific to Sweep 2 of Next Steps, during which the event happened, rather than also present in Sweep 4, when measures of mental health were again collected. Astorga-Rojas (2023) uses a similar set-up and shows lower reported expectation to remain in full-time education after compulsory schooling among Muslim cohort members post 7th July.

Multiple papers have used data from individuals in social housing, exploiting the quasi-random nature of housing offers for those at the top of waiting lists at various times in the UK in the 20th and 21st centuries (Fumagalli & Fumagalli, 2019; Patacchini & Zenou, 2011; Silva et al., 2025). For instance, Fumagalli & Fumagalli (2019) use data from Next Steps and find lower social participation (operationalised as 'hanging around') in areas with greater ethnic diversity (which housing offers produce exogenous variation in). Note, a key issue with this design is that, although allocation to neighbourhood may be random, neighbourhoods vary on many different dimensions simultaneously. Though neighbourhood of residence may not be related to individual (baseline) characteristics among social housing tenants, it is not clear what aspect of neighbourhoods may be driving effects.

Several papers have used month of birth as a potentially exogenous determinant of relative age in one's school year – in the UK, people born in September are the oldest in their class, whereas those born in August are usually the youngest, but parents have limited control over the time of year they conceive. Month of birth is often analysed in an RDD framework (Section 5.10) with data from multiple year groups pooled (e.g., comparing students born in August or September in a given calendar year), but due to the design of CLS's cohorts, comparing individuals within a school year is the only feasible approach, and only for MCS and Next Steps (which roughly span one school year), rather than NCDS and BCS70 (which only span a single week).[29] Crawford et al. (2011) examine month of birth effects on multiple developmental outcomes in MCS and Next Steps cohort members, and show that, among other factors, cohort members born

[29] Cohort members' parents, siblings and children, however, have greater variation in dates of birth.

early in the school year have greater educational attainment and higher scores on cognitive tests than those born late in the school year.

5.8.3 Opportunities

As noted, the UESD is relatively underutilised in CLS's cohorts, especially given the cumulative timespan of data collection represented. Several cohorts also contain data on accidents, though as discussed, the occurrence of accidents can be related to participant characteristics. Data on income and wealth shocks may also be available. We recommend reading the original UESD paper (Muñoz et al., 2020) as it is filled with practical advice and may spark ideas. A research-enabling project would be to produce databases of (a) policies and headlines detailing important events that have occurred during CLS survey fieldwork and (b) interview date ranges and (sub)samples where interview timing is plausibly exogenous (identified, e.g., by assessing sample balance around an index date). Note, however, exogeneity of interview timing only needs to be met with respect to potential outcomes of interest – there may be systematic reasons why individuals were interviewed on given dates, but if unrelated to potential outcomes, the assumption of (conditional) exchangeability would not be violated.

5.9 Instrumental Variables

5.9.1 Method

One issue with RCTs is that not everyone may comply with the intervention to which they have been assigned – in a medical trial, some of the 'treated' may stop taking the drug they have been given, and some of the 'control' may find the drug through other means, instead of remaining on placebo. Analogously, in natural experiments not everyone may receive or follow their expected intervention. For instance:

1. In the UESD, some people may be unaware of the event occurring, in extreme cases like the reclusive Lykov family or Hiroo Onoda and other Japanese holdouts, who were unaware of the end of World War II.
2. In the Vietnam lottery, some draftees absconded, while some non-draftees signed up to serve.

3. In maternity wards, some mothers breastfeed or don't regardless of whether advice from a midwife is received or not (Fitzsimons & Vera-Hernández, 2022).

Simply comparing outcomes by assignment (known as an *Intention to Treat* [ITT] analysis), while often useful in policy terms, will not necessarily identify the causal effect of the treatment itself. Individuals are typically only randomised to receive *encouragement* to follow a certain treatment, rather than to adopt the treatment, per se – for instance, they are provided a box of medication and given instructions to take it, rather than administered it through force. To identify the causal effect of treatment, we instead need to look at those who *comply*.

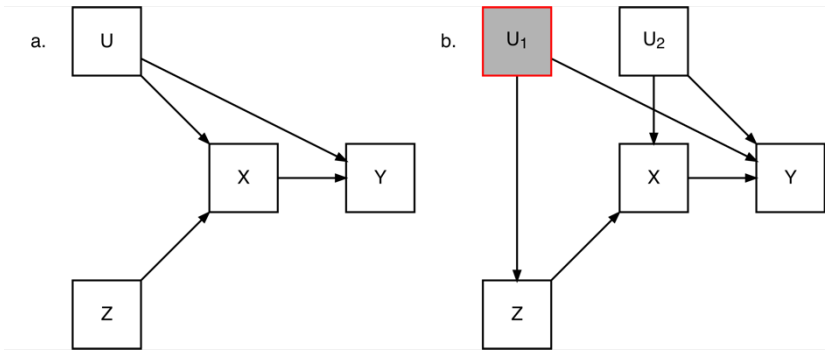
This is not as simple as comparing outcomes among those who used the treatment because it is <u>not</u> *use* of treatment that is randomised. For instance, if we compare, among those entered into the Vietnam lottery, those who served and those who didn't, we would compare on the one hand those who (1) served *because* they were drafted (*compliers*), (2) served because they would serve regardless of being drafted (*always takers*), or (3) served because they were *not* drafted (*defiers*) with, on the other hand, those who (a) did not serve *because* they weren't drafted (*compliers*), (b) did not serve despite being drafted (*never takers*), or (c) did not serve because they *were* drafted (*non-compliers*). These are not equivalent groups and they are not necessarily identical according to their potential outcomes. See Table 5.5, which continues this worked example.

Table 5.5: Numerical example showing that even with random assignment of treatment encouragement, comparing outcomes by treatment used can yield a biased estimate of the causal effect if there is non-compliance (i.e., not all individuals use the treatment they were assigned to). The example uses the Vietnam draft lottery where people either serve or don't serve (treatment used), depending on whether they are drafted or not drafted (encouragement assigned). People can be split into four groups depending on how they behave based on treatment assignment. For simplicity we assume that people are equally split across these four groups to make the arithmetic simpler to follow. For each group, the causal effect of serving is 2 (to verify, compare 'Potential Outcome' columns). However, people who choose to serve when drafted have higher potential outcomes than those who do not. The difference in observed outcomes among those who serve and those who don't is 4 (final two rows) rather than 2, a biased estimate.

Group ($X^{Drafted}$, $X^{Not\ Drafted}$)	Activity			
	Serve		Don't Serve	
	Potential Outcome	# Enacting Activity	Potential Outcome	# Enacting Activity
Serve, Serve (Always Takers)	9	N	7	0
Serve, Don't Serve (Compliers)	10	N/2	8	N/2
Don't Serve, Don't Serve (Never Takers)	5	0	3	N
Don't Serve, Serve (Defiers)	6	N/2	4	N/2
E(Y | X = x)	= (9×N + 10×N/2 + 5×0+6×N/2)/2×N = 8.5		= (7×0 + 8×N/2 + 3×N+4×N/2)/2×N = 4.5	
E(Y | X = Serve) – E(Y | X = Don't Serve)			= 8.5 – 4.5 = 4	

Accounting for non-compliance in the estimation of causal effects can instead be achieved using the Instrumental Variable (IV) design (Figure 5.10a). This design utilises variables that create – or are associated with (Goetghebeur et al., 2020) – exogenous variation in treatment, nudging subjects to a particular level of the exposure, rather than another, analogous to the role of a coin in an RCT. For instance, being picked in the Vietnam draft raised the probability a person would serve, but did not make it inevitable, while giving birth on a weekday makes it more likely, but by no means certain, that a mother will breastfeed.

Figure 5.10: DAG representing the instrumental variable design. The instrument Z is an exogenous cause of variation in the exposure, X, and can thus be used to determine the effect of X upon Y (at least, for the variation induced in X by Z). The instrument meets three criteria: relevance (Z → X), exclusion restriction (Z → X → Y is the only directed path from Z to Y), and (conditional) independence – i.e., no open backdoor path from Z to Y. In (b), conditional independence is achieved by controlling for the confounder U_1.



In addition to meeting the assumptions of positivity, consistency and no interference, to be valid, an instrumental variable must have four properties (Goetghebeur et al., 2020). First, it must be associated with exposure (*relevance*; Z → X in Figure 5.10a). Second, it must be unrelated to any unmeasured confounders of the exposure-outcome relationship (*independence*; no Z ← U → Y). Third, it must be independent of the outcome conditional on the treatment and unmeasured confounders (*exclusion restriction*; Z → X → Y, the only directed path from Z to Y), unless these are conditioned upon (Figure 5.10b); the lottery used in the Vietnam War draft is an instrumental variable but only among males eligible to serve.[30] Fourth, the size of the treatment effect must be independent of the instrument (*homogeneity*) *or*, where there is effect heterogeneity, there cannot be both compliers (people who use treatment only when encouraged) and

[30] Typically, and as displayed in Figure 5.10, the three IV assumptions are stated referring to causal effects rather than (conditional) associations. For instance, relevance is defined as the IV having a causal effect on the exposure and exclusion restriction as the IV not having causal effects on outcomes except via the exposure, X. In the causal framing, not conditioning on U_1 in Figure 5.10b may be classified as a violation of independence (i.e., presence of open back-door paths) rather than exclusion restriction as above. In the causal framing, selection paths (e.g., adding Z → S ← Y to Figure 5.10b) do not fall easily under any assumption's definition, whereas they sit straightforwardly under exclusion restriction as the assumptions are defined above.

defiers (people who use treatment only when *discouraged*) within the same study population (*monotonicity*).[31]

Under these assumptions, it is possible to use estimates of the association between the instrument and the exposure and the instrument and the outcome (in Figure 5.10, paths Z → X and Z → X → Y, respectively) to obtain an estimate of the causal effect of the exposure upon the outcome (X → Y); in effect, exploiting exogenous variation in the exposure related to the IV (Z → X). Several IV estimators exist. The most commonly used, two-stage least squares regression (IV-2SLS), uses the instrument to predict the exposure variable, and then uses these predicted values to estimate associations with the outcome variable. Unlike ITT analysis, this procedure provides an estimate of the change in the outcome per unit change in the exposure. Similarly, the widely used Wald estimator scales the association between the instrument and the outcome by the instrument's association with the exposure.

Where there is treatment effect heterogeneity and the monotonicity assumption is met, IV methods estimate a specific quantity termed the Local Average Treatment Effect (LATE). This is the average effect of the exposure within the variation induced by the instrumental variable, i.e., among *compliers* – people whose level of the exposure is sensitive to (changes according to) the level of the instrument (for instance, those who served in the Vietnam War only because they were drafted). The LATE, which is sometimes referred to as the Complier Average Causal Effect (CACE), can differ from a global average treatment effect because causal effects may be heterogeneous according to why someone was exposed. For instance, the effect of military service on income among those who only served due to the draft could differ from the effect among those who served voluntarily (especially if the latter did so in expectation of serving's beneficial effect on their outcomes).

Discrepancies between observational associations and IV estimates may therefore arise not just because of bias in the observational association, but also because a

[31] There are various (relatively technical) versions of the homogeneity assumption. See Goetghebeur et al. (2020) and Hernán & Robins (2023) for more detail.

different quantity is being estimated. It is generally not possible to identify which individuals are compliers because the 'fundamental problem of causal inference' (Holland, 1986) reappears: we cannot see both states of the world, where a person receives the instrument and doesn't. However, characteristics of the population of compliers can be described (assuming monotonicity) by comparing the attributes of the exposed among those with the instrument and those without. Nevertheless, this population may not be of particular scientific or policy interest; we may wish to know the effect of military service on recruits in general, for example. The price of internal validity is potentially lower external validity.

Where treatment effects are heterogeneous, monotonicity is required because, otherwise, the total effect of the instrument on the exposure becomes a net combination of opposite responses, making the resulting instrumental variable estimate difficult to interpret (for worked examples demonstrating this, see Cunningham, 2021; Huntington-Klein, 2022). IV methods can exploit multiple IVs, which can increase statistical power and precision, but make interpretation more challenging where treatment effects are heterogeneous as the estimated treatment effect is then an average across different groups. Additionally, where the exposure is continuous, rather than binary, the LATE is a weighted average where compliers are weighted in proportion to how affected they are by the IV (i.e., those whose exposure changes the most are weighted more highly), again making interpretation challenging.

Many instruments have been proposed and used in the empirical literature, particularly in economics. Examples include (Baiocchi et al., 2014; Cunningham, 2021; Hernán & Robins, 2023; Huntington-Klein, 2022; Mellon, 2025):

- Having a girl as a first birth (due to a preference for boys, a first-born girl means a family is more likely to have further children)
- Measures of access (e.g., distance to hospitals, as travel time provides an incentive to use the nearest service, regardless of its quality)
- Preferences (e.g., the judge fixed effects design relies on random allocation to judges who have different predilections for handing out long prison sentences)
- Weather (e.g., extreme weather events can mean, among other things, planned events are cancelled)

- Calendar time (e.g., the roll-out of cable networks in the US)
- Offspring traits (e.g., direct effects of genes that are shared between parents and offspring lead to correlation in parent-offspring pairs, and the offspring trait may be less confounded than the parent trait; Davey Smith et al., 2009)

However, the popularity of these IVs can itself be a cause for concern (Gallen & Raymond, 2020; Mellon, 2025). For instance, weather has been linked with over 190 treatment variables, such as income and food prices, through which it may plausibly affect outcomes (Mellon, 2025). Such disparate effects suggest exclusion restrictions do not hold, though Mellon (2025) discusses conditions under which this may not be the case (e.g., the exclusion restriction holds for the specific population under study), as well as outlining a procedure for determining whether this matters for a planned IV study. Note, IV estimates can be especially sensitive to violations of assumptions (Hernán & Robins, 2023), in particular where the instruments are 'weak' (have small association with the exposure. Baiocchi et al. (2014) and Cinelli & Hazlett (2025) describe quantitative bias analysis methods for IV designs.

Another class of IVs that are routinely used are genetic variants – i.e., as in Mendelian Randomisation (MR) analysis. (Almost) every human trait is heritable (Polderman et al., 2015; Turkheimer, 2000). Mendelian inheritance means parents pass on a random half of their genomes to their children, and thus, conditional on parents' genomes, a child's genes potentially represent a source of exogenous variation in a trait. For instance, variants of the ALDH2 gene are related to the comfort with which one metabolises alcohol and have thus been used to instrument alcohol consumption and determine its effect upon health (L. Chen et al., 2008), a canonical example of a difficult question to answer with standard observational data. An advantage of using genetic variants as instruments is that they are inherited at conception and so are not subject to reverse causality. However, this also comes with the disadvantage that genes are carried across life and so it is difficult to determine or isolate specific timing of exposure (though, see Richardson et al., 2020).

Due to data unavailability, MR analyses are generally undertaken without genetic data from multiple family members, so they cannot condition on parental genotype and lack the property that motivates MR. (MR with family-level data is discussed further in

Section 5.14.) Genetic variants are not randomly distributed around the population (Abdellaoui et al., 2019, 2022), and population-level MR analysis can be biased by (population) genetic phenomena, including population stratification and assortative mating, as well as genetic nurture (effects of parental genes on offspring via parental behaviour) and other indirect genetic effects. Population stratification refers to differences in particular genes across groups, which can lead to (non-causal) association between genes and environments – famously, genes that are more prevalent among individuals of East Asian ancestry will be associated with chopstick use, despite this having a cultural explanation (D. Hamer & Sirota, 2000). Assortative mating refers to the phenomenon of individuals not mating randomly, which can lead to associations between phenotypes (and thus the genetic variants they are influenced by) without one phenotype having a causal effect upon another. For instance, if highly educated people are more likely to mate with taller individuals, then genes that influence either will become correlated and a population MR analysis would suggest education influences height (or vice versa) when this is not the case. Both population stratification and assortative mating represent violations of the *independence* assumption.

An additional issue with MR analyses (including those using family-level data) is *horizontal pleiotropy,* where a genetic variant has an effect on multiple phenotypes, not due to the main trait of interest having an effect on the others (*vertical* pleiotropy): many genes have effects on multiple phenotypes, a necessary consequence of the limited number of protein-coding genes (~20,000) in the human genome. Horizontal pleiotropy represents a potential violation of the exclusion restriction assumption. Horizontal pleiotropy can also arise through *linkage disequilibrium* (LD), whereby variants located close together on a chromosome are correlated because recombination of DNA during meiosis (formation of sex cells, gametes) rarely separates nearby loci, so DNA is inherited in blocks. Thus, a genetic variant which has a causal effect on one trait may be in LD with a variant that has an effect on another trait, creating a correlation between the two traits.

The risk of horizontal pleiotropy is heightened when polygenic scores (PGS) are used as instruments. PGS aggregate the effect of many genetic variants (e.g., hundreds or more)

into a single weighted summary variable, with the individual weights determined from Genome-Wide Association Studies, which use a hypothesis-free approach. Unlike with single genetic variants like ALDH2, the pathways linking PGS to a trait may thus be poorly understood. This is a particular concern when studying complex traits, such as education, that are genetically influenced through multiple other traits (e.g., cognitive ability, conscientiousness, and physical health), which may themselves have independent effects on outcomes. As individual genes often have only weak effects, there is a trade-off between using genetic instruments that most plausibly satisfy the exclusion restriction (and independence) assumptions and using instruments that are well-powered. Researchers often strike a balance by using PGS created from the top 'hits' in a GWAS (e.g., variants exceeding the genome-wide significance, $p <= 5x10^{-8}$, level). Multivariable MR, in which PGS for related traits are also used (Davies et al., 2019), may also reduce bias in this situation. PGS also make it harder to meet the *consistency* assumption.

Note, however, whether population stratification, assortative mating, or horizontal pleiotropy matter in a particular MR application is an empirical question – some traits are not subject to assortative mating, nor do some genes vary in prevalence in important ways across populations, and horizontal pleiotropy may not be relevant in a specific situation. Further, as discussed in Section 5.16, clever cross-context designs may be used to elaborate the theory that the gene has an exogenous effect on the exposure and provide compelling causal evidence.

5.9.2 Applications

Fitzsimons & Vera-Hernández (2022) use data from the MCS to assess the effects of breastfeeding on various measures of child development using time and day of birth, which affect the breastfeeding support mothers receive in hospital, as instruments for breastfeeding. Major strengths of their study are their ability to demonstrate that day of birth was not related to mothers', babies' and pregnancy characteristics and that associations were confined to low-SES individuals for whom breastfeeding support was particularly likely to have an effect on breastfeeding rates. Further, the authors augment their analysis using an external dataset containing detailed data on postnatal care to

demonstrate that the relevance assumption is plausible (e.g., that day of birth is related to receiving advice and active support with breastfeeding).

Hyppönen et al. (2022) use data from parents in the NCDS to assess the effect of BMI upon mortality using cohort members' (i.e., offspring) BMI as an instrumental variable; the authors' argument is that offspring BMI is unlikely to be confounded by underlying disease, a major confounder of the BMI-mortality association in observational studies (i.e., reverse causality; Davey Smith et al., 2009). The authors find that associations between BMI and mortality are stronger when using the IV. However, they find evidence that offspring BMI is confounded by socioeconomic position (confounding bias can be amplified when using IVs; J. W. Jackson & Swanson, 2015). It is also possible the exclusion restriction is violated if, e.g., offspring socioeconomic position (which could have a causal effect on offspring BMI) affects parental mortality. Further, because the instrument captures the stable familial component of BMI, the analysis disregards temporary variation in parental BMI, and so estimates target the effect of long-run – rather than contemporaneous – adiposity.

5.9.3 Opportunities

The temporal and geographical range of the CLS cohorts, as well as linkage with administrative data, potentially opens up many opportunities for novel instrumental variable analyses. However, the discovery of plausible IVs depends on deep domain-specific knowledge. Given the wide range of data that are available in the cohorts, there may be a case for extending previous successful IV applications – e.g., to extend the Fitzsimons & Vera-Hernández (2022) work to outcomes captured later in life (e.g., educational attainment) or other domains (e.g., to personality traits); though whether IV assumptions are met for these additional outcomes would need to be examined. Data on offspring and/or parents are available in each cohort and can enable IV analyses using family members' traits. However, again, whether the IV assumptions are met will need to be assessed for the exposure and outcome under study. Finally, PGS for a wide range of traits are now available in each cohort (Shireby et al., 2025) and could be used in MR analyses. For researchers wishing to exploit single genetic variants, raw genetic data can be accessed via application to CLS.

5.10 Regression Discontinuity Designs

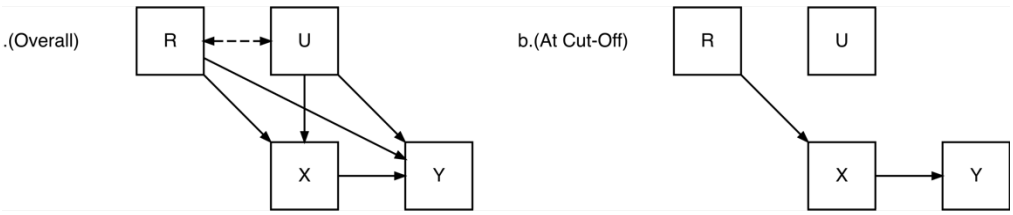
5.10.1 Method

Instrumental variable designs use variation in a third variable (the instrument) as a source or marker of exogenous variation in an exposure. Sometimes there are variables that have an effect on an exposure and vary over a wide range. While overall these variables may be associated with confounding factors, narrow parts of their variation can lead to large differences in the likelihood of exposure. Within this narrow range, variation is therefore strongly related to the exposure, but may be only weakly related – if at all – to the confounding factors.

Take democratic elections – does becoming a Member of Parliament (MP) increase one's salary after the end of one's parliamentary career (Eggers & Hainmueller, 2009)? We can't simply compare people who win elections with those who lose them because rational political parties field their most promising candidates in safer seats and voters give their support to candidates they perceive as more able. People who win large majorities differ from those who lose handily in ways that could allow them to obtain higher wages absent becoming an MP. But winners take all in a first-past-the-post system, whether the majority is one or twenty thousand. People who lose by a small amount are likely similar, on average, to those who win by a small amount. Within a small margin, random factors can be decisive (Figure 5.11).[32]

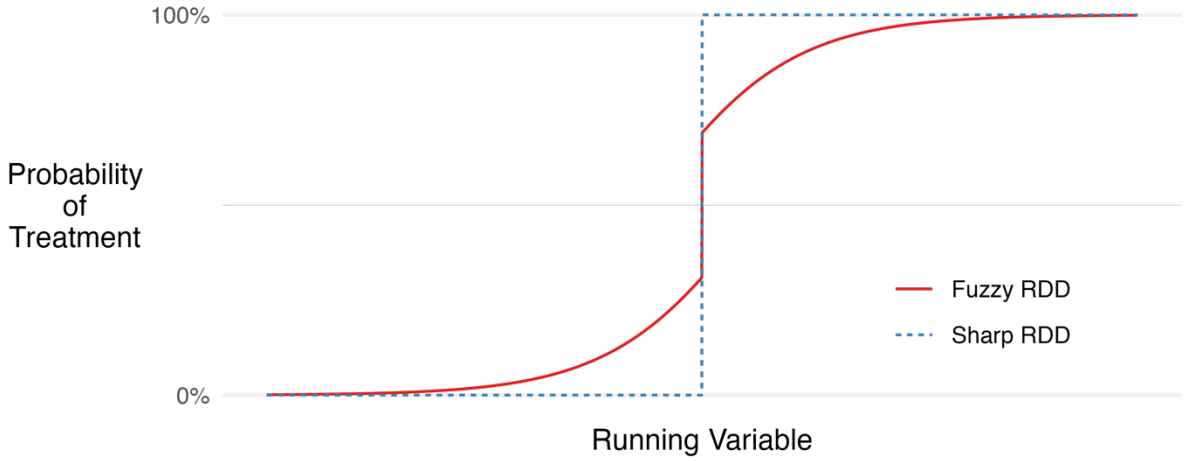
[32] In the 1997 General Election, Conservative MP for Torbay, Rupert Allason, lost his seat by twelve votes. A week before, he failed to tip a waitress, leading fourteen waiters who had intended to vote Tory to cast their votes for the Liberal Democrat candidate instead (Ronson, 2001).

Figure 5.11: DAG representing the Regression Discontinuity Design (RDD). In the RDD, there is a running variable R that influences exposure X. However, across the total span of the running variable (Panel A), this variable is itself a cause of the outcome Y or is correlated with other causes (U) of the outcome Y. Within a narrow range around the cut-off (Panel B), where the running variable strongly affects exposure, variation in the running variable is too small to affect the outcome or to be associated with its other causes. DAGs adapted from Cunningham (2021).



As its name suggests, the Regression Discontinuity Design (RDD) takes advantage of small perturbations around discontinuities such as this. RDD comes in two flavours: sharp and fuzzy (Figure 5.12). In sharp RDD, exposure is zero on one side of the discontinuity and complete on the other. The first-past-the-post election system represents a sharp discontinuity; one is voted in, or one isn't. In fuzzy RDD, probabilities of exposure are intermediate, but around the discontinuity, the likelihood of exposure changes substantially. Age-based eligibility criteria for vaccination represent a fuzzy discontinuity; some people below the threshold may be eligible for vaccination for other reasons, while those above the threshold may opt against vaccination. But ageing a matter of days can still generate large jumps in vaccination rates and by itself would not notably affect outcomes of interest (Basta & Halloran, 2019).

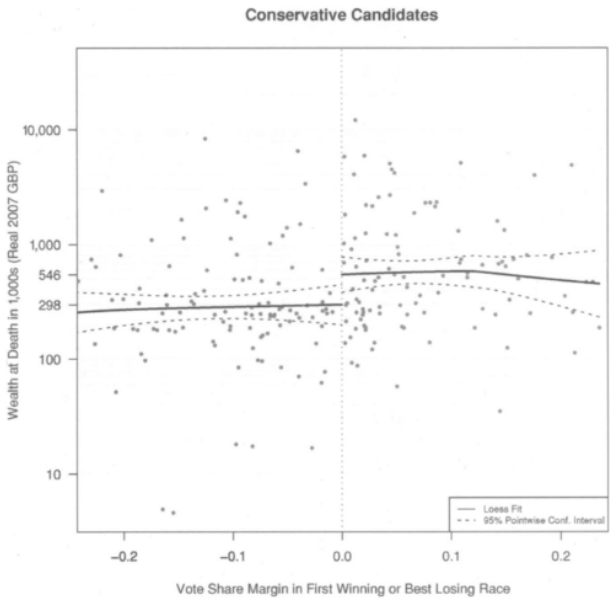
Figure 5.12: Graph representing sharp and fuzzy regression discontinuity designs (RDD). In the RDD, there is a running variable (x-axis) that influences exposure and at a cut-off there is a jump in the likelihood of being treated. In sharp RDD, no one below the cut-off is exposed and no one above the cut-off is unexposed. In fuzzy RDD, some people below the cut-off are exposed and some people above the cut-off are unexposed. However, there is a discontinuous jump in the probability of exposure around the cut-off, such that the probability of exposure is markedly higher for those just above the cut-off than those just below it.



Discontinuities are common wherever there is some form of rationing or where definitive criteria are used to make eligibility transparent or implementation more straightforward. This is a feature of most bureaucracies and administrative systems, including clinical settings where treatment protocols are used. Discontinuities graft sharp changes onto otherwise continuous characteristics. While a person may be identified as requiring medical attention based on strict criteria, underlying risks change more gradually; hospitals may use a cut-off for low birth weight of 2.5 kg but weighing 2.51 kg is little better than 2.49 kg. Discontinuities also arise through cleavages in time – a policy is enacted whereas before it was not. People born on 31st August are the youngest in their school year, while people born on 1st September are the oldest. People born on 31st August 1957 were able to leave school at age fifteen, if they so wished, but people born a day later had to wait until age sixteen because of an Act of Parliament, which raised the minimum school leaving age.

In RDD parlance, the variable determining exposure is referred to as the 'running variable' (Cunningham, 2021).[33] The basic identification strategy of RDD is to estimate the difference in the outcome around the discontinuity (for instance, the difference in wealth at death either side of a zero vote margin; Figure 5.13), because the units of analysis are, in theory, exchangeable at the time the exposure is assigned. This can be achieved using parametric or non-parametric methods.

Figure 5.13: The association between relative vote share received by Conservative candidates for UK Parliament and their wealth at death. Candidates receiving positive relative vote shares were elected to Parliament. The discontinuity at zero, and the flat relationship on either side, are evidence that being a Member of Parliament increased wealth at death among these candidates. Figure taken from Eggers & Hainmueller (2009).



In a simple parametric set-up, the relationship between the running variable and the outcome is modelled with continuous terms, and a binary variable for whether the running variable is above the discontinuity is also included in the regression (see equation below). The continuous terms capture the association between the running variable and the outcome that would appear absent the variable influencing exposure. The binary term captures the effect of the exposure on the outcome at the cut-off.

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot D(x_i \geq c) + \epsilon_i$$

[33] The running variable is also referred to as the 'assignment' variable.

where *i* indexes individuals, x_i represents the level of the running variable for person *i*, and $D(x_i \geq c)$ is a binary variable equal to 1 where x_i is at or above the discontinuity cut-off point (c).

The parametric design uses data from a wide range of the running variable. It relies on modelling the association between the running variable and the outcome appropriately – incorrectly modelled non-linearities may yield a spurious non-zero coefficient for the binary discontinuity term. In the non-parametric set-up, only data from around the discontinuity are used, with weighting possibly used to place greater emphasis on observations particularly close to the cut-off. The non-parametric approach is more data-intensive as it requires sufficiently large samples close to the discontinuity but has the advantage that it does not rely on appropriately modelling the correct functional form. The sample size requirement can be relaxed by increasing the width around the discontinuity from which observations are used, but at the cost of increasing the risk that confounding factors – including the running variable itself – explain the association with the outcome.

In fuzzy RDD, the running variable is associated with exposure above and below the discontinuity, and being above the cut-off is effectively an 'encouragement'. Instrumental variable methods (Section 5.9) are therefore used to rescale the effect by the cut-off's influence on probability of exposure. The causal effects estimated in sharp and fuzzy RDD are local to the cut-off, with the added proviso that in fuzzy RDD this is the causal effect among those who 'comply' with the encouragement – i.e., were exposed due to just sitting above the cut-off (as such, fuzzy RDD requires homogeneity or monotonicity assumptions discussed in Section 5.9). These estimands can reflect relevant research and policy quantities if the causal effect arising from the imposition of the cut-off is of research interest.

A major strength of the RDD approach is that some of its assumptions can be falsified (though not verified). Several tests present themselves to the budding RDD analyst. First, similar to RCTs, balance in baseline (observed) covariates around the cut-off can be examined. Second, models can be repeated using placebo cut-offs. These cut-offs should not be related to jumps in the probability of treatment and so any association

would be evidence that the model is misspecified or its assumptions are violated. For instance, one could use an imaginary decision rule for elections where the 'winning' candidate wins if they receive 20% more votes than their closest rival or imagine the school leaving age was raised in 1971 or 1973 instead of 1972.

RDD is not without its limitations, however. First, as intimated, the relationship between the running variable and the outcome needs to be modelled appropriately; mismodelling can lead to discontinuities appearing which aren't there in practice, or vice versa (Gelman & Imbens, 2019; Huntington-Klein, 2022). A key issue with sharp RDD is that there is no 'common support' – i.e., treated and untreated individuals do not overlap in their value of the running variable. Counterfactual values therefore have to be extrapolated, rather than directly drawn from observed data from those in the comparator group.

A second limitation is that the same threshold may be used in multiple systems, analogous to violations of the exclusion restriction in the IV design (Section 5.9). For instance, in the UK, turning eighteen allows one to buy alcohol legally but also to gamble and buy cigarettes, each of which could affect health outcomes. Similarly, being the oldest in one's school year may increase academic confidence, but also physical maturity and social confidence. Each might impact educational attainment.

A third limitation of RDD is that, as regression discontinuities are most frequently found in administrative systems, the 'treatments' one can examine are limited by those implemented by these systems. These treatments may be relatively narrow (Bann et al., 2024). For instance, just meeting eligibility criteria for more generous welfare payments may be plausibly exogenous, but the total effect of these payments on income or wealth may be smaller, shorter-lived, and easier to compensate for than changes in income observed more generally and which are of greater interest (e.g., in social epidemiology top and bottom quintiles of income are often compared).

Finally, and most subtly, individuals, or the people acting on their behalf, may be aware of cut-offs and their implications and may act to ensure they are eligible for the treatment – or ineligible, if the treatment is undesirable. For example, some individuals who are close to being classified as long-term unemployed might accept a job offer

more readily in light of looming cuts to benefits (Card et al., 2007). Businesses that would be subject to increased regulation if they grow might opt against hiring an extra employee (Garicano et al., 2016). In these cases, there may be confounding even close to the cut-off; checking the distribution of covariates and density of observations either side can help assess this. Another subtle issue is misreporting or other fraudulent behaviour to meet eligibility criteria.[34] In this case, there will be differential measurement error in the running variable.

5.10.2 Applications

Several studies have applied RDD to CLS's cohorts. One popular object of study in the RDD literature is the 1972 Raising of the School Leaving Age (ROSLA, discussed above). In fact, the UK (or rather its constituent countries) has undergone several ROSLAs, including in 1947 and more recently in 2013 and 2015, with the raising of the minimum 'participation age' in education or training. As discussed in Section 5.8.2, though there is little variation in birth dates among CLS cohort members, there is plenty among their family members, including their parents. Sabates & Duckworth (2010) exploit the 1947 ROSLA and use data from the NCDS to assess the effect of increased maternal education upon cohort members' developmental outcomes, finding positive effects on maths ability but no effects on reading or behaviour. One drawback of this approach is that studying parental effects within a birth cohort necessarily conditions upon fertility and the parent's age at birth. It is possible that both are influenced by parental education.

Another discontinuity often exploited in education research involves test scores as these are widely used to determine entry into specific institutions (e.g., minimum A-Level grades for acceptance into top universities or degree requirements for entry into competitive graduate schemes).[35] In the UK, entry into selective grammar schools is largely based on results on the 'eleven-plus', a test similar to an IQ test. Pastore & Jones (2023) use data from cognitive tests at age 11 years in the NCDS to proxy for eleven-plus

[34] In stellar work, Saul Newman (2024, 2025) catalogues evidence that cases of extreme old ages are more likely to reflect pension fraud.
[35] In fact, the original RDD analysis used this design (Thistlethwaite & Campbell, 2017, Reprint).

scores to assess the effect of attending grammar school on several educational, labour market, and health outcomes. However, an issue is that, as they only use a proxy measure and (unobserved) qualifying eleven-plus scores for grammar school varied across England and Wales, they have to estimate the threshold determining grammar school entry from the NCDS data themselves. They do, however, perform several 'negative control' tests (Section 5.15), which support a causal interpretation (e.g., no association between grammar school attendance at the discontinuity and cognitive ability scores at age 7 years).

5.10.3 Opportunities

We have noted that discontinuities most frequently arise when an individual interacts with a (government) service. Specific criteria are frequently used to determine eligibility for services or duties (e.g., for income support) and hard geographical boundaries determine the bodies under whose purview one sits – for instance, residents of Bassetlaw in Nottinghamshire are eligible for three rounds of IVF, whereas their neighbours in surrounding districts (including across the same street, in some cases) are only eligible for one (Patel & Sissons, 2024). Accordingly, geographic and administrative data linkages have the potential to be particularly useful for performing RDD in CLS's cohorts. For instance, educational administrative data linkages may enable examination of the effects of educational access or attainment on various outcomes – universities offer places based on letter grades, though the underlying exam scores are continuous. An additional advantage of geographic (or other) variation is that some policies may be only rolled out in specific areas or groups over time. In this case, temporal variation could be used to identify causal effects with unaffected groups used for comparison – for instance, eligibility for the latest increase in the education participation age depends on birth date but has only been implemented in England and not Scotland or Wales. An issue with performing RDD in the cohorts is that sample sizes may be relatively limited around a cut-off, and so parametric methods – or pooling data across cohorts and with other data sources – may be necessary.

5.11 Target Trial Emulation

5.11.1 Method

So far, we have mainly discussed the benefit of randomised controlled trials (RCTs) that comes from randomisation. But RCTs have additional features that are beneficial for causal inference (Hernán et al., 2025). Specifically, they encourage the individuals involved in designing the RCT to ask well-defined causal questions by specifying study design and analysis plans that realise these questions as *estimands* (quantities one wishes to estimate); as discussed in Section 2.1, clearly stating the estimand is a key component of any good causal research.

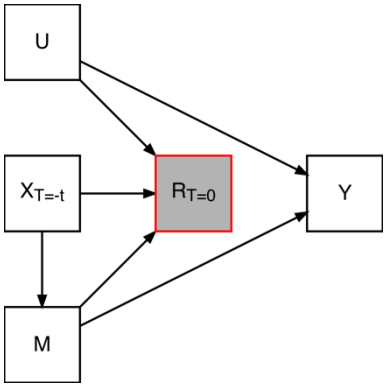
RCT study designs detail the eligibility criteria for the study participants, the treatment strategies that will be compared, the method of treatment assignment, the outcomes of interest, the start and end of follow-up, the causal contrast of interest, and the analysis plan. These components, when combined (and written well), enumerate the population, exposures and outcomes of interest for which a causal comparison is to be made. For instance, an RCT of COVID-19 vaccination on mortality may include individuals if they have not had a vaccine or COVID-19 infection so far, compare two dosages of active vaccine vs. placebo, assigned randomly, with mortality rates determined five years post-randomisation. Comparing survival rates in vaccination vs. placebo groups would provide an intention-to-treat estimate of the effect of the vaccine on mortality assuming no differential loss to follow-up.

Not every failure of observational studies arises due to problems with the underlying process that generated the data (e.g., confounding, measurement, and response biases). Some are "self-inflicted" due to decisions that are made by the data analyst in the design of their study (Hernán et al., 2016). Notable among these are decisions relating to eligibility, treatment assignment, and the definition used to begin follow-up time ('time zero'). Where these do not coincide, there is the potential for one of two related but distinct forms of selection bias: immortal time bias and prevalent (or 'healthy') user bias (Hernán et al., 2016).

Immortal time bias occurs when treatment assignment is determined *after* the start of follow-up (*time zero*). Most commonly, this occurs whenever the treatment is defined in relation to a minimum amount of time, e.g., if participants taking anticoagulants only count as 'treated' when they have been taking the medication for at least 5 days. Anyone who dies between the start of treatment and the designated time required to achieve 'treated' status will not be eligible for classification as 'treated', making this time 'immortal' as no deaths (or other outcomes of interest) can occur in this group during this time. In contrast, the 'untreated' group does not require a minimum amount of time to count as 'untreated' and all events occurring during this period are counted as normal. Since the 'treated' group has already survived a minimum required period that does not apply to the 'untreated' group, any apparent benefit of treatment may be overestimated as an artefact of this immortal time.

Prevalent user bias (also known as 'depletion of susceptibles') arises when treatment assignment occurs *before* the start of follow-up. Most commonly, this occurs in observational settings where comparisons are made between 'exposed' ('treated') and 'unexposed' ('untreated') individuals (e.g., those with and without heart disease). In this setting, the 'treated' must remain 'treated' long enough to make it into the sample at the start of follow-up. Any events occurring due to treatment before the start of follow-up would be missed and undercounted in the 'treated' group. For instance, when studying the effect of unemployment on depression, those whose mental health deteriorates soon after job loss may make greater effort to find a job quickly or may leave the labour force entirely. The individuals observed as unemployed at the start of follow-up (rather than the start of unemployment), then, would be disproportionately made up of those who are unaffected by unemployment's deleterious effects (Figure 5.14).

Figure 5.14: Causal DAG representing prevalent user bias in an observational study where treatment assignment is determined prior to follow-up time (time zero; T=0). In the DAG, exposure is measured at time-zero, but exposure can be initiated before this point ($X_{T=-t}$). Remaining exposed ($R_{T=0}$), which is implicitly conditioned upon in the study design, may be related to confounders as well as mediators on the path between $X_{T=-t}$ and Y.



The more harmful the exposure is and the longer the time between exposure initiation and start of follow-up, the more likely it is for those susceptible to the outcome to be depleted from the treated group sample. Prevalent user bias is a particular challenge with cohort data because of the length of time between data collections – participants may be more likely to experience a harmful exposure and then experience the outcome, or drop out of the study, before their exposure status is collected at the next wave. Note, prevalent user bias can affect both internal and external validity as (regarding internal validity) remaining exposed is a potential collider variable (Figure 5.14) and (regarding external validity) treatment effects may vary depending on 'susceptibility'.

These self-imposed analytical mistakes have generated several notable failures in observational research. For instance, Hernán et al. (2008) show that the supposedly protective effect of hormone replacement therapy on coronary heart disease, identified in observational studies but not RCTs (which found the opposite), disappears when prevalent user bias is avoided. It is thought that those most susceptible to coronary events after HRT initiation will have been depleted from the sample before the start of follow-up, meaning the protective effect seen in many observational studies is due to prevalent user bias. Dickerman et al. (2019) show the apparent protective effect of statins against cancer arises due to a combination of immortal time bias and prevalent user bias. First, existing studies compared long-term statin users with non-users, where

long-term users were defined as those with at least 4 years of statin use. Those who go on to be classified as long-term users must have therefore survived during this period without developing cancer (introducing immortal time bias). In addition, existing studies also included those already on statins before the start of follow-up, which may have led to depletion of susceptibles (introducing prevalent user bias).

The target trial emulation (TTE) framework (Hernán & Robins, 2016) specifically attempts to avoid such design biases by emulating the features of an RCT.[36] In particular, TTE involves specifying the protocol for a hypothetical ('target') trial, including eligibility criteria, treatment assignment procedures, follow-up, outcome, estimand, and analysis plan.[37] This protocol is then used to design a second protocol, detailing the design of the observational study which will be used to emulate the hypothetical target trial (Hernán et al., 2025). This mapping is iterative and based on what is possible with the data – it does not recreate an ideal RCT that one may wish had been run, and some features of an RCT, such as treatment allocation blinding and explicit randomisation, will not be possible to emulate (i.e., the target trial may often be a 'pragmatic trial'). Preparing data for a target trial emulation can involve significant data wrangling (e.g., restricting to an eligible sample). For instance, to assess the effect of job loss on mental health, data may be restricted to those in continuous employment before exposure 'assignment'.

The TTE does not map to a single statistical method – instead it is a framework for designing studies and defining estimands. As such it can be used alongside – or rather, specify the use of – any of the methods (e.g., regression adjustment), study designs (e.g., Regression Discontinuity Designs), and falsification tests (e.g., negative controls) discussed elsewhere in this document. Just like an RCT, TTE can be used to estimate a range of estimands. In an RCT, the intention-to-treat effect estimates the effect of being *assigned* to a treatment, regardless of adherence. Typically, the observational analogue

[36] TTE is sometimes referred to as the pseudo-trial design.

[37] Note, as the allocation procedure in observational data is not randomisation, the target trial protocol sets out the steps designed to estimate the target estimand (e.g., regression adjustment or use of an IV).

of the ITT is the effect of *initiating* a treatment, regardless of adherence during follow-up. In an RCT, there may also be a range of estimands of interest which involve being assigned to treatment *and* adhering to it for the duration of follow-up (traditionally referred to as 'per protocol' effects). However, just like an RCT, there is no single 'per protocol' effect and the exact adherence strategy and therefore estimand must be pre-specified. This requires appropriately accounting for post-treatment confounding and selection, including any time-varying confounding. This is discussed further in Section 5.12.

TTE is popular in pharmacoepidemiology and with large-scale health record data, but it is increasingly used in other fields, including social epidemiology and criminology (see, e.g., Cornish & Brennan, 2025; Matilla-Santander et al., 2023). It has also been applied in cohort data, including the CLS cohorts (see below). Cohort data have the advantage that variables are measured repeatedly which can make it possible to restrict eligibility to individuals without exposure prior to time zero, though as mentioned previously, the time gap between data collections can cause issues. Repeated measurement also allows for a nested trial design, where individuals are entered into multiple successive 'trials' until they become exposed.

A final benefit of the RCT design that is naturally encouraged by the TTE framework is specifying a well-defined intervention that can, hypothetically, be assigned experimentally. In observational studies, it is often unclear what the equivalent intervention may be, and hence the interpretation of the causal effect one is actually estimating (Hernán & Robins, 2023). For instance, a researcher may wish to estimate the effect of body mass index (BMI) on stroke. However, it is not straightforward to 'assign' a certain BMI, even in a trial. Instead, different levels of BMI can be achieved via numerous interventions (e.g., diet, exercise, Ozempic, amputating a limb). The same change in the exposure (e.g., transitioning from BMI of 30 to BMI of 25) can therefore be achieved through different means, each of which may have distinct effects on the outcome (stroke), making the estimand (the causal effect being estimated) unclear. Since the effect will be a mixture of different possible interventions, it has limited practical utility and limited transportability. This problem is much more easily avoided when we attempt to conceptualise the exposure as a hypothetical intervention as it

explicitly encourages us to think about well-defined causal contrasts with clearer interpretations.

5.11.2 Applications

Vagni (2026) uses data from BCS70 to estimate the causal effect of becoming a mother upon social class (the 'Motherhood Class Penalty'). The author uses an explicit TTE framework to define the exposure, outcome, eligibility criteria, time zero, and the estimand of interest. Comparator women (women who remain childless at age 46) are identified by matching upon Mahalanobis distance with a novel use of sequence analysis (Gabadinho & Studer, 2010) to identify pairs of treated and untreated individuals with similar pre-motherhood class trajectories. The author finds that motherhood has a large effect on class mobility, with mothers facing long-term lower likelihood of attaining or remaining in professional occupations following childbirth. Strengths of the study include: the clear and transparent statement of the study design, estimands, and assumptions; the use of external data (Annual Population Survey and British Household Panel Survey) to assess potential bias from attrition; and the use of data visualisations to display the results and provide compelling evidence that estimates are causal (e.g., showing estimated differences in occupational class are confined to the peri- and post-childbirth period).

5.11.3 Opportunities

The TTE approach is relatively underutilised in CLS cohort data, despite the cohorts having several features making the approach tractable (as noted above), including longitudinal assessment to determine eligibility and to implement nested trial designs to increase statistical power.

5.12 G-methods for Time-Varying Exposures

5.12.1 Method

Most of the methods outlined so far focus on the effect of exposures occurring at a single timepoint. In many applications, however, researchers are interested in how the effect of the same exposure depends on when it occurs or in how exposure sustained or

accumulated over time influences later outcomes. CLS's cohorts are particularly well suited to answering such questions. However, time-varying exposures introduce additional challenges for causal inference, some of which cannot be addressed by conventional statistical methods (e.g., ordinary least squares regression). In this section, we outline these challenges and briefly discuss potential solutions.

Before turning to time-varying confounding – which was briefly introduced in Section 2.2.3 – it is important to note that the problems and methods discussed here apply specifically to causal questions about the effects of exposure patterns over time. This is more specific than simply studying an exposure that may vary over time, since such exposures may still be analysed as time-fixed depending on the causal question. For example, for some target population, one might ask about the effect of high versus low depressive symptoms at age 16 on blood HbA1c concentration at age 30, a biomarker of diabetes. (For simplicity, we treat the exposure as binary and set aside concerns about consistency – that is, we assume "high" and "low" depressive symptoms are 'sufficiently well-defined'.) The causal question is valid, but it concerns exposure at a single age and its total effect on later HbA1c through all subsequent pathways. Although depressive symptoms and other influences may change between ages 16 and 30, the causal contrast is defined only in terms of symptoms at age 16. From an interventionist perspective, this corresponds to asking what the average HbA1c concentration at age 30 would be under hypothetical interventions that set everyone's depressive symptoms to high vs. low levels at age 16, while allowing later symptoms to evolve naturally thereafter. This matches the definition of the average treatment effect as provided in Section 2.2.1, and its estimation does not require methods for time-varying exposures that we discuss in this section.

Alternatively, a researcher may be interested in the effects of different exposure patterns over time. We will refer to these as 'regimes'. Extending the previous example, one might ask about the effect on HbA1c at age 30 of experiencing the treatment regime of high levels of depressive symptoms at age 16 and again at age 24, compared with the regime of experiencing low levels of depressive symptoms at both timepoints. This defines a different type of causal question. Here (again using an interventionist perspective), we are interested in the joint effect of intervening twice: first in

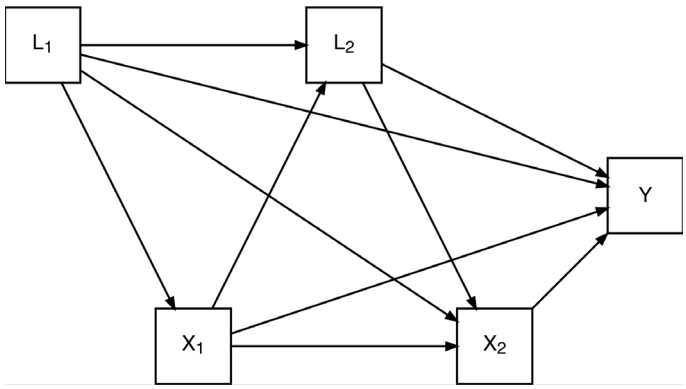
adolescence and again in early adulthood. More formally, for this contrast, the average treatment effect could be expressed as $E[Y^{X_1=1,\,X_2=1}] - E[Y^{X_1=0,\,X_2=0}]$, where X_1 and X_2 denote exposure levels at the two timepoints. Other causal contrasts are also possible, depending on which exposure regimes are compared. For example, one could consider HbA1c levels under high depressive symptoms at age 16 but low symptoms at age 24, and vice versa. Naturally, the very same principle could be extended to more exposure timepoints, giving rise to an even larger number of possible contrasts.[38] For this reason, defining the causal contrast of interest becomes more challenging in time-varying settings. Despite this additional complexity, comparisons of different exposure regimes allow researchers to formalise pressing causal questions. They are especially relevant when the aim is to learn whether exposures have different effects at different stages of the life course, whether accumulated exposure has distinct effects, or which exposure regimes may have the greatest effect and as such be targeted in policy making.

When estimating the effects of exposure regimes, a new complication arises with respect to confounding. We illustrate this using the simplified example introduced earlier. Consider the DAG in Figure 5.15, which represents the setting introduced above. In this DAG, X_1 and X_2 represent the two exposure instances and L_1 and L_2 represent sets of time-varying confounders (e.g., one's overall health).[39] Suppose we wish to estimate the effect of high depressive symptoms at both timepoints, compared with low symptoms at both timepoints, on later HbA1c. The graph is best read sequentially, from left to right over time. At the first exposure timepoint, the situation looks familiar: to estimate the effect of depressive symptoms at age 16 (X_1) on HbA1c at age 30 (Y), we would wish to condition on common causes of X_1 and Y – in this case, L_1. The situation is also familiar at the second timepoint: to estimate the effect of depressive symptoms at age 24 (X_2) on HbA1c at age 30 (Y), we would wish to condition on common causes of X_2

[38] The same principles also apply when one is interested in the history of different exposures, rather than repeated measures of the same exposure. For example, one might be interested in the joint effect of parental income in childhood and unemployment in adulthood on later health outcomes.
[39] Time-invariant confounders could also have been added to this DAG, but weren't for readability.

and Y, including prior exposure (X_1), and the time-varying confounders taking place before the second exposure (L_1 and L_2).

Figure 5.15: A DAG representing a simple setting with a time-varying exposure at two timepoints (X_1 and X_2). L_1 and L_2 are time-varying confounders. To estimate the effect of $X_2 \rightarrow Y$, one must condition on the full history of the covariates and the exposure, including L_1, L_2, and X_1. To estimate the effect of $X_1 \rightarrow Y$, one needs to condition on L_1, but must <u>not</u> condition on L_2. If one wishes to estimate the joint effect X_1 and X_2 have on Y, a problem arises, because the two instances require mutually exclusive adjustment sets (that is, one needs L_2 to be conditioned upon, whereas the other needs that it is not conditioned upon). Standard estimation methods cannot accommodate this structure, and g-methods are required.



So, we have identified two different, though partly overlapping, sets of confounders, one for X_1–Y and the other for X_2–Y. If we were to estimate these effects separately, we could simply fit two separate models, each focusing on one exposure instance and conditioning on the respective set of confounders. However, we wish to estimate an effect that compares different exposure regimes, so X_1 and X_2 must be considered jointly. Consequently, confounding relevant to both the earlier and later exposure–outcome relationships must be addressed when estimating that joint effect. The difficulty arises when confounders of the later exposure–outcome relationship (here L_2) are themselves affected by the earlier exposure. This scenario is known as treatment-confounder feedback and confounders affected by past exposure levels are referred to as intermediate confounders. In our example, health status in early adulthood could be an intermediate confounder: it may affect both later depressive symptoms and later HbA1c but may also itself be affected by earlier depressive symptoms. Intermediate confounders have a dual role: they confound the later exposure–outcome relationship

($X_2 \leftarrow L_2 \rightarrow Y$) but also lie on the causal pathway from the earlier exposure to the outcome ($X_1 \rightarrow L_2 \rightarrow Y$).

In the presence of treatment-confounder feedback, conventional methods will generally fail to estimate the effects of exposure regimes correctly, even when all relevant confounders are measured. To see why, consider what the dual role of intermediate confounders implies for the adjustment set. To address confounding of the X_2–Y relationship, L_2 must be conditioned upon. But for the X_1–Y relationship, conditioning on L_2 would block part of the earlier exposure effect. Standard methods therefore face a problem with no satisfactory solution: conditioning on L_2 removes part of the causal effect X_1 has on Y, whereas **not** conditioning on L_2 leaves the X_2–Y relationship confounded.[40]

To address this problem, a family of methods has been developed, which are collectively referred to as *g-methods* (the letter *g* stands for 'generalised'). The three main approaches are: g-computation, also known as the parametric g-formula (Robins, 1986); inverse probability of treatment weighting (IPTW; Robins et al., 2000), already encountered for single time points in Section 5.3; and g-estimation of structural nested models (Robins, 1989; Vansteelandt & Sjölander, 2016). The naming convention can cause some confusion. IPTW belongs in the g-methods family despite containing no signposting letter g, while g-estimation is, despite its name, a method that is fully distinct from g-computation. Nevertheless, all three methods share the ability to estimate effects of exposure regimes even in the presence of treatment-confounder feedback.[41] While discussing the details of the three g-methods is beyond the scope of this work, we briefly outline the intuition behind g-computation and IPTW. We do not discuss g-estimation further, as it would require additional technical context not well

[40] Note, the software DAGitty is less developed for handling time-varying exposures and time-dependent confounding.

[41] Apart from g-methods, related developments have also emerged within the structural equation modelling (SEM) literature (De Stavola et al., 2015), though the focus is on SEM as a parametric framework for causal modelling rather than on the broader range of SEM applications familiar from applied research.

suited to the brief treatment possible here. Readers interested in more detail on all three g-methods are referred to Hernán & Robins (2023).

In the most common implementation, g-computation uses a series of regression models to simulate, step by step and forward in time, what each person's intermediate confounders and outcome would be under a particular exposure regime, given the covariate history up to the respective timepoint. Using these simulations, one could first estimate, for example, the expected outcome under a regime in which everyone was exposed at each timepoint. Next, the process is repeated under a different regime (e.g., no one exposed at any timepoint). The simulated outcomes are averaged under each regime, and the resulting means can be compared, for example by taking the mean difference.

This approach naturally accommodates the dual role of intermediate confounders. As the simulations are carried out forward in time under each exposure regime, they preserve the influence of earlier exposures on later intermediate confounders and their resulting effect on the outcome. The final comparison is made directly between the estimated mean potential outcomes under different regimes, so intermediate confounders contribute to the differences in the estimated mean potential outcomes only insofar as they are changed by the exposure regime. The resulting comparison can be interpreted as the estimate of the respective causal effect, provided that the identifying assumptions hold (discussed below) and that all models for the intermediate confounders and outcome are correctly specified.

Unlike the usual implementation of g-computation, IPTW does not attempt to simulate counterfactual trajectories. Recall from Section 5.3 that, in the single-exposure setting, the weights are defined as the inverse of the probability of receiving the observed exposure level, conditional on confounders. Reweighting observations by these weights aims to create a pseudo-population in which the exposure is independent of the measured confounders. For time-varying exposures, IPTW extends this same principle: if successfully implemented, it reweights observations so that, at each exposure timepoint, the exposure is independent of the covariate history (conditional on past exposure levels). In its simplest form, weights can be derived separately for each exposure instance, for example using regression models for the exposure conditional on

the covariate and exposure history up to that point. These timepoint-specific weights are then multiplied across timepoints to give a single weight for each individual. The resulting weights can then be used to estimate the effects of exposure regimes, which can be obtained through methods such as weighted regression. For a simple worked example, see Thoemmes & Ong (2016).

IPTW generally relies on the same identifying assumptions as g-computation, but the two methods have different requirements for statistical modelling. The standard version of g-computation requires a correctly specified model for each intermediate confounder as well as the outcome, whereas the implementation of IPTW described here requires a correctly specified model for each exposure instance. In most analyses, there are many more potential confounders than there are exposure instances, so IPTW typically has the benefit of making fewer parametric assumptions than g-computation. That said, this may come at a cost of lower efficiency of IPTW, producing wider confidence intervals. In practice, IPTW can also be more prone than g-computation to numerical instability, especially in settings with many timepoints and complex covariate histories, where extreme weights are more likely to arise. Importantly, this comparison applies to the standard versions of the two methods, but other implementations can be used to relax some of the modelling assumptions (see, for example, Huffman & van Gameren, 2018; Karim et al., 2017; Zivich et al., 2024). The strengths of the two methods can also be combined in doubly robust estimators (Bang & Robins, 2005; van der Laan & Gruber, 2012). The flexible and doubly robust approaches typically come at the cost of greater complexity.

While progress has been made in statistical methods, identification of exposure regime effects relies on strong assumptions. The methods outlined above generally require a sequential version of the conditional exchangeability assumption. In practical terms, this typically implies that, at each exposure timepoint, one must account for common

causes of the exposure and the outcome.[42] Even when only a single exposure instance is considered, it is rare that all confounding has been fully accounted for, and some bias due to unmeasured confounding may remain. As additional timepoints are introduced, the number of confounders can only increase, so this assumption becomes more difficult to justify. Other identifying assumptions, including positivity, consistency, and no interference, also generalise to sequential versions and may likewise become harder to satisfy (see Hernán & Robins, 2023, for details).

The strong identifying assumptions, together with the added analytical complexity, have an important practical implication. When many exposure timepoints are available, researchers may be tempted to define exposure regimes that involve all of these instances, hoping to make full use of the available information. Yet it may be preferable to focus on fewer exposure timepoints than are available in the data, so that the required assumptions remain more plausible. For example, one should consider the availability of confounders at or before each exposure timepoint. Another reason for limiting the number of timepoints is so that the causal question becomes clearer. As additional exposure occasions are introduced, the number of possible exposure regimes increases rapidly, and it may become unclear which contrasts are substantively meaningful. Therefore, it may be more informative as well as practical to focus on a few timepoints that correspond to meaningful developmental periods or plausible intervention points, rather than modelling every observed measurement occasion.

5.12.2 Applications

Hope et al. (2019) use MCS to examine the effect of time-varying maternal distress upon childhood mental health problems across ages 3, 5, 7, and 11 years. Using marginal structural models and allowing for treatment-confounder feedback (prior child mental health influencing maternal distress) and time-varying confounding via family

[42] More precisely, the exact form of sequential conditional exchangeability depends on the exposure regime under consideration. The implication stated in the main text generally holds for static regimes, in which exposure is set according to a fixed sequence that does not depend on covariate history. Hernán & Robins (2023) discuss more complex regimes for which this assumption may need to be formulated differently, and in some settings more stringently.

circumstances, they show that maternal distress increases the risk of future mental health problems in children, even net of future maternal distress. A strength of the study is the use of QBA (Section 5.20) to examine sensitivity of the results to unobserved confounding, showing that – according to their reading of the literature – implausibly strong unobserved confounding would be necessary to explain associations.

Pinto Pereira et al. (2020) use data from the NCDS and the National Survey of Health and Development (NSHD, the 1946 British Birth Cohort Study; for a cohort profile, see Kuh et al., 2011; Wadsworth et al., 2006) to assess the causal effect of early- and mid-adulthood obesity upon physical functioning in mid-life. Using the parametric g-formula and an explicit counterfactual framework, the authors compare the causal effect of multiple treatment regimes: never obese, persistent obesity, and obesity beginning at particular ages. Their results suggest that obesity increases the risk of low physical functioning in mid-life, with greater effects for obesity that began earlier in the life course. The authors also perform a causal mediation analysis; we delay discussion of this until Section 5.13. Strengths of the study include: sharing of analytical code, the reduction of a complex exposure history to a small number of substantively interpretable regimes, and explicit discussion of study estimands, the assumptions upon which these are based, and the causal interpretation that can be made, subject to the assumptions being met.

5.12.3 Opportunities

Given repeated measurement of many factors across the life course, there is ample opportunity to assess the effects of time-varying exposures in CLS's cohorts, as well as to include time-varying confounders in such analyses. Further, unlike many other cohort studies, data collection in CLS cohorts began at birth or early in life. This allows for testing of a wider range of life course theories, but also averts an issue with cohorts that have data collection starting later. Specifically, such cohorts implicitly assume that exposure begins at the point at which variables are measured (e.g., at recruitment in adulthood), though in practice exposure would likely begin before then (Gilsanz et al., 2022, term this late-study start confounding).

As discussed, a key limitation is the age gap that can appear between consecutive sweeps – for instance, ten years elapse between the age 23 and 33 year sweeps in the NCDS. This could warrant focusing upon life course stages where measurement was more frequent (i.e., from age 42 years onward in the NCDS), or upon MCS and Next Steps, whose earlier sweeps are more closely spaced. Data linkages offer a further route around the problem, as administrative records may capture relevant information more frequently than survey sweeps.

To our knowledge, most existing applications of g-methods in CLS's cohorts have been causal mediation analyses (Section 5.13) rather than comparisons of exposure regimes. Exposures that are measured repeatedly but to date have been studied largely with descriptive trajectory methods – e.g., on poverty spells, parental mental health, employment histories, housing tenure, area deprivation – could be fruitfully reassessed within a counterfactual framework.

5.13 Causal Mediation

Note: This section builds on the concepts of joint exposure effects and treatment-confounder feedback (intermediate confounding). Readers unfamiliar with these concepts are strongly encouraged to read Section 5.12 on G-methods for Time-Varying Exposures first.

5.13.1 Method

In the previous sections, we have considered settings united by a common motivation: to understand what would happen to some (population-average) outcome if the exposure were changed, whether at a single time point or across multiple time points. Such questions concern a particular end result of interest under specified (hypothetical) interventions on the exposure and are typically agnostic about the processes set in motion by those interventions. For example, one may learn that, in some target population, prescribing one type of medication reduces the five-year risk of developing heart disease by 20%, compared to an alternative prescription. This may be sufficient to conclude that prescribing the medication is beneficial over the alternative in preventing heart disease, but it does not explain how that benefit is produced. For

instance, the medication could be more effective because patients are more likely to take it than the alternative, or because it acts on different physiological pathways involved in heart disease, or because of some combination of these and other pathways. Questions about intermediate pathways linking the exposure to the outcome have motivated the development of mediation analysis.

Besides being of inherent scientific interest, understanding the mechanisms by which a causal effect is produced can provide several real-world benefits. First, there may be no known way to intervene on the exposure itself in the real world, or such interventions may be costly, impractical, or otherwise undesirable. Downstream processes may be more feasible to intervene on and provide alternative levers for producing the desired causal effects. For instance, policymakers may not have latitude to fully prevent adverse childhood experiences (primary prevention), but do have more scope to implement interventions to avert some of their negative psychological sequelae (secondary prevention; e.g., talking therapies; Ning et al., 2024). Second, understanding intermediate processes may help refine interventions. Rather than treating an intervention as a black box, researchers may wish to know which components, behaviours, or biological changes are most important for producing the outcome of interest. For example, if obtaining a degree increases wages simply by signalling human capital, rather than increasing it, a policy goal of maximising university attendance might be withdrawn. Third, identifying intermediate pathways can help inform tailoring: different pathways may be more relevant in different populations, settings, or stages of life. For instance, if the effect of job loss on mental health operates via financial strain, resources could be focused on those without savings or other sources of income.

In general, mediation analysis has often been understood as an attempt to answer such questions by decomposing the total effect of an exposure on an outcome into direct and indirect components. In the simplest setting, an exposure X is thought to affect an outcome Y, at least partly through an intermediate variable M, referred to as the mediator. The indirect effect is then understood as the part of the effect of X on Y that operates through M, while the direct effect is understood as the part that operates through paths that do not involve M.

To make this discussion more concrete, we return to the example introduced in Section 5.12, with some changes. Suppose we are interested in the effect of high versus low depressive symptoms at age 16 (X), on HbA1c at age 30 (Y), a biomarker of glycaemic control. We suspect that depressive symptoms may also influence health behaviours, including smoking, and smoking may in turn affect HbA1c. We may therefore be interested in whether the effect of depressive symptoms on HbA1c operates, at least in part, through smoking frequency (M), measured at age 24. As in the previous section, we set aside consistency issues for simplicity, assuming that both "high" and "low" depressive symptoms and smoking frequency are sufficiently well-defined.

Although the motivation for mediation analysis may seem clear at first glance, translating a general interest in "understanding mechanisms" or the notion of "direct" and "indirect" paths into well-defined causal effects is challenging. Traditional approaches to mediation developed from path analysis (S. Wright, 1934), an early framework in which hypothesised relationships between variables are represented through a system of equations (this was later generalised within structural equation modelling, SEM). These approaches to mediation define effects through statistical models and make no reference to formal causal frameworks. While a full description of traditional mediation frameworks is beyond the scope of this work, we provide a brief illustration of the simplest approach, where the three effects are defined through linear regression coefficients (Baron & Kenny, 1986; MacKinnon et al., 2007):

1. **Total effect:** The first step is to regress the outcome on the exposure and a baseline set of covariates $C_1, C_2, \ldots, C_n$. In our example, this is done by simply running a linear regression model with HbA1c as the outcome, depressive symptoms as the exposure, and including some baseline covariates, giving $Y = \alpha_0 + \alpha_1 \mathrm{X} + \alpha_2 \mathrm{C}_1 + \ldots + \alpha_{n+1} \mathrm{C}_n + \varepsilon_1$. The coefficient for the exposure α_1 represents the total effect, which, in our example, gives the total effect of depressive symptoms on HbA1c.
2. **Direct effect:** The outcome model is refitted with the mediator (smoking frequency) also included: $Y = \beta_0 + \beta_1 \mathrm{X} + \beta_2 M + \beta_3 \mathrm{C}_1 + \ldots + \beta_{n+2} \mathrm{C}_n + \varepsilon_2$. The exposure coefficient β_1 estimates the direct effect of the exposure on the outcome. The idea underpinning this step is that by 'controlling' for M, we hold it

constant, so the regression coefficient for the exposure gives the X–Y association that is independent of M.

3. **Indirect effect:** The mediator is regressed on the exposure and covariates: $M = \gamma_0 + \gamma_1 \mathrm{X} + \gamma_2 \mathrm{C}_1 + \ldots + \gamma_{\mathrm{n+1}}\, \mathrm{C}_n + \varepsilon_3$. Following the path-tracing rule, the effect along a path is obtained by multiplying the coefficients that make up that path. Therefore, the indirect effect can be defined as $\gamma_1 \cdot \beta_2$ (because γ_1 represents the X–M path and β_2 represents the M–Y path).

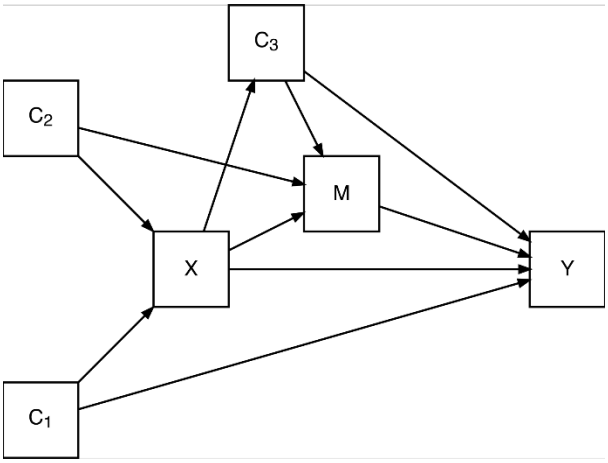
We have outlined the simplest regression-based approach to mediation, but many extensions have been developed, including mediation in traditional SEM, longitudinal mediation models, multilevel mediation, and others (see MacKinnon, 2008, for an overview). However, greater modelling complexity does not solve the central problem: these frameworks do not specify what these quantities represent outside the statistical models in which they are defined. It is therefore unclear what precise causal questions they answer and under what conditions a causal interpretation is justified.

Work using counterfactual frameworks shows the quantities only correspond to certain mediational causal effects under restrictive conditions. For example, regression-based definitions rely on a simple linear structure, and the usual decomposition breaks down in the presence of exposure-mediator interaction. Intermediate confounding (partly due to problems discussed in Section 5.12) and collider bias (Section 2.2.4) can also create issues. Formal causal frameworks are therefore needed both to define the causal estimands and to clarify the conditions under which they coincide with these regression-based definitions (De Stavola et al., 2015; VanderWeele & Vansteelandt, 2009).

Despite what the term might suggest, causal mediation analysis is not a single analytical method. Instead, it can be understood as an umbrella term for analyses that explicitly target direct and/or indirect effects defined through counterfactual frameworks. Because many formulations of these effects exist, we briefly outline some of the most widely used options and discuss their assumptions, focusing primarily on issues of confounding.

We begin with an illustrative causal diagram in Figure 5.16. In this diagram, X influences Y both through M and through other, unspecified pathways that do not involve M. For pedagogical clarity, we depict three distinct sets of confounders that may be relevant depending on the causal effect of interest: C_1 for the exposure–outcome relationship, C_2 for the exposure–mediator relationship, and C_3 for the mediator–outcome relationship. Notice that C_3 is influenced by X, making it a set of intermediate confounders. This separation is a simplification, as the same variable may confound more than one of these relationships; that is, many pre-exposure confounders may influence X, M, and Y, and causal relationships also exist among the confounders themselves.

Figure 5.16: Illustrative DAG for causal mediation analysis, comprising exposure X, mediator M, and outcome Y. Depending on the causal effect of interest, three sources of confounding may need to be considered: confounding of the exposure–outcome relationship by C_1, the exposure–mediator relationship by C_2, and the mediator–outcome relationship by C_3.



As in traditional approaches, we begin by defining the total effect. The total effect captures the net effect of the exposure on the outcome, without separating the pathways through which this effect occurs. In the previous sections of this guide, we have already encountered such a quantity: the average treatment effect. In the context of mediation analysis, the average treatment effect is usually referred to as the total average causal effect, or just total effect (TE). For a binary exposure, this is the difference in mean potential outcome had everyone been exposed compared to no one having been exposed:

$$TE = ATE = E[Y^{X=1}] - E[Y^{X=0}]$$

In our example, this would be the difference in mean HbA1c had everyone experienced high depressive symptom levels, compared to everyone experiencing low depressive symptoms. We have already encountered the assumptions required for identifying this effect from data. Specifically: (conditional) exchangeability (i.e., by accounting for all common causes of the X–Y relationship), (conditional) positivity, consistency, and no interference.

Several attempts have been made at defining the direct effect. Intuitively, to conceptualise the effect "not through M", one might imagine changing the exposure while preventing any corresponding change in the mediator. The controlled direct effect (CDE) formalises this idea through joint hypothetical interventions on the exposure and mediator, where the exposure is changed while the mediator is held fixed at a specified value (Pearl, 2001; Robins & Greenland, 1992).

For example, one might compare mean HbA1c if everyone experienced high rather than low depressive symptoms at age 16, while preventing everyone from smoking at age 24 in both scenarios (thus setting everyone's smoking frequency to 0). Because smoking frequency is fixed at the same value under both exposure scenarios, any difference in mean HbA1c cannot operate through depression-induced changes in smoking frequency. More generally, the mediator need not be set to 0. A CDE can be defined at any fixed mediator value $M = m$:

$$CDE(m) = E[Y^{X=1,M=m}] - E[Y^{X=0,M=m}]$$

In this equation, $E[Y^{X=1,M=m}]$ denotes the mean potential outcome had everyone been exposed and had their mediator set to value m. This is compared with $E[Y^{X=0,M=m}]$, the mean potential outcome had no one been exposed while also having their mediator set to the same value m. The CDE can also be viewed as a special case of the joint exposure regime effects described in the previous section, with the mediator treated as a second exposure occurring after X.

To identify the CDE, we require the assumptions needed to identify the effect of the joint intervention on X and M. This includes all the assumptions invoked for the TE, but also additional assumptions concerning the mediator intervention. In particular, one must account for confounders of the mediator-outcome relationship (C_3) as well as

confounders of the exposure-outcome relationship (C_1). Much like the effects of exposure regimes, the estimation of CDEs is more complex in the presence of treatment-confounder feedback (i.e., when some members of C_3 are influenced by X and are therefore intermediate confounders). In that case, the CDE can still be identified if these confounders are measured and accounted for, but g-methods are typically required for its estimation. In addition to assumptions about confounding, identification requires sequential versions of positivity, consistency, and no interference, which are detailed elsewhere (Daniel & De Stavola, 2019).

While CDEs are appealing, a conceptual challenge is that researchers must select the value at which the mediator will be fixed. This matters in the presence of exposure-mediator interaction on the chosen effect measure scale. For example, suppose depression has a more detrimental effect on HbA1c among smokers, perhaps because depressive symptoms activate inflammatory pathways that amplify the metabolic consequences of smoking. The CDE would then differ according to the smoking frequency at which M is fixed. It may also not be obvious which value is most scientifically or practically relevant. A further consideration is that the CDE has no corresponding indirect effect, and therefore does not attempt to decompose the TE (VanderWeele & Vansteelandt, 2009). Nevertheless, CDEs can answer policy-relevant questions when fixing the mediator at the chosen value corresponds to a feasible intervention, such as whether intervening on X would provide any additional benefit once M had already been fixed through a separate intervention (see the example in T. Q. Nguyen et al., 2021).

Another related approach to conceptualising the direct effect is the natural direct effect (NDE, Pearl, 2001; Robins & Greenland, 1992). Earlier, we described the general idea behind a direct effect as a thought experiment in which the exposure is changed while any resulting change in the mediator is prevented. The NDE follows this intuition more literally: rather than setting everyone's M to the same arbitrary value, it sets each person's M to the value it would have naturally taken under non-exposure (or under another comparison exposure level for non-binary exposures).

In our example, this means comparing mean HbA1c under two hypothetical scenarios. In the first, everyone experiences high depressive symptoms at age 16, but each

person's smoking frequency at age 24 is set to the value it would have taken had they instead experienced low depressive symptoms. In the second, everyone experiences low depressive symptoms at age 16, and each person's smoking frequency is set to that same person-specific value, namely the value it would naturally have taken under low depressive symptoms. More generally, we write:

$$NDE = E\left[Y^{X=1,M=M^0}\right] - E\left[Y^{X=0,M=M^0}\right]$$

where M^0 denotes the mediator value under non-exposure. The first term $E\left[Y^{X=1,M=M^0}\right]$ is the mean potential outcome had each person been exposed while having their mediator set to its person-specific value under non-exposure. The second term $E\left[Y^{X=0,M=M^0}\right]$ is the mean potential outcome under non-exposure, with the mediator set to that same person-specific value. Since M^0 is precisely the value the mediator would naturally take under X = 0, this second term reduces to the mean potential outcome under non-exposure, $E[Y^{X=0}]$.

Importantly, the NDE has an indirect counterpart, the natural indirect effect (NIE). To understand the intuition behind the NIE, it is helpful to revisit the fundamental logic of mediation: the exposure shifts the mediator, and this exposure-induced shift in the mediator then affects the outcome. The NIE attempts to quantify the effect of this shift on the outcome. It starts from a scenario in which everyone is exposed, and then asks how the mean outcome would change if each person's mediator were shifted from the value it would have taken under non-exposure to the value it would have taken under exposure:

$$NIE = E\left[Y^{X=1,M=M^1}\right] - E\left[Y^{X=1,M=M^0}\right]$$

Here, $E\left[Y^{X=1,M=M^1}\right]$ is the mean potential outcome under exposure and mediator set to its corresponding value under exposure, and $E\left[Y^{X=1,M=M^0}\right]$ is the mean potential outcome under exposure while the mediator is set to its value under non-exposure. In our example, this would be the effect of changing each person's smoking frequency from the value it would have taken under low depressive symptoms to the value it would have taken under high depressive symptoms, while keeping depressive symptoms themselves high in both scenarios.

The NDE and NIE align with some of the intuition underpinning traditional mediation analysis. This is reflected in their mathematical properties; for example, they decompose the TE, so that $TE = NDE + NIE$ for effects on the difference scale. Importantly, the NDE and NIE are best considered thought experiments, as some of the hypothetical interventions they represent could not realistically be enacted in the real world.

$$\begin{aligned} NDE + NIE &= E\left[Y^{X=1,M=M^0}\right] - E\left[Y^{X=0,M=M^0}\right] + E\left[Y^{X=1,M=M^1}\right] - E\left[Y^{X=1,M=M^0}\right] \\ &= E\left[Y^{X=1,M=M^1}\right] - E\left[Y^{X=0,M=M^0}\right] \\ &= E[Y^{X=1}] - E[Y^{X=0}] \\ &= TE \end{aligned}$$

While some properties of the NDE and NIE may seem conceptually appealing, identification of these effects rests on strong assumptions that are implausible in many applied settings. Both the NDE and NIE involve what are referred to as cross-world counterfactuals – quantities that combine exposure and mediator values from different hypothetical scenarios that cannot be observed together for all individuals. For example, the NDE contains $E\left[Y^{X=1,M=M^0}\right]$, where each person's exposure is set to 1, but their mediator is set to the value it would have taken had their exposure been 0. This quantity cannot be observed for any individual whose level of the mediator differs according to the exposure, because we cannot simultaneously observe that person under both exposure and non-exposure. Identifying natural effects therefore requires extrapolating from observed data to these cross-world quantities. This relies on the assumption of cross-world independence. The general formulation of this assumption is beyond the scope of this guide, but in practice, the main implication of this assumption is a requirement of no intermediate confounding: natural effects cannot generally be identified when any mediator-outcome confounder is affected by the exposure.[43] The violation of this assumption is represented by an arrow from X to C_3 in

[43] The requirement of no intermediate confounding is the main practical implication of cross-world independence. However, cross-world independence is a broader assumption and is not identical to the requirement of no intermediate confounding. For technical details, see Andrews & Didelez (2021).

Figure 5.16. Even if all intermediate confounders are known and measured, natural effects are generally not identified, and thus they also cannot be estimated, even if g-methods are used.

Apart from the absence of intermediate confounding, identifying natural effects requires accounting for confounders of the exposure–outcome relationship (C_1), and the mediator–outcome relationship (C_3), much like identifying the CDE. In addition, one must also account for confounders of the exposure–mediator relationship, C_2. This is because the NDE and NIE require the conditional distribution of mediator values under exposure and non-exposure (i.e., reflecting the causal effect of the exposure on the mediator) to be recoverable from the observed data. Finally, identifying natural effects also requires extended versions of positivity, consistency, and no interference, which are detailed elsewhere (Daniel & De Stavola, 2019).

In applied research, the assumption of no intermediate confounding required for the standard identification of natural effects is often difficult to justify.[44] Therefore, in many settings, other formulations of the direct and indirect effect that do not make this assumption are more relevant. A popular approach is provided by (randomised) interventional analogues of natural effects, including the interventional direct and indirect effect (IDE, IIE; Didelez et al., 2006; VanderWeele et al., 2014). We provide only a brief introduction to these effects, but readers interested in further intuition and examples are referred to Nguyen et al. (2021).

The interventional analogues follow a similar logic to their natural counterparts but differ in an important way. Intuitively, the NDE and NIE involve personalised hypothetical interventions on the mediator: each person's mediator is set to the value it would naturally have taken *for them* under a particular exposure scenario.

---

[44] In some cases, cross-world independence may be replaced with additional parametric assumptions instead (Petersen et al., 2006), although these alternative assumptions may be equally implausible in many applied settings. Alternatively, one may forgo point identification and instead estimate bounds or run a sensitivity analysis to obtain a model-based range of plausible effects (Breum et al., 2025; Daniel et al., 2015). These approaches can be more demanding to implement and interpret compared to approaches relying on point identification, and they remain uncommon in applied research.

Interventional effects instead involve changing the distribution of the mediator, without requiring each person to receive their own person-specific value under that exposure level.

To define these effects, let $G_{x|C}$ be a random draw from the mediator distribution that would arise under exposure level $X = x$ in stratum $C = c$. Then, for the binary exposure, we define:

$$IDE = E[Y^{X=1,M=G_{0|C}}] - E[Y^{X=0,M=G_{0|C}}]$$

$$IIE = E[Y^{X=1,M=G_{1|C}}] - E[Y^{X=1,M=G_{0|C}}]$$

To illustrate, much like the NIE, the IIE starts from a scenario in which everyone is exposed. It then asks how the mean potential outcome would change if the mediator distribution shifted from that under non-exposure to the distribution that would arise under exposure, within strata of relevant covariates. The IDE instead keeps the mediator distribution at that which would arise under non-exposure (within strata of C) and compares the mean potential outcome when everyone's exposure is changed from 0 to 1.

Under the assumption of cross-world independence, natural effects equal their interventional analogues. When intermediate confounding is present, IDE and IIE can still be identified, as they do not require cross-world independence, but they will no longer coincide with natural effects and no longer decompose the TE.[45] Additionally, all other assumptions listed for natural effects are still required, and intermediate confounders must be measured and properly accounted for. It is also worth noting that interventional effects answer their own causal questions (Miles, 2023; Moreno-Betancur & Carlin, 2018), and this should be reflected when selecting the target effect and interpreting the findings. Although they may sometimes serve as approximations to natural effects, this correspondence does not hold in general, so using them as proxies should be explicitly justified (see Gervais et al., 2025) or at least noted as a limitation.

[45] Specifically, because $E[Y^{X=x,M=G_{x|C}}] \neq E[Y^{X=x}]$ when there is intermediate confounding, the IDE and IIE no longer sum to the TE.

Apart from interventional analogues of natural effects, other estimands have also been developed. While discussing all alternatives is beyond the scope of this guide, we note separable effects as an option gaining popularity especially in medical research (Stensrud et al., 2022). These effects may be relevant when one can plausibly separate the exposure, or treatment, into two distinct components that act along different causal pathways: one component that may affect Y only through M, and another component that may only affect Y through pathways not involving M. This assumption may be more plausible for some exposures or treatments that act through relatively well-understood and distinct biological mechanisms (for example, see Andrews et al., 2023).

Another response is to reformulate the question more fundamentally. Natural effects have been criticised not only because of the assumptions they require, but also because they rely on nested counterfactual quantities that do not correspond directly to interventions that could be enacted in a single real-world setting (Robins et al., 2022). Related concerns can arise for interventional analogues of natural effects, as the implied intervention on the mediator distribution is challenging to connect to a concrete action. From this perspective, such effects may serve as conceptual tools, but they may not always be the most appropriate targets for actionable input. Researchers may instead focus on causal effects corresponding to more explicitly defined interventions on the exposure, the mediator, or both. CDEs provide one example of such effects, but other joint interventions can also be studied, as outlined in Section 5.12 (for example, see Stacey et al., 2019). Alternatively, other interventional direct and indirect effects may be formulated to correspond explicitly to a hypothetical target trial of specified mediator interventions.

We have covered several options for effects one may focus on when conducting causal mediation analysis and discussed some of their challenges. Importantly, researchers need not choose a single causal effect. If several effects answer substantively meaningful questions, they may be reported alongside one another, provided that their interpretation and identifying assumptions are clearly distinguished. For example, researchers may be interested in the thought experiments represented by natural effects or their interventional analogues, but may also estimate the effects of specified

interventions on the exposure, mediator, or both that could plausibly be implemented in practice, thereby providing more directly actionable evidence.

Depending on the target causal effect, different statistical methods may be used for estimation. Although several algorithms and software implementations are available, the choice is more limited in practice when intermediate confounding is present. In these settings, g-methods are typically required. For example, CDEs, IDE, and IIE may be estimated using g-computation, for example with `gformula` in Stata (Daniel et al., 2011) or `CMAverse` in R (B. Shi et al., 2021). A causally informed SEM approach has also been proposed for some settings with intermediate confounding, under additional parametric assumptions (De Stavola et al., 2015), although g-methods typically offer greater flexibility and are more commonly used in this context.

Research questions may also involve multiple mediators of interest, and extensions are available for such settings (for example Daniel et al., 2015; Vansteelandt & Daniel, 2017). However, broadening the question to distinguish effects involving additional mediators adds further requirements to an already demanding set of assumptions.

Much like analyses of time-varying exposure effects, causal mediation analyses can place substantial demands on the time resolution available in the data and related assumptions. Researchers must carefully consider the temporal ordering of the exposure, mediator, outcome, and any relevant confounders, as well as whether the assumptions required for the chosen causal effect are plausible. Even if only a single mediator is studied, we encourage users to start by asking whether (some form of) causal mediation analysis provides answers that match their substantive questions better than alternatives. If it does, they should then consider whether the available data can support the required assumptions, and whether the research team has sufficient expertise to implement the analysis accurately and interpret the findings appropriately. In some cases, it may be preferable to reformulate the research question around less demanding causal effects, especially when the assumptions required to identify those effects are more defensible.

5.13.2 Applications

Ning et al. (2024) perform a particularly impressive study of the effect of early-life adversity upon adolescent mental health using data from MCS. Testing the predictions of a theoretical model, the Dimensional Model of Adversity and Psychopathology (DMAP), the authors estimate indirect effects of two aspects of adversity (deprivation and threat) via cognitive ability and emotional dysregulation, which, according to the DMAP, they are differentially expected to affect. Though they find limited evidence that either dimension of adversity has a net non-zero effect on adolescent mental health, the authors find evidence in support of the DMAP, with deprivation having a stronger indirect effect upon mental health via cognitive ability than via emotional dysregulation, and threat having a stronger indirect effect via emotional dysregulation than via cognitive ability.

The study has several strengths. The authors follow a pre-registered analysis plan, with deviations noted and justified. They also clearly state their estimands and the assumptions under which a causal interpretation can be made. They develop a DAG paying particular attention to the relationships between variables – for instance, they separate exposure-mediator-outcome and mediator-outcome confounders, and they treat threat as an intermediate confounder of the deprivation pathway. Because natural effects are not identified in the presence of intermediate confounding (see above), they accordingly target the interventional indirect effect for deprivation, estimated using the g-formula, while targeting the natural indirect effect for threat, where no such confounder arises. Finally, they use a child-reported, rather than parent-reported, measure of adolescent mental health, given concerns with the accuracy of the latter; they note this decision may have meant they observe no net (total) effect of adversity upon the outcome.

Hope et al. (2021) use MCS data to assess the effect of hypothetical interventions to reduce social inequalities in childhood mental health problems (MHP). Though it may not immediately appear so, this can be treated as a mediation question. Using marginal structural models, the authors first estimate the controlled direct effect of income upon childhood MHP with parenting practices treated as a mediating factor and fixed at

observed levels. They then simulate the effect of interventions that improve parenting practices by re-estimating the CDE at predicted values for the mediator; the intervention may thus narrow inequalities by reducing differences in this mediator between groups. A key strength of their study is that the hypothetical interventions they evaluate vary in their effectiveness, uptake, and targeting (e.g., offered to families on means-tested benefits only), with these parameters selected based upon reviews of the literature (e.g., meta-analyses of the effectiveness of parenting interventions) and expert discussions (e.g., on plausible targeting strategies). A limitation is that they do not account for potential reverse causality between (early) child mental health problems and parenting behaviour.

Pinto Pereira et al. (2020) – discussed in Section 5.12.2 – assess mediation of the effect of early- and mid-adulthood obesity upon mid-life physical function using data from the NCDS and the NSHD. Using the parametric g-formula to estimate IDE and IIE for different treatment regimes (never obese, persistent obesity, and obesity beginning at particular ages), they find little of the total effect is mediated via physical activity. However, as the authors note, an issue is their dichotomous measure of physical activity, which may have little discriminant validity, thus biasing estimates of the indirect effect towards the null.

5.13.3 Opportunities

The attributes of CLS's cohorts discussed in Section 5.12.3 also make the data well suited for causal mediation analyses. Specifically, the availability of multiple data collections and the breadth of data collected overall mean that temporal ordering can be established and there are variables for many important mediators and time-varying confounders. Because data collection began at birth and the same characteristics are measured repeatedly, many baseline confounders (e.g., pre-exposure measures of mediators) are also available.

5.14 Family Designs

5.14.1 Method

In Section 5.5, we discussed the longitudinal fixed effect design which exploits repeated measurement of the same individuals to control for time-invariant confounders by design rather than through explicit statistical control. Family designs similarly exploit features of the data-generating process to control for broad classes of confounders. There are several family designs that can be used with CLS's cohort data, including designs that precede the molecular genetics revolution and those that use genetic data directly.

Multiple CLS studies have data on siblings, either from a small number of multiple births or from data collections on a larger number of cohort members' siblings; for instance, Sweeps 2 and 3 of the MCS included a questionnaire for older siblings. Like longitudinal fixed effects, the sibling comparison design exploits variation in exposure within sibships, this time to account for factors that are common between them (Sjölander et al., 2022). This includes aspects of family background (e.g., parenting styles, family socioeconomic position) and shared genetics (50% on average for dizygotic twins or full siblings; 100% for monozygotic twins).

The usefulness of the sibling design depends on the extent to which confounding factors are in fact shared. The behaviour genetics literature suggests the contribution of shared environments to traits and other outcomes is relatively weak and declines in predictive power as individuals age (Davey Smith, 2011; Plomin, 2018; Plomin & Daniels, 1987). Importantly, this literature also shows that children elicit responses from their environment, including the parenting they receive (see, e.g., Selzam et al., 2018; Wertz et al., 2019) – in other words, less of the family environment may be 'shared' than initially expected. The range of influences accounted for by the sibling design plausibly decreases as the age gap between siblings becomes greater, given differences in childhood environment will be larger and as common events will be experienced at different developmental stages. Again, as with longitudinal designs that exploit within-person variation, the sibling design in fact only captures common environments that make siblings more similar to each other than to other people – to

the extent that a common exposure has a heterogeneous effect (e.g., because its effects interact with other exposures, such as prior unique developmental experiences) it will not be fully controlled for by design in this set-up.

Moreover, in some circumstances, the sibling design can in fact increase bias, relative to a standard observational design. This is partly due to problems which affect fixed effects designs more broadly (see Section 5.5): where there is measurement error, a greater component of within-sibling variation in the measured factor will be noise, and where a relatively strong confounder causes the variation within siblings, exploiting just within-sibling variation can increase bias, analogous to the issue of Z-bias discussed in Section 5.1. Further, some issues are (arguably) unique to the sibling design, including 'carryover' effects (Sjölander et al., 2016), where one sibling influences another, and the existence of shared mediators (Sjölander et al., 2022), which are absorbed into the fixed effect along with shared confounders – financial resources could be shared within families, for example. As sibling designs only exploit siblings who are discordant on exposure, they can also suffer from power issues and low generalisability to the general population, which includes singletons and families with siblings who are not discordant on exposure.

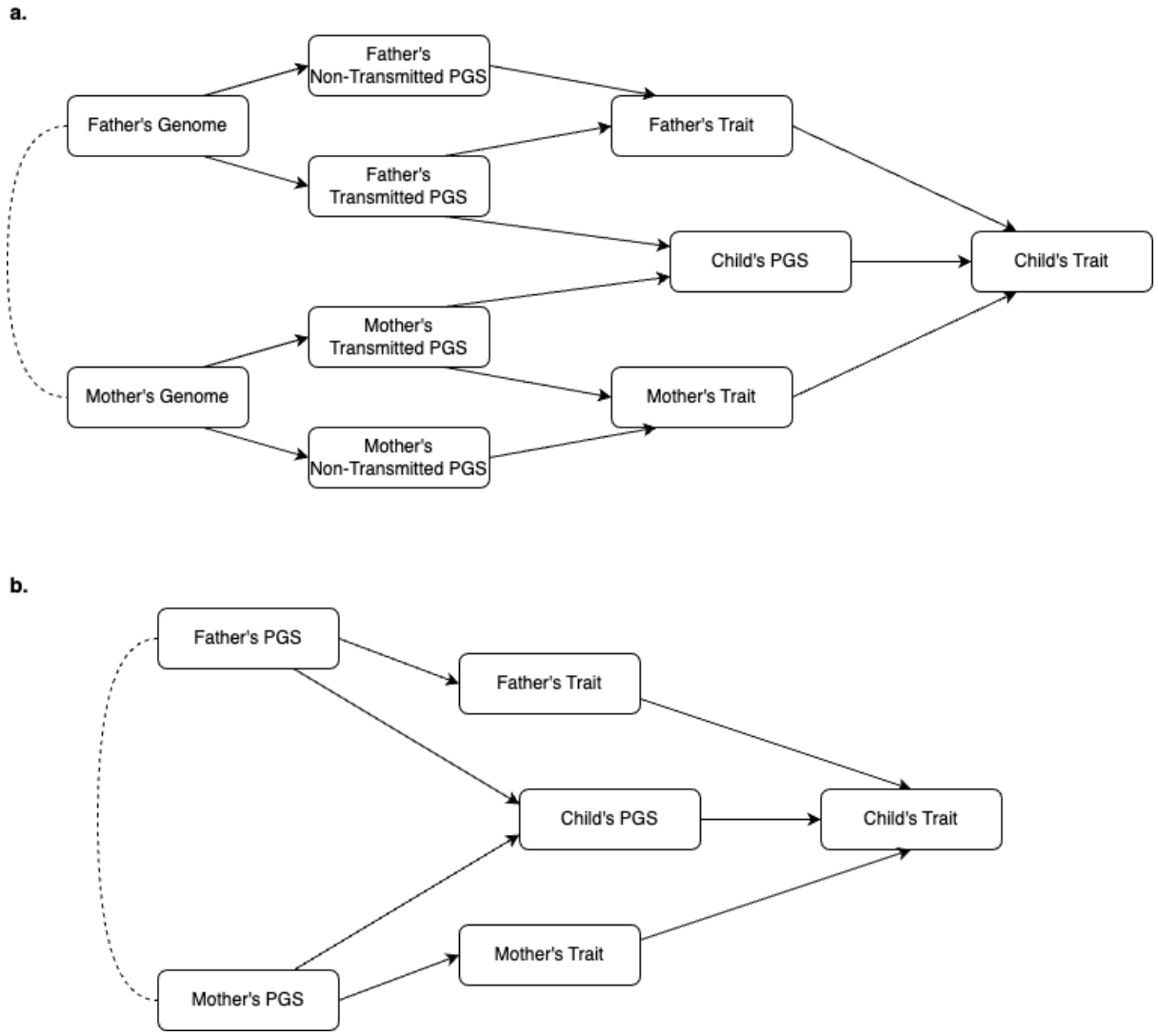
Another family-based method that can be used with CLS data is the adoption design. NCDS and BCS70 cohort members were born during periods when adoption was relatively common (Rushton et al., 2020) and contain data from ~ 300 adoptees in total, who can be combined, including with adoptees identified in other studies, to increase total sample sizes. As discussed in Appendix Section 8.2, a major and consistent finding in the behaviour genetics literature is that almost all human traits are heritable: parents and their (biological) offspring resemble each other at least in part because they share the same genes. The adoption design exploits the fact that children and their adoptive parents are not expected to be genetically related. Therefore, associations between adoptive parental traits and family environments and children's outcomes should not be explained by direct genetic transmission. However, it is still possible for children to elicit specific parenting behaviours and other environmental responses (evocative gene-environment correlation), a form of reverse causality. Moreover, adoption placements can be selective, including where children are adopted by

relatives. It is necessary to confirm non-familial adoption as otherwise genetic relatedness would still be present, albeit to a lesser extent than with one's biological parents.

The adoption and sibling designs use facts of genetic inheritance indirectly, but the advent of molecular genetics has led to the development of designs that use measured genetic variation. Biological parents and offspring in the MCS were genotyped in Sweep 6 (age 14 years) of the survey, yielding data from more than 2,700 sets of genetic 'trios' (mother-father-offspring) and ~ 8,500 (overlapping) sets of 'duos' (mother-offspring or father-offspring pairs). An increasingly popular design is the genetic nurture design (Bates et al., 2018; Kong et al., 2018). This exploits the fortuitous feature of Mendelian inheritance that offspring inherit a random half of each parent's DNA. With genetic data from parents and offspring, parental genetic data can be decomposed into variants that were inherited by the offspring and variants that were not. Any causal effect of the 'non-transmitted' genetic variants upon offspring can only be via the environment – e.g., through the impact on parental behaviour (so-called genetic nurture). Any association cannot reflect direct genetic transmission, meaning – quite remarkably – that genes can be used to study the environment, even if one isn't interested in genetics *per se*. Conversely, the effect of transmitted genes on offspring's outcomes is a combination of genetic nurture and direct genetic effects, and so, analysed alongside non-transmitted variants, transmitted variants can be used to determine the impact of direct genetic effects.

Genetic nurture analyses proceed in one of two ways. With access to raw genetic data, polygenic scores (summary measures of genetic liability to a trait capturing the effect of many genes; Choi et al., 2020) can be constructed for non-transmitted alleles and transmitted alleles, separately, and used in analyses (Kong et al., 2018). Alternatively, with existing polygenic scores (PGS) constructed from all alleles, transmitted and non-transmitted (as available on the UK Data Service), researchers can condition on child PGS and use parental PGS to identify indirect genetic effects (Bates et al., 2018). In both cases, the parental PGS can be used as instrumental variables for a trait of interest in Mendelian Randomisation (MR) analyses. See Figure 5.17 for DAGs reflecting these designs. Note, it is not necessary that the parent's (exposure) and child's (outcome) trait are the same – see, for example, Hughes et al. (2024), who use genetic trio data to assess the effect of parental education on child's mental health.

Figure 5.17: Genetic nurture designs that utilise genetic trio data. In observational data, parental traits and offspring traits are correlated but this may be due to direct genetic transmission, rather than an environmental effect of parental traits/behaviours upon their children. Molecular genetic data can be exploited to account for this. (a) Polygenic scores (PGS) are derived from parents' and children's genetic data to determine which alleles from each parent have been transmitted to their child and which haven't, to construct PGS of transmitted and non-transmitted alleles (Kong et al., 2018). Non-transmitted PGS can only have a causal effect via the 'environment' (e.g., due to parental traits) and so (under specific conditions – e.g., exclusion restriction) can be used as instruments for parental traits. (b) Conditioning upon child's PGS closes the direct genetic transmission path, leaving the indirect 'genetic nurture' path between parental PGS and children's traits (Bates et al., 2018). Again, parental PGS can be used as instruments for their own traits (under appropriate conditions).



As Mendelian Randomisation is a form of IV, under treatment effect heterogeneity (and meeting the monotonicity assumption), the estimand in this analysis would be a

Complier Average Causal Effect (CACE; Section 5.9). The exact interpretation of who compliers are may be difficult or impossible to ascertain and this has consequences for generalisability.

Another issue with the genetic nurture design is that genetic nurture effects can be biased by population-level phenomena such as assortative mating and population stratification. But the design does allow one to determine direct genetic effects accounting for these phenomena. As with other Mendelian Randomisation analyses (Section 5.9), this design can be biased in the presence of horizontal pleiotropy, a particular problem when looking at complex traits or when using PGS rather than single well-understood genetic variants, for both indirect and direct genetic effects. Nevertheless, accounting for direct genetic transmission may be an improvement on what has gone before. Genetic nurture studies are also power hungry – current PGS typically only explain a small amount of variation in their relevant traits – so combining MCS with other data sources containing genotyped families (e.g., Born in Bradford, the Avon Longitudinal Study of Parents and Children [ALSPAC], and the Norwegian Mother, Father and Child Cohort Study [MoBa]) may be necessary.

5.14.2 Applications

Adoption and genetic nurture designs have each been used in CLS's cohorts. Dearden et al. (1997) use data from the NCDS to calculate father-son correlations in educational attainment, comparing these correlations in natural and adoptive families to assess whether intergenerational stability is genetic, though due to small sample sizes, estimates are imprecise. Sacerdote (2002) performs a similar analysis with the same data, showing associations between adoptive families' SES and adopted children's reading test scores, likelihood of university attendance, and marriage.

Wright et al. (2025) use trio data from MCS to examine genetic nurture of BMI, finding that, although there is widespread observational evidence that parental BMI is related to offspring BMI, when accounting for direct genetic transmission, there is only evidence that mother's BMI is related to offspring BMI. However, BMI is a complex phenotype, influenced by multiple traits and behaviours, and there is a substantial possibility for horizontal pleiotropy to explain maternal results. Nevertheless, the study

shows that associations between fathers' and children's adiposity may be explained by direct genetic transmission, rather than father's behaviour or related environments. Tanksley et al. (2025) use a similar set-up but focus on unbiased estimation of the direct genetic effect of an externalising PGS upon phenotypic externalising behaviour (as measured by the Strengths and Difficulties Questionnaire) – associations between the child's PGS and the SDQ scores are similar whether parental PGS are included in regression models or not, suggesting the variants captured in the PGS have a causal effect on the child's behaviour.

5.14.3 Opportunities

The MCS sibling questionnaires are presently underutilised and contain several items also asked of cohort members, enabling sibling analysis. The MCS genetic trio and duo data are also underutilised. Rectifying this is straightforward: ready-made PGS for a wide array of traits are available via Special Licence Access on the UK Data Service; see the CLS Genetics GitHub website for more details.

5.15 Negative Control Outcomes and Exposures

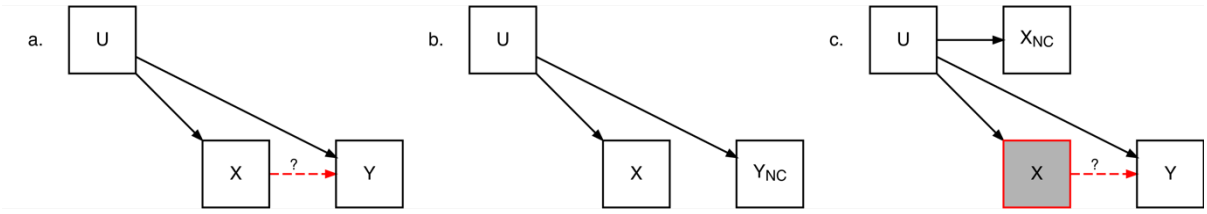
5.15.1 Method

In Section 3.3, we discussed making theories "elaborate" as a way to support causal inference. A broad concern in observational research is that an exposure, X, may be associated with an outcome, Y, due to an unobserved – or poorly measured – confounding factor, U. That the exposure and outcome are materially confounded by U is a competing theory to the possibility that the exposure has a causal effect on the outcome.

Competing theories should be treated seriously; one way of treating a competing theory seriously is to ask what predictions it makes about the world. A given confounding process may imply a larger set of associations than simply a spurious one between X and Y. For instance, the confounding factor may generate an association between X and a larger set of outcomes, or between Y and a larger set of exposures. But these associations may not be predicted by the central theory that X causes Y, and there may

be no plausible reason the alternative exposure would cause Y or X would cause the alternative outcome.

Figure 5.18: Negative control outcome and negative control exposure design. (a) We are interested in the causal effect of X upon Y, but anticipate there may be confounding via a common (unmeasured) cause U. Any association between X and Y could reflect this confounding pathway. (b) U may also be a common cause of another 'negative control outcome' (Y_{NC}) which we have good reason to expect X would not have a causal effect on. Any association between X and Y_{NC} would therefore reflect the confounding by U. (c) U may additionally be a common cause of another 'negative control exposure' (X_{NC}) which we have good reason to expect does not have a causal effect on the outcome of interest, Y. Conditional on the exposure of interest (X), any association between X_{NC} and Y would therefore reflect the confounding by U. Note, in practice, some confounders may be measured, and these would be conditioned upon in the negative control analysis.



The negative control exposure and outcome designs encapsulate this logic (Lipsitch et al., 2010).[46] In the negative control *outcome* design (Figure 5.18b), a 'placebo' outcome, Y_{NC}, is identified which should only have an association with X through a confounding factor U that is thought to bias the association between X and Y. If an association is observed between X and Y_{NC}, this may indicate the main study analysis is biased by U.[47] To take an example, let's say we are interested in the effect of breastfeeding on child cognitive development (Lawlor et al., 2016). Breastfeeding is strongly socially patterned in the UK, with more socioeconomically advantaged parents more likely to breastfeed; an association between breastfeeding and child cognitive development could therefore be expected due to higher socioeconomic position (SEP) of parents or even direct

[46] Another negative control design – negative control populations (Piccininni & Stensrud, 2024) – is discussed further in Section 5.16.

[47] Sanderson et al. (2021) note the negative control outcome is a renaming of Austin Bradford Hill's (1965) 'specificity of association' criterion for assessing causation. The flip-side of specificity is coherence among outcomes (Rosenbaum, 2019) – some variables can be used as "positive controls" to determine that statistical models perform as expected (Gutierrez et al., 2025).

heritability of cognitive ability. But SEP is difficult to measure accurately, so residual confounding may persist even if measured SEP is controlled for. Home invasion by mice during childhood is also socioeconomically patterned, but we would not expect it to be influenced by breastfeeding. Home invasion by mice is therefore a plausible negative control outcome because any association with breastfeeding should be spurious rather than causal. What's more, it would be spurious for similar reasons that a link between breastfeeding and child cognitive development could be spurious.

The negative control *exposure* design is similar (Figure 5.18c), but in this case a 'placebo' exposure, X_{NC}, is identified that should not have a causal effect on the outcome but would be likely confounded by the same factors as X, the exposure of interest. An example of a negative control exposure design is using paternal smoking to assess whether maternal smoking during pregnancy has a causal effect upon child health, in cases where the smoking is hypothesised to have an effect by influencing the intrauterine environment. As mothers' but not fathers' behaviour has a direct effect on the intrauterine environment, one would expect to see specific associations between mother's smoking and child health. On the other hand, as both mother's and father's smoking are patterned socioeconomically, a countervailing theory that associations are driven by confounding through SEP would also predict an association between father's smoking and child health. Thus, if paternal smoking is related to offspring outcomes (conditional on maternal smoking), it would suggest an association between maternal smoking and child outcomes is also biased (Brion et al., 2007).[48]

Negative control designs are sometimes referred to as placebo, specification or falsification tests (Eggers et al., 2024). The success of the negative control outcome and exposure design depends on the negative control sharing the same confounders as the true exposure or outcome of interest – referred to as 'U-comparability' (Lipsitch et al., 2010). Since it refers to unobserved confounders, the assumption of U-comparability is untestable. In practice, the proposed negative control(s) may only share part of the confounding structure. In the absence of full U-comparability, an observed association

[48] Conditioning on maternal smoking is required as the X_{NC}-Y association is additionally confounded by the path $X_{NC} \leftarrow U \rightarrow X \rightarrow Y$ if a causal effect of X upon Y exists.

with the negative control will not necessarily indicate residual confounding in the X–Y association, and a null association with the negative control will not necessarily indicate absence of residual confounding. Nevertheless, the negative control design may still provide useful information on the presence or absence of bias from major confounders – e.g., on SEP, as in the examples above.

It is not just the presence of the confounders, but similarity in the strength of their relationship with the true and negative control exposures or outcomes that is important. For instance, given the additional possible health risks of smoking during pregnancy, one may expect greater selection into smoking among mothers, compared with fathers. For this reason, a negative control test may represent only a weak test of the competing theory. Showing an association between the negative control and the true exposure or outcome would suggest the true exposure-outcome association is biased, but the reverse may not be true: not finding an association with the negative control is not necessarily dispositive evidence that the true exposure-outcome association is unbiased.

In practice, it may be difficult to identify negative controls that have no causal association with the true outcome or exposure, whatsoever. For example, it is possible that father's smoking influences the intrauterine environment, albeit indirectly via the mother inhaling smoke second-hand. In this case, the pattern and overall size of associations would be informative, rather than only observing a null association for the negative control: given father's effects on the intrauterine environment should be weaker than mother's, we would expect (a) child health to have a weaker association with father's smoking than mother's smoking and (b) the absolute size of the association with father's smoking to be small. Otherwise, strong confounding bias is likely to be present.[49]

[49] The utility of examining patterns of association ("pattern matching") was discussed in Section 3.3, where we noted how associations between measures of socioeconomic position (SEP) and health should follow a particular pattern of association if explained by cognitive ability, a hypothesised confounder of that relationship.

The negative control method can therefore involve a degree of uncertainty for the analyst and require some persuading of the reader. Operating under this uncertainty requires discipline and transparency as it can be tempting to spend "researcher degrees of freedom" (Simmons et al., 2011) in light of inconvenient results – one should not jettison a negative control analysis after seeing the results, if the results motivate one to think 'maybe the confounding structures are not quite the same' or 'perhaps there is a causal relation the association represents'.

Nevertheless, there are general strategies for identifying plausible negative controls. These focus on finding situations where a hypothesised *active ingredient* linking exposure and outcome is missing or an essential mediating pathway is not present (Lipsitch et al., 2010; X. Shi et al., 2020) or, less restrictively, is weaker. The *parental comparison design* (Davey Smith, 2008; Richmond et al., 2014) used in the smoking example above follows this logic, and is frequently used in settings where maternal exposures are expected to operate via the intrauterine environment given fathers do not influence this directly. In pharmacoepidemiology, negative control outcomes can be relatively straightforward to identify as treatments may be expected to have causal effects on only a limited set of health outcomes. For instance, flu vaccination, which would plausibly reduce hospitalisation rates due to respiratory problems, should not lower rates of hospitalisation due to injury or trauma (L. A. Jackson et al., 2006).

Another strategy is to take measurements from periods where the exposure should not have an effect on outcomes – for instance, flu vaccination should not affect hospitalisation outside the flu season, when there is no influenza for it to prevent (L. A. Jackson et al., 2006). Using outcome measurements prior to exposure or exposure measurements post outcome can also represent expressions of this idea – for instance, a future measurement of environmental pollution when looking at hospital admissions for asthma (Flanders et al., 2011). Height, which has been repeatedly used as a negative control outcome in the health and social sciences (Cohen-Cole & Fletcher, 2008; L. Wright, 2020; L. Wright et al., 2021), can have this property as it is largely set from late adolescence. It is also very highly heritable, so large associations with environmental factors (within a population) are likely spurious. Cohen-Cole & Fletcher (2008) show

that in addition to obesity 'spreading' between friends in social networks (Christakis & Fowler, 2007), so do height, acne and headaches.

Negative control tests are routinely used in some particular study designs. For instance, as discussed in Section 5.10, sharp RDD studies typically exploit a placebo test, where arbitrary cut-offs in the running variable are chosen that are known to have no bearing on treatment, but could still be related to unmodelled confounders. Sanderson et al. (2021) discuss using phenotypes determined prior to outcome and exposure as negative controls in MR studies, e.g., tanning ability and hair colour.

So far, we have described negative control designs for detecting bias from unobserved confounding. But negative controls can also be used to detect selection and measurement biases (Arnold et al., 2016); regardless of the source of bias, the purpose of the negative control exposures and outcomes is to identify variables that have similar (hypothesised) biasing structures to the main association of interest. For instance, recall biases can generate spurious associations between exposure and outcome if the accuracy of recall of exposure is related to the outcome (e.g., disease cases more likely to remember a childhood ailment). Yet, recall bias may also apply to negative control exposures that have no plausible causal effect on outcomes. Zaadstra et al. (2008) examine recall bias in a study of multiple sclerosis (MS) using negative control exposures for broken limbs, concussion, and tonsillectomy, each of which is not expected to have a causal effect upon MS. Returning to genetics studies, sex is a particularly useful negative control when using polygenic scores as these are typically calculated from autosomal (i.e., non-sex-specific) chromosomes; any association with observed sex must therefore be due to selection bias.

Negative controls can be used not just for bias detection, but also bias correction or reduction (X. Shi et al., 2020; Yang et al., 2024). Again, strength of association between negative controls and true outcomes and exposures can be revealing. For instance, VanderWeele (2011) notes that although height, acne and headaches are correlated within friendship groups (Cohen-Cole & Fletcher, 2008), the size of this correlation is lower than that for obesity, and so greater confounding would be required to explain the obesity spreading hypothesis.

5.15.2 Applications

Libuy et al. (2024) use MCS data to assess the effect of changing proximity to fast food restaurants on weight gain. Given fast food restaurants are spatially patterned, they use proximity to several other types of commercial establishment (e.g., employment and career agencies) as negative control exposures, arguing that time-varying changes in these might reflect unobserved confounding processes, such as gentrification. However, the plausibility of the assumption behind the negative control analysis is not explored and other types of commercial establishment not examined could plausible share confounding structures more closely with fast food restaurants (e.g., betting shops). Hamer & Batty (2020) use data from BCS70 to investigate the association between infection requiring hospital admission and later diabetes risk, using non-infectious reasons for admission as a negative control exposure. This has a more plausible common confounding structure (e.g., health consciousness).

Tabor et al. (2025) use data from MCS, Next Steps, and BCS70 alongside data from other cohorts to examine the effect of sexual minority status on asthma. Their theory suggests effects should only occur in adulthood so they use childhood asthma as a negative control outcome, an application of the idea that measures prior to plausible causal effects occurring can be used as negative controls. Wright et al. (2021) use data from Next Steps to assess the effect of youth unemployment on later mental health, using height and patience as negative control outcomes. A question in both papers is whether confounding structures are sufficiently similar to the focal associations of interest.

5.15.3 Opportunities

Given data collection from parents and children, including on characteristics prior to the cohort members' births, there are ample opportunities to employ the parent comparison design in CLS cohorts. The cohorts have each collected data on a wide variety of factors and traits, which may be useful more generally (e.g., height, suncream use, etc.). Data linkages also increase the possibility for negative control designs, including links with health record data, enabling assessment of multiple health outcomes, which can serve as negative control outcomes. A variety of data on health

outcomes is also collected in standard survey data collections. Longitudinal measurements add further possibilities as there is ample pre-exposure measurement of many variables. Finally, there is also an opportunity for research-enabling activity reviewing plausible negative controls and determining their possible use in CLS data.

5.16 Cross-Context Analysis

5.16.1 Method

In Section 5.15, we saw that one way to make a theory elaborate is to enumerate which variables an exposure should have a causal effect upon (and which it should not) or which variables an outcome should be causally affected by (and which it should not), given the set of predictions a theory can make. Another way to make a theory elaborate is to draw out the settings or individuals a hypothesised causal effect should exist in or in which the effect would otherwise be stronger or weaker than usual. A competing theory may be able to explain the presence of an association between X and Y in general, but not specific patterns of association between these different cases. Alternatively, a causal effect may be similar in different settings, but (hypothesised) confounding structures may change, so consistency of association across these settings is evidence compatible with the causal theory being true. Cross-context analysis encapsulates this logic by performing analyses in different (subsets of) data or populations (Eggers et al., 2024; Gage et al., 2016; Piccininni & Stensrud, 2024).

There are three types of comparison that are particularly relevant for CLS's cohort studies. First, time – a causal effect may be anticipated in some periods, cohorts, or stages in the life course, but not others. Second, space – a causal effect may be anticipated in some geographies, but not others, or the confounding structure may differ fundamentally between geographies. Third, population subgroups – there may be certain sections of the population for whom a causal effect would not be anticipated. These are best explained by way of examples.

Time

- Cognitive ability is hypothesised to have a beneficial effect on health due to greater knowledge and better decision making among those with higher cognitive

ability (Gottfredson, 2004). Knowledge about diseases – and how to prevent them – varies across time as scientific discoveries are made and technologies are introduced (Clouston et al., 2016). Given this, if cognitive ability causes health, we should expect inequalities in health according to cognitive ability to grow after facts about a disease are first discovered and to decline as technologies are introduced that reduce the difficulty of avoiding or managing a disease.

- In the UK, maternal smoking during pregnancy has become less prevalent but more strongly associated with socioeconomic adversity over time. If associations between maternal smoking and offspring outcomes are merely driven by confounding through socioeconomic position, we would therefore expect stronger associations in more recent cohorts (Sellers et al., 2020).

Space

- Between 1999 and 2010, the UK Government implemented a well-funded strategy to halve teenage pregnancies in England (Hadley, 2017). Teenage pregnancies did fall over this period (Wellings et al., 2016) but may have done so absent the strategy. Was the Teenage Pregnancy Strategy successful? At the very least, we would expect pregnancy rates to fall faster than in other UK countries that did not implement such strategies (Baxter et al., 2021; Craig et al., 2016).
- The 2013 Removal of the Spare Room Subsidy (the so-called 'Bedroom Tax') reduced benefits for working-age individuals in social housing with an unoccupied bedroom. This reduction in entitlements for individuals already more likely to be on low incomes may be expected to reduce mental and physical health (cf. Reeves et al., 2016). However, a curious feature of the policy's implementation is that, in Scotland, the government compensated any reduced benefits with Discretionary Housing Payments, meaning the bedroom tax did not affect incomes for residents in that country. We should therefore not expect any association between being subject to the bedroom tax and mental health in Scotland but should expect to see one in England and Wales.
- The social patterning of breastfeeding varies across the world. For instance, in the UK, socially advantaged mothers are more likely to breastfeed, but this is not

the case in Brazil. If the association between breastfeeding and child development is simply a product of socioeconomic confounding, we would expect a positive association in the former, and little association in the latter (Brion et al., 2011).

Population Subgroups

- As discussed in Section 5.9, certain variants of the ALDH2 gene are related to better cardiovascular health. This is hypothesised to be due to the effect of ALDH2 on reducing alcohol consumption by determining the comfort with which one metabolises alcohol. In East Asian cultures, women typically drink little, regardless of which ALDH2 variant they carry, while men are more likely to drink and for this to be related to ALDH2. If ALDH2 has a causal effect only via alcohol consumption, we should therefore expect associations with cardiovascular disease in East Asian men but not women (L. Chen et al., 2008).
- Unexpected events (Section 5.8) should only have effects among those who are aware of them. For instance, the release of the Royal College of Physicians' (1962) report on smoking and health was a watershed moment in public appreciation of the harms of smoking, but if access to information is sufficient to alter the decision to smoke, then smoking among TV owners or newspaper readers – likely more aware of the RCP report – should have declined the most (James, 2024).
- Cholera being transmitted via infected water, rather than miasmic air, predicts that people who live near the Broad Street pump, but do not take its water, will be less likely to succumb to the disease, and that those who live far away, but still have its water delivered, will become infected (Section 3.3).

The examples where the causal effect is expected to differ share a common feature: an essential mechanism transmitting the effect is missing, or is otherwise weaker, in one setting than another. For instance, for the bedroom tax, in Scotland, the essential ingredient – lower income – was not present. Readers may notice the similarity between the cross-context design and the negative control outcome design – rather than choose different variables for analysis, different sections of the data are used for which, e.g., causal effects should not exist. Cross-context designs where there is an absence of a

causal effect in the comparator population but biasing processes are otherwise the same have been termed negative control population designs (Piccininni & Stensrud, 2024).

Like the negative control design, convincing cross-context analyses require knowledge of the subject area. However, here the barrier to entry may be lower, and the arguments easier to make. For instance, the ALDH2 example largely requires identifying a group with very low alcohol consumption regardless of the ALDH2 variant they carry. Also, like the negative control design, the cross-context design can still be useful when the strength of a causal mechanism varies across settings, rather than the mechanism being absent or blocked entirely. The elaboration lies in predicting an overall pattern of results. In their discussion of cross-context designs, Gage et al. (2016) describe a test of the hypothesis that cannabis has a causal effect on psychosis due to its tetrahydrocannabinol (THC) content: a corollary is that as levels of THC in cannabis have increased, associations between cannabis use and psychosis should have similarly increased, if this is the causal mechanism.

Where the cross-context design is motivated by an argument that the causal effect of the exposure of interest will vary across populations, plausibility relies on the extent to which biasing processes are similar, akin to the assumption of U-comparability in the negative control outcome and exposure designs (Section 5.15). Additionally, where it is the biasing processes that are hypothesised to vary across populations, an assumption is that the causal effect should be similar in each population, but effects may be heterogeneous and could themselves lead to cross-context differences in associations.

Cross-context designs are extensively used in the genetics literature in the study of gene-environment interaction (GxE; Pereira et al., 2022). For instance, studies have examined the association between polygenic scores (PGS) and attained education as access to education has expanded (Herd et al., 2019; Morris et al., 2026), or between PGS and body mass index before and amidst the obesity epidemic (Conley et al., 2016; L. Wright, Davies, et al., 2024) or in other groups for whom genetic effects should be stronger (e.g., among those genetically predisposed to obesity and living near fast food outlets; de Roo et al., 2024). Again, adding a cross-contextual element supports causal interpretations by testing for a predicted pattern of associations. However, a drawback

of GxE studies is that the environment may be 'endogenous' (*determined in the system*; e.g., people may choose where to live based on their genetic predispositions). Further, for complex traits, the associations between PGS and phenotypes may reflect pleiotropic effects (also see Westerman & Sofer, 2024, for discussion of issues in GxE studies). Note, birth year – and the cohort effects it implies – more plausibly represents exogenous environments, as genetic changes occur at a relatively slow rate. A drawback, however, is the lack of specificity in which aspects of the environment have changed (e.g., aspects of the obesogenic environment in general).

5.16.2 Applications

Sellers et al. (2020) compare associations between maternal smoking in pregnancy and offspring development (birthweight, conduct problems and cognitive ability) in the NCDS and MCS – as discussed, social patterning of maternal smoking was stronger in the more recent cohort. Associations between maternal smoking and birthweight were similar in each cohort, whereas associations were stronger in the MCS for other outcomes assessed, consistent with material bias from confounding. However, causal effects becoming larger in more recent cohorts is also compatible with the results.

Wright et al. (2024) examine the association between a PGS for BMI and phenotypic BMI across life in the NSHD, NCDS, BCS70, and MCS, which collectively span a period before and during the obesity epidemic – the NSHD cohort members grew up during post-war rationing, while childhood obesity rates reached 20% in the MCS. They find stronger associations in more recent cohorts, particularly comparing the MCS with earlier generations, consistent with the theory that the PGS reflect genes that have effects on adiposity via modifiable behaviours. However, this is evidenced only indirectly using birth year as a proxy for exposure to the 'obesogenic environment' – it is unclear which aspects of the environment are operative (if any).

5.16.3 Opportunities

Opportunities for cross-context analysis arise from the temporal and geographic span represented in CLS's cohorts. For instance, education systems have historically differed in England, Scotland and Wales, e.g., on rules on minimum school leaving ages, which

(over some time frames) entail that individuals born earlier in a school year in England are able to leave school at a younger age than if they went to school in Scotland (Del Bono & Galindo-Rueda, 2007). Geographic range can also be extended by harmonising with studies from other countries where, e.g., different confounding structures for exposures and outcomes of interest may be present.

There are numerous opportunities for GxE in the cohorts, including comparing genetic effects across cohorts, given the generational differences they represent (Morris et al., 2026; L. Wright, Davies, et al., 2024). Opportunities also exist linking genetic data with geographic data (e.g., to identify differences according to pertinent features of the local environment) and with other survey data, given the extensive phenotypic data available (e.g., association between genetic risk and depression among victims of crime), notwithstanding potential endogeneity of these latter environments. Again, harmonisation with other international data could also expand possibilities for GxE cross-context analysis.

5.17 Biases of Known Direction

5.17.1 Method

Though it would be optimal and most useful to identify a precise and unbiased causal estimate, sometimes only determining the overall direction of a causal effect or that it exceeds a minimum effect size is sufficient or still insightful. A valid test of a theory may be that X increases Y, however large or small (though it is of course usually better to know that X is an important causal factor of Y), and one may only need to know that a causal effect exceeds some defined quantity to make a good decision – for instance, knowing that an intervention will yield higher benefits than the alternative. Health economists have a useful concept, 'value of information', which denotes the higher expected utility one would obtain if able to make decisions on more accurate data (Tuffaha et al., 2014); if decisions are unlikely to change following further data collection, then the information collection does not have much value instrumentally. Analogously, in some circumstances, one may be satisfied with a biased estimate of a causal effect if removing the bias would not yield further valuable information.

While the precise extent of a bias may not be known, its direction may be guessed at with high certainty, and this can be used to draw causal inferences even where biases are present and acknowledged (Rosenbaum, 1989, 2020). For instance, an observed positive association would only get larger if a confounding factor positively related to X but negatively to Y (or vice versa) were controlled for. Readers will be aware of this type of reasoning – it appears frequently in discussion sections of papers where authors note the possibility of a source of bias but dismiss its importance as it implies the observed association is a conservative (under)estimate of a true causal effect. One oft-repeated case of this is independent, non-differential (i.e., classical) measurement error in the exposure variable which, in expectation, should bias an association towards the null, in proportion to the size of the measurement error (i.e., regression dilution).

Biases of known direction need not only enter at the interpretation stage of a study – they can also motivate study design. One example of this is to choose a setting in which confounding structures run opposite to a hypothesised causal effect – for instance, in many low- and middle-income countries, obesity is related to higher socioeconomic status (Dinsa et al., 2012), so observational estimates of the causal effect of obesity upon health conditions are likely biased towards the null. Another possibility is the judicious choice of a comparator group that provides a greater hurdle for a causal effect to surmount. For instance, in a study examining the effect of banning individuals with prior convictions from purchasing guns (and discussed in Rosenbaum, 2020), Wright et al. (1999) compare rates of violent crime among those affected by the law against individuals who had previously been arrested for a crime but found not guilty; one may expect that, if both were able to purchase guns, those who have been convicted of crimes would have a greater likelihood of committing further crimes than those only previously suspected of criminality.

Exploiting biases of known direction can be especially compelling when it is possible to obtain two or more estimates biased in opposite ways, in which case these estimates can provide an upper and lower bound to “bracket” the true causal effect. The main way this can be achieved is through the use of multiple comparator groups (Rosenbaum, 1987). For instance, in the difference-in-differences design (Section 5.6), such a bracketing analysis would work by identifying comparator groups who violate the

parallel trends assumption, one anticipated to have stronger time trends than the treatment group and one to have weaker time trends than the treatment group (Ye et al., 2024).[50] Hasegawa et al. (2019) bracket the effect on homicide of the repeal of Missouri's handgun purchaser licensing law, which required a permit (and hence a background check) to buy a handgun, using two groups of comparator states that had either higher or lower homicide rates than Missouri prior to the repeal of the law.

Exploiting biases of known direction is not without its problems, however. Most obviously, in many cases it may not yield any useful information because it would suggest an estimate is anti-conservative (i.e., if anything an association is biased away from the null). Examples of this abound – for instance, in studies of the effect of socioeconomic position on health, both causal and confounding (health-related selection) explanations predict poorer health in more socioeconomically deprived groups. Additionally, it may not be straightforward to identify comparator groups biased in opposite directions (and thus obtain a bracket for the causal effect), or the resulting bracket may be too wide to be useful.

In other cases, only one bias may be known but others not, and it may not make sense to emphasise the one known factor. This is particularly the case where biases interact, yielding unclear direction of bias in aggregate. For instance, bias from classical measurement error is not necessarily towards the null where there is unobserved confounding (Davey Smith & Phillips, 1996).[51] Given incentives within academia to overclaim evidence for a proposition, it is possible for motivated reasoning to lead authors to focus attention on those known biases that lend support, rather than those that don't (Fox et al., 2021).

[50] Ye et al. (2024) provide formal estimation procedures for difference-in-differences bracketing analyses.

[51] It is also worth noting that for classical measurement error the phenomenon of attenuation towards the null is true in expectation but may not hold in a given study, so the direction of bias is not in fact known (Fox et al., 2021).

5.17.2 Applications

Ermisch & Pevalin (2005) use data from the BCS70 to assess the effect of teenage parenthood upon partnership status and partner quality at age 30y. They use multiple comparison groups, including women who did not have a teenage birth and – using miscarriage as an IV – women whose teenage pregnancy ended in miscarriage. Though not reported so explicitly, using miscarriage as an IV could bias towards a "benign view" of teen motherhood (Ashcraft et al., 2013). For instance, miscarriage can occur due to factors related to poorer outcomes in general (e.g., smoking and alcohol consumption), whereas teen mothers may be expected to have worse outcomes than other women more generally, even if they had not had a child. Provided no other biases are present, the OLS and IV estimates therefore arguably bracket the true causal effect.

5.17.3 Opportunities

Whether the method can be applied in CLS's cohorts depends on the specific research question as the direction of known biases depends on information that is *not* captured in the data collection. Assessing this requires domain-specific knowledge. A key benefit of CLS's cohorts is the depth of data collection and the sample size, which mean multiple comparator groups can be defined.

5.18 Differential Effects

5.18.1 Method

The negative control exposure design exploits situations where the same set of confounders has effects on multiple 'exposures', only some of which plausibly have a causal effect upon the outcome of interest. Confounders like these are said to be "generic" (Rosenbaum, 2019), having multiple impacts on the world. For instance, a generalised disposition to risk may influence consumption of several drugs, meaning that people who are more likely to smoke crack cocaine may be more likely to inject heroin. A generalised interest in health could lead one to exercise more and follow a good diet.

The effects of a generic bias may be competing or the overlap between them incomplete – for instance, one may become addicted to one drug, but not another. Those who perform just one of these behaviours may be more similar in terms of the confounding factor to those who perform only the other than to those who perform neither, but in standard observational research both would appear in the comparison group. Instead of comparing exposed vs. non-exposed groups, the differential effect design (Rosenbaum, 2006, 2013) compares individuals exposed to only one of two factors in lieu of the other to reduce imbalances on generic confounders – e.g., people who exercise but eat poorly may be more similar to people who eat well but do not exercise than exercisers in general are to non-exercisers in general.

The differential effects design is another expression of the principle Zubizarreta et al. (2018) term 'isolation' – a smaller comparison is made that is less biased than the overall association in the wider system. Two examples discussed by Paul Rosenbaum in *Design of Observational Studies* (Rosenbaum, 2020) clarify the differential effects design as well as highlight its limitations:

1. Non-steroidal anti-inflammatory drugs, such as ibuprofen, may reduce the risk of Alzheimer's, but pre-clinical Alzheimer's may reduce pain sensitivity leading to lower ibuprofen use. However, pain sensitivity should affect the use of other analgesics, too, including ones that are not anti-inflammatory, such as paracetamol. Comparing exclusive ibuprofen users with exclusive paracetamol users is useful for gauging the effect of anti-inflammatory drugs on Alzheimer's as this would likely reduce imbalance on pain sensitivity, though other confounders may remain.[52]
2. Understanding the effect of cannabis use during pregnancy on child development is challenging as mothers who smoke cannabis during pregnancy differ sharply from those who do not (Columns 2, 3 & 5, Table 5.6). Cannabis users are more similar to those who use crack cocaine than those who use no drug, however; comparing exclusive cannabis users with exclusive crack cocaine

[52] Note, preferences are sometimes used as Instrumental Variables (e.g., in the judge fixed effect or physician preference designs).

users leads to an analytical sample less imbalanced on potential confounders (Columns 3-4, Table 5.6). Nevertheless, maternal crack cocaine use likely has an independent effect on childhood development.

Table 5.6: Characteristics of mothers by drug use during pregnancy, from an example differential effects study (figures from Rosenbaum, 2020). The differential effects design compares mothers who used only cannabis with those who used only crack (shaded columns), who resemble each other more closely than either resembles mothers who used neither drug.

	Neither	Cannabis	Crack	Both
Number of Babies	**931**	**11**	**39**	**5**
Cigarette use during pregnancy	22%	100%	92%	100%
Alcohol use during pregnancy	23%	64%	74%	100%
Late or no prenatal care	10%	27%	48%	40%
Less than high school education	23%	54%	49%	60%
Married	34%	0%	3%	0%
Did not want to get pregnant	32%	46%	59%	80%

By comparing two exposures, the 'effect' identified is not directly informative of the effect of an exposure vs. its absence. However, this may still be useful if the comparison factor does not plausibly have an effect on the outcome (i.e., it is a type of placebo test; Example #1) or where the effect of one is known and so the effect of the other can be back-transformed, or where the direction of effect of the comparator is known and exceeding this is informative (similar to reasoning in Biases of Known Direction, Section 5.17). The differential effect design can be particularly beneficial in drug safety research where side-effects for a specific health complaint may only be expected to exist for one

treatment under study (again, as with Example #1). Regardless, the information gleaned from the differential effects design may be useful as part of a strategy of triangulation. An additional limitation of the design is that, where behaviours are rare, by focusing on subgroups of the population, statistical power may be much lower.

5.18.2 Applications

Applications of this method using any data appear to be few and we identified none using CLS data – according to Google Scholar, the Rosenbaum (2006) paper only has 47 citations (as of September 2026), many of which are other methodological papers.

5.18.3 Opportunities

Several behaviours plausibly reflect common traits and impulses – e.g., health-seeking behaviour or risk aversion, in general. Given the variety of behaviours (e.g., on drug use) measured in the cohorts, there may be many possible applications. However, as the differential effects design uses only subsets of the data, the opportunities to obtain sufficient sample sizes with the method in CLS data may be fewer, though will increase if the cohorts are combined with each other or with similar studies.

5.19 Reasons for Effects

5.19.1 Method

Theories that predict a causal effect of an exposure on an outcome typically invoke a set of mechanisms linking the two. One way of making a theory elaborate is to consider intermediate effects. For instance, the theory that cognitive ability improves health by influencing health-related decision making assumes that (a) cognitive ability influences health-related decision making and that (b) decision making influences health. While it may be difficult to directly empirically assess the causal effects of cognitive ability upon health, in general, these 'reasons for effects' provide extra targets for empirical study (Rosenbaum, 2005, 2020). A causal theory may fail at the first step, if no link between intermediate factors can be shown. For instance, one may discard a hypothesis that blood alcohol limits reduce car fatalities by disincentivising drink-driving, if these limits have no observed effects on drivers' drinking behaviour.

One advantage of subjecting reasons to empirical scrutiny is that certain links in a causal chain may be more amenable to causal effect estimation or to more persuasive testing than others. For instance, in their sociological study of diabetes patients, Lutfey & Freese (2005) provide compelling evidence of the difficulties individuals with lower cognitive resources have in managing the illness. From there, if one is interested in the effect of cognitive ability upon diabetes survival, one needs to establish that poor disease management negatively affects prognoses, which some might consider self-evident. Consider also papers that repeatedly show that youth unemployment is related to poorer mental health later in life (see L. Wright, 2021, for a review). This association is hypothesised to partly reflect youth unemployment having a deleterious long-term effect on labour market outcomes – i.e., individuals who were previously unemployed may have poorer mental health later because they are more likely to earn low wages at the point of measurement of mental health. One would therefore expect youth unemployment to negatively affect labour market prospects. This can be tested experimentally, albeit in a somewhat artificial way, with correspondence studies – a type of field experiment in which curricula vitae (CVs) are sent to real-world employers and interview offers recorded. These CVs are manipulated and sent randomly to determine the effect of applicant attributes on job prospects. For instance, CVs may vary in whether an individual has experienced an episode of unemployment (Eriksson & Rooth, 2014). Thus, while the main estimate of interest may be biased, one can still have good, empirically well-supported reasons for thinking a causal effect exists. Indeed, because of their expense, RCTs frequently use surrogate outcomes – outcomes preceding the true outcome of interest (e.g., blood pressure in an RCT of a drug designed to reduce cardiovascular disease mortality) – on the logic that established 'reasons for effects' will translate into clinical outcomes.

Reasons for effects are conceptually similar to Bradford Hill's (1965) 'criterion' of "biologic plausibility" and form further pieces of evidence to assess whether an association of interest is causal. To be clear, here the object of inquiry is independent assessment of the effect of the exposure upon intermediate outcomes, or intermediate outcomes upon final outcomes or even other intermediate outcomes, in the causal chain, rather than performing a mediation analysis.

There are four issues with solely relying on 'reasons for effects'. First, there may be several mechanisms of action – cognitive ability is proposed to influence health not just through decision making, but also by affecting socioeconomic position (Deary et al., 2021). Ruling out one mechanism does not rule out transmission through other pathways (Gerring, 2010), including pathways that have not yet been thought of. Second, mechanisms may counteract, producing an unknown aggregate effect – this indeterminacy may in fact be motivation for assessing total causal effects empirically in the first place. Third (and related), assessing reasons for effects does not straightforwardly translate into determining effect sizes. Rather, it most straightforwardly informs qualitative judgements about a causal effect being present or not, limiting its usefulness, especially given effects may need to be of a certain minimum size to be practically important.

Fourth, the populations used to estimate effects at each step along a causal chain may vary and results may not generalise across these. It is possible that the overall chain may be broken for the target population of interest. For instance, one may observe a causal effect of youth unemployment on short-term employment prospects and a later effect of contemporaneous income on obesity at age 50 years, but by age 50 years, any labour market disadvantage arising from an episode of youth unemployment may be compensated for (e.g., through faster within-firm job promotions once a job is found).

5.19.2 Applications

Decades of research has shown that socioeconomic and psychosocial adversities are related to poorer health outcomes (Marmot, 2016). Major theories conceptualise adversities as stressors that, experienced chronically, alter and cause "wear and tear" of multiple systems within several of the body's regulatory systems (cardiovascular, metabolic, immune, etc.; Delpierre et al., 2016) and lead to a state of chronic inflammation. Wear and tear is operationalised as measures of "allostatic load", which combine measures of biomarkers across multiple systems, while chronic inflammation is most commonly indexed by elevated levels of C-reactive protein (CRP). Several studies have used data from the biomedical sweep of the NCDS to examine associations between socioeconomic (Barboza Solís et al., 2016) and psychosocial

(e.g., adverse childhood experiences; Barboza Solís et al., 2015; M. Chen & Lacey, 2018) adversity and allostatic load or CRP, in an effort to provide mechanistic understanding of the association between these adversities and clinical health outcomes. While these studies show such associations, issues are that (1) chronic stress is only one pathway and (2) the same factors that may confound the association between adversity and health plausibly also confound the association between adversity and allostatic load and CRP. Counterclaims would make the same set of predictions; biases are replicated (Rosenbaum, 2001).

5.19.3 Opportunities

Given the rich data collection reflected in CLS's cohorts, data related to causal mechanisms may be available (e.g., on biomarkers, health behaviours).

5.20 Quantitative Bias Analysis

5.20.1 Method

> There is a quantitative question. Cigarette smokers have a ninefold greater risk of developing lung cancer than nonsmokers, while over-two-pack-a-day smokers have at least a 60-fold greater risk. Any characteristic proposed as a measure of the postulated cause common to both smoking status and lung-cancer risk must therefore be at least ninefold more prevalent among cigarette smokers than among nonsmokers and at least 60-fold more prevalent among two-pack-a-day smokers. No such characteristic has yet been produced despite diligent search.
>
> Cornfield et al. (1959)

In Section 5.17, we discussed approaches to exploiting biases of known direction to support causal inference. These approaches incorporate information on bias qualitatively – e.g., to identify appropriate groups for study, contexts to draw data from, or to ascertain in which direction an estimate is likely to be biased. A drawback with this approach is that it does not allow the analyst to determine by how much an estimate is likely to be biased or what an unbiased estimate would plausibly look like. It is also not useful in situations where bias runs in the same direction as the anticipated causal

effect or where the causal effect is hard to predict *a priori* – in these cases, a given estimate would be compatible with a null or opposite-signed effect.

As the name suggests, quantitative bias analysis (QBA) instead incorporates or produces quantitative information on bias. QBA methods allow analysts to produce bias-corrected estimates or to ascertain the size of bias necessary to overturn a substantive result. This makes better use of the information the analyst may have – for instance, analysts may be able to make reasonable assumptions about the size of bias, as well as its direction. By providing data on the likely material effect of bias, QBA can support better decision making, enable more plausible comparisons of alternatives, and inform the direction of research resources to areas where the expected consequences of reducing uncertainty in a biasing factor would be greatest. QBA can also focus disagreement (Fox et al., 2021): instead of unproductive debates over whether an estimate is or is not biased, discussion can centre on whether bias is plausibly sufficiently strong in practice or whether it has been appropriately operationalised by the analyst performing the QBA.

The term QBA denotes a broad aim rather than a single method. An arsenal of QBA methods exists, though not all are advertised under this name – sensitivity analysis is a popular alternative term (Cinelli & Hazlett, 2020; Rosenbaum, 2020; VanderWeele & Ding, 2017), though this is also sometimes unhelpfully used synonymously with 'robustness checks' – additional analyses performed to check that incidental analytical decisions (e.g., the coding of variables) do not influence results. There are QBA methods for measurement error, selection bias, and confounding; some can be combined to deal with each within a single analysis (Fox et al., 2021). QBA methods vary in their complexity, ranging from *post hoc* deterministic methods applied to summary data (e.g., regression coefficients) to simulation-based stochastic methods that adjust record-level data repeatedly using random draws from hypothesised distributions for biasing factors.

Methods also vary in what they produce. Some methods – like the E-Value (VanderWeele & Ding, 2017) – produce bounds on the size of bias required to entirely explain an observed association; larger bounds mean stronger – potentially implausibly strong – bias is required to explain an association absent a causal effect. Other

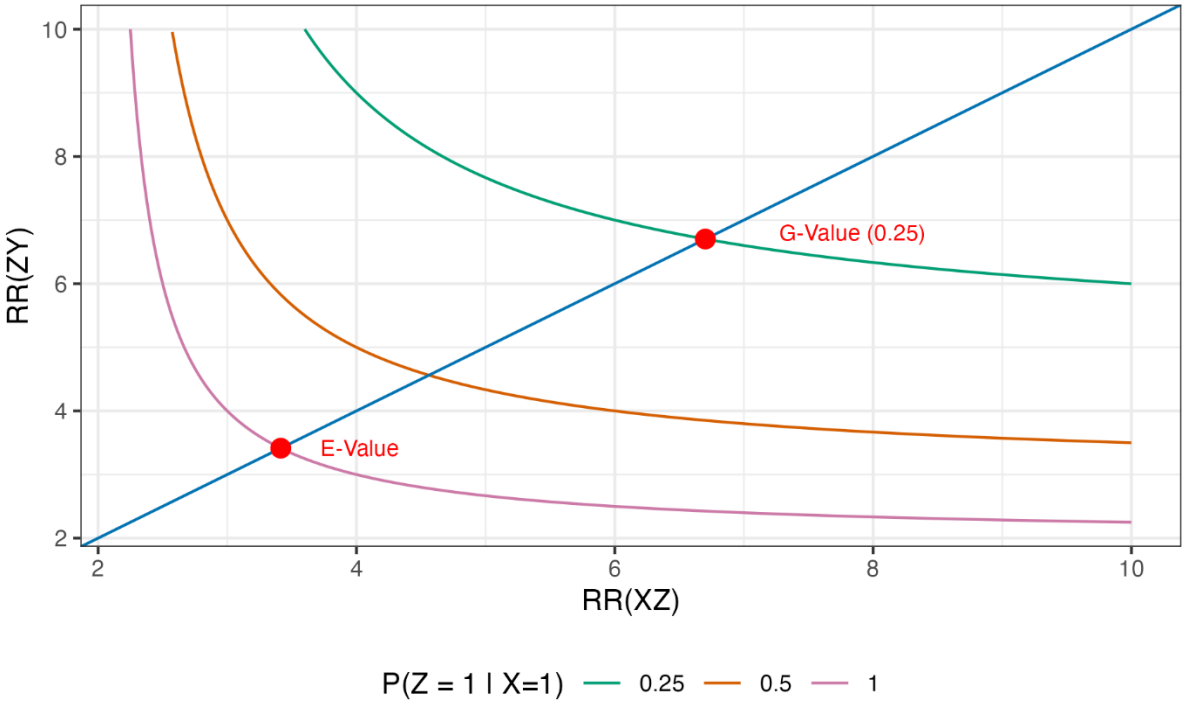
methods – like Pattern Mixture Modelling (Leurent et al., 2018; C. D. Nguyen et al., 2025; Tompsett et al., 2018) – can produce a bias-corrected estimate reflecting the researcher's best guess at what the causal effect is, given assumptions put into the QBA model. These methods can also be used for *tipping-point* type analyses to assess whether only a large degree of bias could explain an association (Y. Liu et al., 2025). Given the large number of QBA methods, what follows is a selective review.

5.20.1.1 Methods using Summary-Level Data

As discussed in Section 1, observational studies formed a major source of the evidence in the discovery of the harmful effect of smoking on lung cancer. One reason observational evidence was compelling was the size of the association between smoking and lung cancer observed. In one of the earliest quantitative bias analyses, Cornfield et al. (1959) derived lower bounds on the associations a single binary confounder must have with a binary exposure and binary outcome to fully explain an observed association between the two. Specifically, they show that the exposure-outcome association must be equal to or exceeded by both the confounder-outcome and confounder-exposure associations, rules now referred to as the 'Cornfield conditions'.

One corollary of their formulation is that the closer the confounder-outcome association is in size to the exposure-outcome association, the stronger the confounder-exposure association must be (and vice versa). This trade-off between confounder-exposure and confounder-outcome associations necessary to explain an exposure-outcome relationship follows a simple equation and can be plotted as contours as in Figure 5.19 – see the figure caption for further details.

Figure 5.19: Graphical representation of the 'Cornfield Conditions' (Cornfield et al., 1959) for a binary exposure (X), outcome (Y) and confounder (Z). For a given risk ratio (RR = 2) between exposure and outcome (i.e., P(Y = 1 | X = 1)/P(Y = 1 | X = 0)) and probability Z = 1 given X = 1 (green, orange, and pink lines), the graph shows the combinations of association between confounder and exposure (RR[XZ], x-axis) and confounder and outcome (RR[ZY], y-axis) required to fully explain the association. If the confounder-exposure association is weaker (e.g., RR[XZ] ↓ or P(Z = 1 | X = 1) ↓), then the confounder-outcome association must be stronger to explain the same result.



The Cornfield conditions formalise several intuitions, including:

1. Stronger exposure-outcome associations require a greater degree of confounding to explain results.
2. Confounders must have at least as large a causal effect on the outcome as the observed exposure-outcome association.
3. The strength of effect of the confounder on the outcome must be greater if the effect of the confounder on the exposure becomes weaker (and vice versa).

Using their bounds, Cornfield et al. (1959) offered an unanswerable test to sceptics of the link between smoking and cancer (among whom was Ronald Fisher): no plausible

confounding factors existed that could explain the very large association between the two.

The Cornfield conditions further formalise the intuition that there must be strong imbalance in absolute terms in the prevalence of the confounder between levels of the exposure; otherwise the confounder must have a particularly strong impact on the outcome to compensate (compare the pink, orange and green lines for a given RR(XZ) in Figure 5.19). To give an applied example, parental problematic drug use may be a confounder in the association between parental SES and childhood development, but it is (thankfully) a rare occurrence, so is unlikely to explain much of the association between the two.

The E-Value (VanderWeele & Ding, 2017), a widely used QBA method, builds upon the Cornfield conditions by making two assumptions: first, that the treatment-confounder and confounder-outcome effects are equal, and second, that the prevalence of the confounder among exposed is 100%. The E-Value is represented as a point at the intersection of the pink (P(Z = 1 | X = 1) = 1) and blue (RR(XZ) = RR(ZY)) lines in Figure 5.19. As Fox et al. (2021) and others (e.g., Poole, 2020) note, however, the assumption of complete prevalence of the confounder among the exposed is generally implausible and conservative. Fox et al. (2021) thus introduce the 'G-Value', which generalises the E-Value to situations where prevalence of the confounder is lower – Figure 5.19 also displays the G-Value(0.25) (intersection of green and blue lines) where the prevalence of the confounder among treated is 25%. As can be seen, the G-Value is larger than a given E-Value; the confounder has to have a stronger effect on the outcome because it is now less prevalent among the exposed group.

Though both the E-Value and G-Value allow bounds to be represented in a single scalar value, the price of this is the assumption that confounder-exposure and confounder-outcome associations are equal. This is not likely to be plausible in practice. It is instead preferable to posit values of each separately and to determine the consequences of this for bias, as enabled by drawing contours as in Figure 5.19. The utility of the Cornfield conditions and the E-Value and G-Value is that they can be used as a basis for rigorous argument about confounders, including providing a frame for reviews of existing evidence on plausible confounding variables. It is not sufficient to

say a confounder exists – it needs to have particular properties. However, this does not appear to be how E-Values and the like are used in practice. In their review of the use of E-Values, Blum et al. (2020) report that few authors assessed whether confounding was likely to overturn results or even noted candidate confounders likely to be missing from analyses.

A further limitation of the E-Value concerns multiple confounders – the E-Value relates to the combined effect of these, so several confounders that are individually too weak to explain away an association may jointly do so (Ioannidis et al., 2019). This joint strength can be particularly difficult to ascertain where individual unmeasured confounders are correlated. On the other hand, the degree of residual confounding depends on the confounders that have been controlled for – some of these (e.g., age and sex) can be strongly related to unobserved confounders, and so partly operate as proxies for these variables (Greenland, 2020). Together, these issues mean the E-Value can give an overly optimistic or overly pessimistic view of sensitivity to confounding bias (Sjölander & Greenland, 2022).

Though developed for settings with binary variables, the E-Value can be produced for continuous variables too, using heuristics for converting standardised beta coefficients to odds ratios (Mathur et al., 2025). Methods have also been developed for settings with continuous outcomes, specifically. Omitted variable bias (OVB) follows a simple formula in linear regression settings: it is the product of the regression coefficient for the confounder had it been included in the regression model and the coefficient from a regression of the confounder on the exposure had it been performed.

True Model: $Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot Z + \varepsilon$

Estimated Model: $Y = \widehat{\beta_0} + \widehat{\beta_1} \cdot X + \hat{\varepsilon}$

Decomposition: $\widehat{\beta_1} = \beta_1 + \beta_2 \cdot \delta_1$ (True Effect + Omitted Variable Bias)

Where δ_1 is the solution to regression: $Z = \delta_0 + \delta_1 \cdot X + \mu$

Note, δ_1 is the solution to the regression of Z upon X, despite this running counter to the direction of hypothesised causality (i.e., X is dependent on Z, rather than the other way round). However, this formulation makes intuitive sense: the observed association

between X and Y is equal to the causal effect of X upon Y (β_1) plus the effect of Z upon Y multiplied by the association a one-unit difference in X has with Z. This formulation has another advantage: even where Z has a large per-unit effect on X, δ_1 can be small if Z accounts for little of the variance in X. Again, absolute effect size is key for driving OVB.

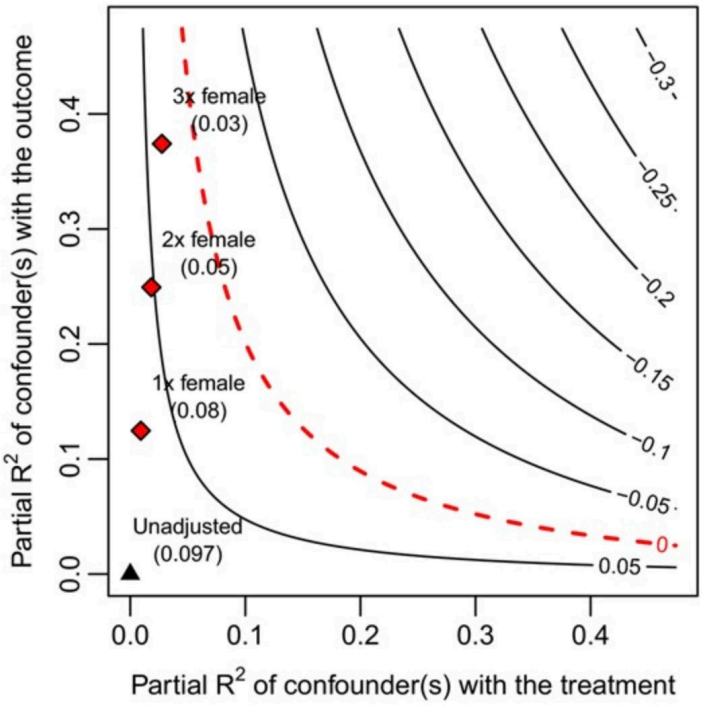
The OVB formula, $\beta_2 \cdot \delta_1$, can be visualised in a contour plot similar to the Cornfield conditions in Figure 5.19. An issue, however, is that the standard OVB formulation depends on the scales used for the exposure and outcome and so identifying appropriate literature on confounder-exposure associations may be difficult. Additionally, the formula does not easily generalise to situations with multiple confounders. To address this, Cinelli & Hazlett (2020) reparameterise OVB in terms of partial R^2 (i.e., the proportion of residual variance in exposure and outcome a confounder explains after conditioning on covariates).[53] With this formula, analysts can determine the required confounding associations to materially change an observed association between exposure and outcome (e.g., attenuate to the null), again assisted aided by contour plots.

Cinelli & Hazlett's (2020) method is implemented in the `sensemakr` R package (Cinelli, Ferwerda, et al., 2024); see Figure 5.20 for an example. The `sensemakr` package outputs additional information to contextualise analysts' consideration of whether unobserved confounders can plausibly explain results. This includes a 'robustness value' (for both the point estimate and lower confidence interval), which – analogous to the E-Value – specifies the minimum partial R^2 that would be required to attenuate the observed association to the null, assuming equal partial R^2 for both the confounder-exposure and confounder-outcome associations. It also includes information on the associations between *observed* confounders and exposure and outcome. For instance, in their working example on the effect of experiencing violence during the Darfur war on attitudes to peace (displayed in Figure 5.20), the authors benchmark the effect of the

[53] As partial R^2 are directly related to the (partial) correlation between two variables, the partial R^2 is symmetric. (Note in the equation above, OVB is expressed in terms of a regression of the confounder upon the exposure, even though the assumption is that the confounder is the cause of the exposure.)

supposed confounder (whether the respondent lived in the *centre* of a village) against sex, showing that *centre* would need to have multiple times stronger association with outcome and treatment than sex to explain results ('[1-3]x female' sit below the red-dashed line of null association in Figure 5.20). This is helpful because, while not all confounders may be observed, major ones may be. Using the latter to calibrate expectations can therefore help interpretation of the QBA.

Figure 5.20: Contour plot of a Cinelli & Hazlett (2020) type sensitivity analysis. Drawn from the Cinelli & Hazlett (2020) running example examining the potential for confounding in the association between experiencing violence during the Darfur war and attitudes to peace. The authors hypothesise that this association is confounded by whether the survey respondent lived in the centre of the village (more likely to be victimised) or at the periphery, and use the variable 'female' to contextualise discussion of whether confounding would be sufficient to reduce the observed association (0.097) to the null. If the confounder were to have 3 times the strength of association with violence victimisation and peace attitudes as female does, this would still not be sufficient to reduce the point estimate to the null (3x female [0.03]). Cinelli & Hazlett (2020) type sensitivity analyses can be performed with the sensemakr R package (Cinelli, Ferwerda, et al., 2024), which also outputs 'robustness value' statistics.



Oster (2019) provides another formulation of bias from omitted variables that, like Cinelli & Hazlett (2020), uses observable information from regressions including

observed confounders, but also uses information on the "stability" of coefficients when adding observed confounders to models. Oster's solution parameterises OVB as a function of four observable and two unobserved quantities: the model R^2 from (a) a "short" regression of Y on X (denoted $\dot{R}$), (b) a regression of Y on X plus observed confounders ($\tilde{R}$), and (c) a regression of Y on X plus observed and unobserved confounders, would it be possible to run this (R_{MAX}); coefficients for X in regressions (a) ($\dot{\beta}$) and (b) ($\tilde{\beta}$); and a parameter, δ, capturing "proportional selection" – the association of unobserved confounders with treatment relative to that of observed confounders (by necessity, this is also unobserved).

$$\beta^* \approx \tilde{\beta} - \delta \cdot (\dot{\beta} - \tilde{\beta}) \cdot \frac{R_{MAX} - \tilde{R}}{\tilde{R} - \dot{R}}$$

... where β^* is the coefficient on X from a regression of Y upon X plus observed and unobserved confounders. δ captures imbalance in the unobserved confounders across exposure levels (though as a function of imbalance in the observed confounders) and $R_{MAX} - \tilde{R}$ captures the impact of confounders on the outcome. All else being equal, greater proportional selection and larger effects of the confounders on the outcome will bias coefficients more.

Using this formula, one can reason about the combinations of R_{MAX} and δ that would reduce β^* to an important value (e.g., the null, $\beta^* = 0$) or to determine what plausible values of R_{MAX} and δ would imply for β^*. A key development in Oster's (2019) paper is that she identifies heuristics for R_{MAX} and δ by applying her formula to the results of RCTs, among other analyses. Oster suggests values of $R_{MAX} = \min(1, 1.3 \cdot \tilde{R})$ and δ = 1 as sensible bounds.

The Cinelli & Hazlett (2020), Oster (2019), E-Value, G-Value, and Cornfield Condition methods each exploit summary-level data and are referred to as 'simple sensitivity analysis' techniques. Each is generally used to evaluate bias from unobserved confounding, though can be extended to assess other types of bias, including selection bias or bias from violations of the assumptions of instrumental variables analysis (Mathur et al., 2025; Mellon, 2025). Other QBA methods to obtain bounds that use summary-level data exist, including for (e.g.) bias from differential measurement error

(VanderWeele & Li, 2019). We will not cover these methods further here (for more, see Fox et al., 2021).

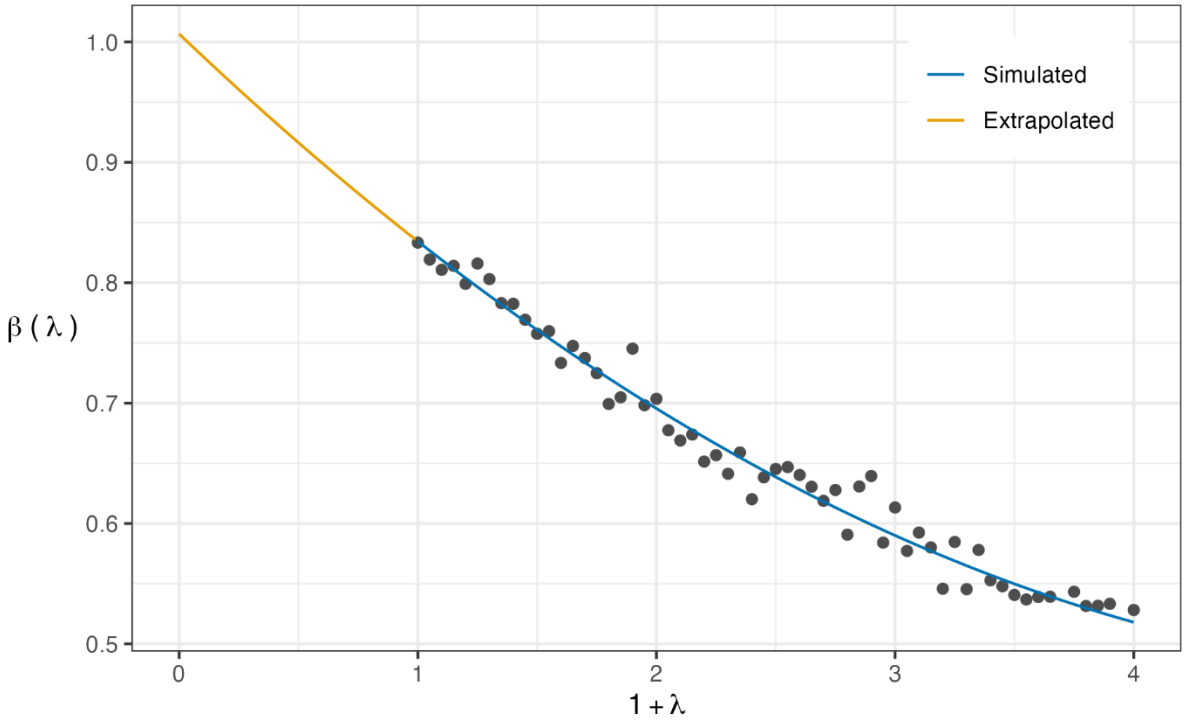
Another summary-level approach that has been used extensively to account for bias is regression calibration (Keogh & White, 2014), though in this case the purpose is to provide a bias-corrected estimate. Regression calibration applies where there is regression dilution – classical measurement error in the exposure, which would bias the estimate, most commonly towards the null. This attenuation is predictable: in linear regression, the observed association will be attenuated by a factor equal to the measurement's reliability (i.e., the correlation between two independent measurements of the same variable). The correction is therefore straightforward – one can divide the estimated coefficient by assumed reliability – though more complex ways of incorporating this information can be performed (e.g., by treating reliability as an uncertain estimated quantity with its own distribution). However, this approach will generate biased results when there are unobserved or mis-measured confounders (Davey Smith & Phillips, 1996; Phillips & Davey Smith, 1991) as the uncorrected estimate will be subject to other biases, rather than just the regression dilution.

5.20.1.2 Methods using Record-Level Data

Another popular method that corrects for measurement error but instead uses record-level data is Simulation and Extrapolation (SIMEX; Cook & Stefanski, 1994). In SIMEX, analyses are re-run *adding* random measurement error to (continuous) variables.[54] Progressively increasing error and then drawing a line of best fit through estimates, before extrapolating back, allows one to predict a situation where no measurement error exists (see Figure 5.21). SIMEX can handle situations where it is confounders, rather than the exposure itself, that are subject to measurement error, but accurate extrapolation requires (a) correct assumptions about the reliability of the initial measurement (which determines how far a line of best fit should be drawn backwards to reflect a situation of no measurement error) and (b) use of an appropriate extrapolation function, which is an inherently uncertain choice.

[54] Küchenhoff et al. (2006) extend SIMEX to categorical data.

Figure 5.21: Results of an example SIMEX analysis. SIMEX consists of two steps. First, one reruns analyses, progressively adding random measurement error to the variable hypothesised to be subject to measurement error, in proportion to the amount of error anticipated in the raw variable already. Next, one estimates a line of best fit through the estimates obtained from Step 1 (blue line) and then extrapolates this line back to the point where no measurement error would exist to obtain the measurement-error-corrected estimate (orange line for x-axis = 0).



Systematic or differential measurement error can be corrected for via simulation using record-level data (Gallop & Weschle, 2019; Pina-Sánchez et al., 2023; L. Wright, Ploubidis, et al., 2024). For instance, to account for 'mode effects' (a type of measurement error where responses vary systematically by mode; Section 2.2.5) in a mixed-mode study in which participants are not randomised to mode, external information on mode effects can be used to simulate values had participants responded in a single mode. For example, experimental evidence may indicate that individuals report 0.2 SD *better* mental health when responding via telephone rather than web (Goodman et al., 2022; Tomova, Silverwood, & Wright, 2026a). Subtracting 0.2 SD from observed mental health for those in the telephone mode would therefore provide (estimated) counterfactual values which can be used when fitting the substantive model of interest. This approach is very flexible – it can handle multiplicative, or other non-additive, errors and differential measurement errors (Pina-

Sánchez et al., 2023), e.g., by incorporating different-sized errors according to participant characteristics. The process can be repeated to explore sensitivity of the findings to different magnitudes of measurement error.

Another record-level simulation-based method is Pattern Mixture Modelling (PMM; Leurent et al., 2018; C. D. Nguyen et al., 2025; Tompsett et al., 2018). PMM is used to correct for violations of the Missing at Random Assumption used in multiple imputation (Section 5.4) – i.e., for selection bias. Post hoc adjustment is used to amend imputed values – for instance, if one anticipates missing values to in fact be 0.3 SD lower than imputed, imputed values can simply be corrected to reflect this and analyses otherwise run as usual (i.e., with estimates pooled with Rubin's Rules). As with simulation for non-random measurement error, the complexity of the PMM procedure can be increased – for instance, by choosing more complex patterns of MNAR (e.g., larger violations in some groups than in others). Corrections to MNAR should be performed within the 'Gibbs sampler' – i.e., within the imputation procedure, rather than once the imputed datasets have been created – so that chained equations models reflect beliefs about MNAR. However, this can be hard to code (some limited functionality exists in the mice R package; C. D. Nguyen et al., 2025) and is more computationally demanding to run as it requires separate imputations for each MNAR assumption posited.

Simulation-based methods can incorporate assumptions on biases stochastically, rather than use single deterministic values – for instance, by reflecting uncertainty as distributions and repeating the simulation using draws. Fox et al. (2021) provide advice on turning beliefs about the size of biases into statistical distributions from which to draw. Often, it may be easier to rule *out* particular-sized biases, than rule biases *in* (for instance, mode effects are unlikely to exceed 0.5 SD, but may be within the range 0-0.2 SD; Tomova, Silverwood, & Wright, 2026a). In this case, simulations can be run in a tipping-point-like fashion, or large and implausible biases used to assess whether implausible levels of bias would be insufficient to change substantive conclusions.

Readers interested in QBA can see the textbook by Fox et al. (2021), which includes recommendations for performing and reporting QBA (see also Lash et al., 2014). However, note, as epidemiologists, the authors focus their exposition and examples on

cases with binary exposures and outcomes. For comprehensive discussion of QBA for measurement error, specifically, see Keogh et al. (2020) and Shaw et al. (2020).

5.20.2 Applications

Several papers have calculated E-Values as part of analyses of CLS data (see, e.g., Gondek, Bann, Brown, et al., 2021; Metherell et al., 2023; Moulton et al., 2023; Ning et al., 2021). For example, Gondek et al. (2021) use E-Values to understand the potential consequences of bias in an analysis of early mental health and later problematic alcohol use using data from the NCDS and BCS70. Though the E-Values were small, they note a paucity of existing evidence on the strength of the confounders they hypothesise, making it difficult to evaluate plausibility. Nevertheless, this evidence may appear in future, so performing the QBA provides information that may still be of use to future readers.

Dräger et al. (2025) perform an Oster (2019) type QBA of the association between school absences and attainment at GCSE using data from the MCS. They show that many of their results are robust to an assumption about relative selection on observables and explanatory power of observables for their outcome. However, this assumption is not well justified and is only compared against a much more conservative assumption, which produces prima facie implausible results for some model specifications.

Bann et al. (2023) use PMM when assessing cross-cohort differences in the association between height and cognitive ability using data from NCDS, BCS70 and MCS, given different levels of non-response and missingness in each of these studies. They perform a tipping-point-style analysis to assess whether differences in missingness mechanisms would be required to explain results.

5.20.3 Opportunities

As QBA methods are designed for situations where data are missing (e.g., on confounders), they typically do not make demands upon existing data. Therefore, there is ample opportunity to use these in CLS's cohorts. That said, as data on many major confounders are available, Cinelli & Hazlett (2020) type QBA may be especially compelling as these confounders can be used to calibrate assessments of unobserved confounding.

6. Discussion

Causal inference is one of the central goals of quantitative health and social science, but also one of its most difficult to attain. The foregoing has shown that there may be ample opportunities for making causal inferences using CLS cohort data and multiple different methods and broad approaches to utilise when doing so.

Viewed individually, each of these methods stands or falls on the validity of its assumptions. Validity depends on the specific situation to which the methods are being applied and usually cannot be ascertained directly. Determining whether assumptions are valid – or close enough to be valid as to still be useful – requires knowledge of a specific research domain and an ability to synthesise evidence and construct arguments, especially when realising that one's readers need convincing, too ('The first principle is that you must not fool yourself—and you are the easiest person to fool'; Feynman, 1974).[55] Argumentation about whether assumptions are met is necessary regardless of whether model-based approaches (*is residual variation as-if random?*), design-based approaches (*does this IV meet the exclusion restriction?*), or Elaborate Theory approaches (*is the mechanism of action weaker in this setting, as hypothesised?*) are used.

In any case, no single study is likely to be perfect. An RCT may accurately balance treated and control groups, but available measures may be inadequate, diluting estimates, or the intervention may be too narrow to generalise directly to situations of interest, necessitating the application of theory to justify any generalisation, perhaps informed by other empirical results (Deaton & Cartwright, 2018). Viewed as one component of a wider literature, however, imperfect studies can still make a contribution. For instance, a study may reduce uncertainty about a hypothesis by testing the effect of a confounder not yet considered in other studies. More formally, one study may have biases unrelated to those in another study, possibly because it

[55] The advice that is made harder to follow by the human brain's capacity to treat results as obvious in hindsight, and to forget uncertainty that existed before (Baratz, 1983; Yudkowsky, 2007).

used a different study design (e.g., a quasi-experiment rather than statistical control). If these studies nevertheless converge upon the same result, this is stronger evidence of the reliability of the causal effect estimate (Gutierrez et al., 2025; Lawlor et al., 2016). Such "triangulation" can involve evidence of different characters that may even, at first glance, appear only tangentially related. Part of the evidence for the effect of smoking on cancer involved the results of prospective cohort studies (Doll & Hill, 1954), but it also included ecological studies of societal trends in lung cancer mortality, lung autopsies of smokers who died from other causes, and experimental evidence from animal models (Cornfield et al., 1959). Added more recently is evidence from individuals with particular genetic disorders, who provide indirect evidence for later steps in a hypothesised biological pathway, which smoking is thought to precipitate (Rosenbaum, 2015).

Rosenbaum (2015) following Haack (2009) likens such reasoning to completing a "causal crossword". Considered in isolation, each piece of evidence supporting a causal inference may be weak, but together they lend mutual support, as two guesses intersecting with the same letter do for crossword players. Meditating on this and advice to "Make your theories elaborate", Rosenbaum (2015, p. 209) concludes "If you took [the advice to make your theories elaborate] seriously, then you would ask of each new study what it contributes to the currently incomplete, partly penciled-in puzzle. You would welcome the completion of a new entry, even a small entry, compatible with the current tentative completion. You would also welcome a compelling new entry that challenged some current entries. You would welcome the suggestion that a particular entry is mistaken and constitutes a barrier to correct completion of the puzzle. A pencil and an eraser would be two tools of equal importance." As with a crossword, two entries may not intersect – may be only tangentially related – but still contribute to the completed whole.

Completing causal crosswords would require quantitative researchers to clearly state their goals and the quantity they are trying to learn with their statistical analysis (the estimand; Lundberg et al., 2021). Several authors have offered powerful critiques of prevailing norms in non-experimental sciences against stating one's causal ambitions (Dahabreh & Bibbins-Domingo, 2024; Grosz et al., 2020; Hernán, 2018b; Labrecque &

Kezios, 2026). These norms encourage the use of imprecise euphemisms such as ‘association’ and allow researchers to claim an interest only description or prediction, when in fact that would motivate different analytical choices to those chosen (Grosz et al., 2020; Labrecque & Kezios, 2026). Hiding one’s aims makes productive disagreement more challenging and holds back cumulative scientific progress (Dahabreh & Bibbins-Domingo, 2024; Grosz et al., 2020). It also enables suboptimal practices, such as controlling for mediators in statistical models (Labrecque, 2023) and formulating poorly conceptualised research questions where the intervention is not well-defined (Hernán, 2018b). This taboo has led to the curious practice of researchers cautioning against causal interpretations in their studies while, at the same time, assuming causality when drawing policy and action recommendations from their work (Bann et al., 2025; Haber et al., 2022) – a practice that has been referred to as “Schrödinger’s inference” (Tennant & Murray, 2021).

Authors critiquing current practices have offered several recommendations for the conduct and reporting of studies that aim to generate evidence for causal inference (Dahabreh & Bibbins-Domingo, 2024; Grosz et al., 2020; Haber et al., 2022; Hernán, 2018b; Labrecque & Kezios, 2026). For instance, Hernán (2018b) argues causal language should be used throughout papers except in the results section, where the language of association should instead be used, reflecting that statistical estimates acquire a causal interpretation only through the arguments made elsewhere in the paper. Others have produced “roadmaps” for designing causal inference studies – checklists and processes going from defining a coherent research question and theoretical estimands to choice of appropriate statistical estimator, quantitative bias analyses and falsification tests, and judicious interpretation of results (Dang et al., 2023; Dang & Balzer, 2023; Petersen & van der Laan, 2014).

All of which is to say: beyond simply being a set of methods, causal inference requires an ability to ask good questions, to synthesise disparate lines of evidence to come to reasoned judgements, to be transparent about the plausibility of underlying assumptions, and to be honest about uncertainty. CLS data have many attributes that can support causal inference, but the data alone are not enough.

7. Statements

7.1 Declaration of interest

All authors declare no conflicts of interest.

7.2 Funding

This work was supported by the Economic and Social Research Council (ES/W013142/1).

7.3 Acknowledgements

Claude (Fable 5.1 and Opus 5.5) and OpenAI (GPT-6 Astra) were used as reviewers, suggesting grammatical, formatting and substantive edits to text, which were reviewed by LW before approval or rejection. Claude Opus 5.5 was also used to generate draft alt-text for figures and tables.

7.4 Author contributions

LW and RJS conceptualised the user guide. LW wrote the first draft, except for Sections 5.12 and 5.13, which were written by MD, and Section 2.1, which was written by GDT. All authors provided critical revisions.

7.5 Data availability

The code used to create plots and DAGs in this document is available at https://osf.io/ta5bk.

8. Appendix

8.1 Correlation ≠ Causation: A Short Empirical Review

Nature is replete with examples of correlations between variables that clearly do not result from causal effects of one variable upon the other; see Vigen (2025) for a humorous list. Associations identified in observational studies that persist even after adjusting for known confounding factors may be more plausible and anticipated by existing theories (or less rigorously, explainable post hoc; Bishop, 2019), but still be incorrect, regardless. We can be confident this is the case because RCTs regularly produce results that are discordant with those in standard observational studies. Moreover, observational studies can produce estimates of causal effects that are implausible *prima facie*, including cases where similar procedures (e.g., statistical adjustments for major hypothesised confounders) are followed that identify estimates that appear more plausible.

Given medicine's widespread use of randomised controlled trials, much of the evidence on the discordance between observational studies and RCTs comes from the medical literature. The history of epidemiological research includes multiple examples of factors that, based on observational studies, were thought to cause, or protect from, disease, but have been subsequently shown in RCTs to act otherwise. This includes factors that, despite showing initial promise, have in RCTs produced only null results; for instance, Vitamin C was thought to reduce the risk of cardiovascular disease but large RCTs have shown no evidence of effect (Lawlor et al., 2004). More troublingly, factors that have appeared protective in observational studies have turned out in RCTs to be harmful, including cases where observational evidence had been used to inform medical recommendations and clinical practice. This includes HDL ("good") cholesterol and β-carotene. For instance, in observational studies, β-carotene was found to be related to lower risk of cancer, but in fact increases risk among smokers (Davey Smith & Ebrahim, 2001). Dispiriting results from RCTs have led prominent epidemiologists to wonder, tongue in cheek, whether it was "time to call it a day" on their field (Davey Smith & Ebrahim, 2001).

High-profile examples comport with more general findings. Meta-analyses comparing observational associations with effects from RCTs in studies of therapeutic interventions (Golder et al., 2011; Papanikolaou et al., 2006; Toews et al., 2024) or disease risk factors (Janiaud et al., 2021) show small average differences between the two designs, but this masks substantial heterogeneity, including cases of statistically significant differences in effects, variations in effect size of double or more, and instances of reversal of sign (Ioannidis, 2001).[56] Similar patterns have been found in economics (Glazerman et al., 2003), while in education, (non-randomised) interventional studies have a mean effect size exceeding 0.4 SD, compared with an average effect size of just 0.06 SD in well-conducted RCTs (Lortie-Forgues & Inglis, 2019).[57] Studies comparing observational and quasi-experimental estimates of the effects of educational attainment also show a variety of discordant effects (Davies et al., 2023). Though there is evidence that longitudinal or prospective cohort studies show smaller discrepancies with RCTs than other observational study designs, bias is observed in these types of study, too (Bröckelmann et al., 2022; Glazerman et al., 2003; Toews et al., 2024).

A crucial feature of studies comparing RCTs with observational studies is that, in some, authors have noted the difficulty of predicting the size and direction of bias *a priori* and of finding statistical methods that reliably eliminate bias (Glazerman et al., 2003; Ioannidis, 2001; Janiaud et al., 2021). While studies often differ in results, this would not matter so much if the size of this difference were predictable – a clock that has yet to be set to daylight saving time can be corrected *post hoc*. Nevertheless, while at an aggregate level, bias may be difficult to predict, there are several research areas where bias is expected to operate in a single direction. As noted, observational studies of educational interventions typically overestimate effect sizes, and there are other fields where this is the case. Notably, social epidemiology studies investigating the effect of

[56] The cited studies are only a selection of those available. The blogger Gwern Branwen (2022) maintains a larger list.

[57] To understand how dispiriting this is, imagine you took a random person who received the intervention and compared them with a random person who did not. With an effect size of 0.4 SD, the intervention recipient would have the better outcome 61.1% of the time. With an effect size of 0.06 SD, this would drop to 51.7%, little better than a coin toss (i.e., 50%).

social status upon health can generally be presumed to find a stronger negative association between disadvantaged social status and health than reflects a causal effect of the former due to, among other things, health-related selection into social position. This presumption is supported by studies exploiting policy experiments that find comparatively small effect sizes of, e.g., income on health (Bann et al., 2024; Courtin et al., 2020), as well as the persistence of health inequalities in relatively economically equal societies (Mackenbach, 2019). Observational studies assessing the effect of the family environment upon children may also be expected to be biased towards finding larger effects than is correct due to the heritability of human traits and direct genetic transmission within biological families (Plomin, 2018; Rich Harris, 1999); this point will be expanded upon in Section 8.2. In both social epidemiology and family research, bias is particularly concerning as observational studies are expected to give estimates that are anti-conservative – i.e., larger than the true causal effect. An unbiased association could therefore be small or even null.

There are, however, sometimes substantively important reasons why results from observational studies and RCTs differ that do not reflect confounding bias in observational studies. This includes differences with *what* intervention is implemented, *how* and to *whom*, and in the target *estimand*. Individuals tasked with providing an intervention in an RCT may be more motivated and well-trained than those so tasked when the intervention is rolled out as policy in the real world, meaning fidelity to protocols may decrease (Kunz & Oxman, 1998). RCTs by necessity also only deal with interventions that are ethically, financially and physically feasible to implement, a constraint that does not apply to observational studies. Interventions in RCTs are thus typically more limited in scope than the corresponding comparisons made in observational studies and are often focused on quite specific groups of individuals (Bann et al., 2024). For instance, rather than examine a change in household income from bottom to top quintile among the general population, which a typical observational social epidemiology study would do, an RCT may instead investigate relatively small and short-term increases in welfare generosity among people, e.g., receiving sickness

benefit.[58] Nevertheless, it is still the case that observational studies can yield results that can only be plausibly explained by residual confounding or other biases – for instance, after controlling for several confounders, smoking remains associated with suicide (heavy vs. non-smokers, OR = 3.33; Davey Smith et al., 1992) and hormone replacement therapy use remains related to lower risk of mortality from accidents (RR = 0.5; Petitti et al., 1986). These findings demonstrate the potential for substantial bias in observational studies, though as noted in Section 5.11, some of this bias can be "self-inflicted" and follow from the analyst's decisions, rather than irredeemable issues of internal validity (confounding, selection, and information biases) in the underlying data.

8.2 Why Does Correlation (Often) Not Equal Causation?

The prevalence of observational studies generating different results to RCTs raises the question of why correlation often does not equal causation (Branwen, 2019). Several disciplinary perspectives are useful for understanding why confounding may arise in observational studies. Here we highlight two: economics and behaviour genetics.

The economics perspective highlights that individuals attempt to achieve their goals but differ in their preferences and the resources they have available (finances, cognitive skills, etc.). All actions have opportunity costs – resources expended on one action cannot be expended elsewhere – and individuals are in competition over finite resources. This competition usually resolves with the person who is willing to pay the most obtaining the resource – e.g., the person who is willing to pay the most for the nice house in the nice area is the one who gets it. In many situations, this will lead to positively valenced things becoming clustered – the family who can afford to move within the catchment area of the high-achieving school could also afford to move to the low-crime neighbourhood as well as afford to insulate and heat their home and to follow

[58] It should also be noted that there are meta-scientific reasons why results can differ between observational and interventional studies: specifically, p-hacking and publication bias. Though there is evidence of p-hacking and publication bias in RCTs (Dwan et al., 2013), funders of RCTs also often require pre-registration and publication of results (Lortie-Forgues & Inglis, 2019). In observational studies, a null result may be more easily dismissed as reflecting bias in the observational data and filed away.

the most healthful diet. In other situations, this will lead to attractive goods becoming anti-correlated as they are, to some extent, substitute goods – for instance, time spent on one activity means less time for another. Both of these processes can lead to confounding in observational data.

Another corollary of the economics perspective is that the people least harmed by a negative situation, or who derive the greatest benefit from a positive one, are the people who are most likely to select into or remain in that situation – for example, those whose disease prognosis is the worst may be most willing to accept the experimental treatment. (Prescribing based on prognosis is known as 'confounding by indication'.) Again, this can lead to confounding in observational data. Indeed, the very fact that, in observational data, two people differ on some exposure variable should give us pause before assuming they are otherwise comparable (Rosenbaum, 2020).

The behaviour genetics perspective highlights genetic influences on human traits and the direct transmission of DNA from parents to their children in (biological) families. This is supported by decades of research using twin and other genetically informed family studies showing that almost all human traits are heritable (Polderman et al., 2015), including unsurprising traits like height, but also complex, social traits like education and even divorce. The behaviour genetics perspective predicts that parents' and children's traits will be correlated, regardless of any environmental effect parents may have; children with highly educated parents are more likely to attain higher education themselves partly because they have inherited education-promoting genes from their parents. For compelling descriptions of this view, see Robert Plomin's (2018) book *Blueprint* and Judith Rich Harris' (1999) *The Nurture Assumption*.[59]

The behaviour genetics perspective also highlights gene-environment correlation (rGE) as a potential source of confounding in observational studies. rGE describes the processes by which genes and environments become correlated and comes in three

[59] Twin studies offer another reason to be sceptical that environmental exposures can have large causal effects – specifically the small role the shared environment (non-genetic factors shared between siblings) has in explaining trait variation, a phenomenon that has been referred to as the 'gloomy prospect' of behaviour genetics (Davey Smith, 2011; Plomin & Daniels, 1987).

forms: passive, active, and evocative. *Passive rGE* describes the process discussed above – children inherit genes from their parents as well as family environments, so these become correlated with no further intervention. This includes correlations between parents and their children for the same trait (e.g., education) but also correlations with other aspects of the family environment that are influenced by parental genes. For instance, parents with high reading ability may have more books in the home but will pass genes related to reading ability on to their children, whether their children read those books or not.

Active rGE describes the process by which people choose environments based on their genetically influenced proclivities – a child with high reading ability may spend their pocket money on books. *Evocative rGE* describes the process by which the environment and people within it respond to a person partly based on their genetic endowment – for instance, a child who is perceived to be struggling academically may elicit more support from teachers. In each of these cases, an observational study that assumed the direction of causality ran solely from the environmental factor to the child's trait would be in error.

One final genetics perspective relates to the relatively limited number of protein-coding genes in the human genome (approximately 20,000) compared with the incredible number of tasks a body carries out, and the wide and often continuous variation observed between individuals in their traits. Genome-wide association studies have demonstrated that most traits are polygenic (influenced by many genes each having small effects) and that genes are pleiotropic (implicated in many traits) (Chabris et al., 2015). Correlations between traits can therefore arise because of the varied action of genes. This includes correlations between traits that otherwise would not have an obvious link. For instance, height and schizophrenia have genetic overlap (Romero et al., 2025). The observation that "everything correlates to some extent with everything else" has been called the "crud factor" in psychological and sociological research (Meehl, 1990).

9. References

Abdellaoui, A., Dolan, C. V., Verweij, K. J. H., & Nivard, M. G. (2022). Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics*, *54*(9), Article 9. https://doi.org/10.1038/s41588-022-01158-0

Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K. E., Nivard, M. G., Veul, L., Holtz, Y., Zietsch, B. P., Frayling, T. M., Wray, N. R., Yang, J., Verweij, K. J. H., & Visscher, P. M. (2019). Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour*, *3*(12), 1332–1342. https://doi.org/10.1038/s41562-019-0757-5

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *The American Economic Review*, *91*(5), 1369–1401.

Adhikari, A., DeNero, J., & Wagner, D. (2024). *Computational and Inferential Thinking: The Foundations of Data Science* (2nd edn). https://inferentialthinking.com/

Anders, J., & Dorsett, R. (2017). What young English people do once they reach school-leaving age: A cross-cohort comparison for the last 30 years. *Longitudinal and Life Course Studies*, *8*(1), 75–103. https://doi.org/10.14301/llcs.v8i1.399

Andrews, R. M., & Didelez, V. (2021). Insights into the Cross-world Independence Assumption of Causal Mediation Analysis. *Epidemiology*, *32*(2), 209. https://doi.org/10.1097/EDE.0000000000001313

Andrews, R. M., Shpitser, I., Didelez, V., Chaves, P. H. M., Lopez, O. L., & Carlson, M. C. (2023). Examining the Causal Mediating Role of Cardiovascular Disease on the Effect of Subclinical Cardiovascular Disease on Cognitive Impairment via Separable Effects. *The Journals of Gerontology: Series A*, *78*(7), 1172–1178. https://doi.org/10.1093/gerona/glad077

Arnold, B. F., Ercumen, A., Benjamin-Chung, J., & Colford, J. M. (2016). Brief Report: Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies. *Epidemiology (Cambridge, Mass.)*, *27*(5), 637–641. https://doi.org/10.1097/EDE.0000000000000504

Ashcraft, A., Fernández-Val, I., & Lang, K. (2013). The Consequences of Teenage Childbearing: Consistent Estimates When Abortion Makes Miscarriage Nonrandom. *Economic Journal (London, England)*, *123*(571), 875–905. https://doi.org/10.1111/ecoj.12005

Astorga-Rojas, D. (2023). Should I stay or should I go? The effect of London's terrorist attack on the educational choices of Muslims. *SERIEs*, *14*(3), 463–501. https://doi.org/10.1007/s13209-023-00282-2

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, *116*(536), 1716–1730. https://doi.org/10.1080/01621459.2021.1891924

Atlas of Longitudinal Datasets. (2024). *Atlas of Longitudinal Datasets*. https://atlaslongitudinaldatasets.ac.uk/

Avendano, M., & Panico, L. (2018). Do flexible work policies improve parents' health? A natural experiment based on the UK Millennium Cohort Study. *J Epidemiol Community Health*, *72*(3), 244–251. https://doi.org/10.1136/jech-2017-209847

Bailey, D. H., Jung, A. J., Beltz, A. M., Eronen, M. I., Gische, C., Hamaker, E. L., Kording, K. P., Lebel, C., Lindquist, M. A., Moeller, J., Razi, A., Rohrer, J. M., Zhang, B., & Murayama, K. (2024). Causal inference on human behaviour. *Nature Human Behaviour*, *8*(8), 1448–1459. https://doi.org/10.1038/s41562-024-01939-z

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, *33*(13), 2297–2340. https://doi.org/10.1002/sim.6128

Banack, H. R., & Kaufman, J. S. (2014). The obesity paradox: Understanding the effect of obesity on mortality among individuals with cardiovascular disease. *Preventive Medicine*, *62*, 96–102. https://doi.org/10.1016/j.ypmed.2014.02.003

Bang, H., & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bann, D., Wang, M., Davies, N. M., Wright, L., O'Connor, M., & Courtin, E. (2025). *Science or Advocacy? The Global Rise of Policy Claims in Population Health Research (1990-2024)* (p. 2025.11.13.25340175). medRxiv. https://doi.org/10.1101/2025.11.13.25340175

Bann, D., Wright, L., Davies, N. M., & Moulton, V. (2023). Weakening of the cognition and height association from 1957 to 2018: Findings from four British birth cohort studies. *eLife*, *12*, e81099. https://doi.org/10.7554/eLife.81099

Bann, D., Wright, L., Hughes, A., & Chaturvedi, N. (2024). Socioeconomic inequalities in cardiovascular disease: A causal perspective. *Nature Reviews Cardiology*, *21*(4), 238–249. https://doi.org/10.1038/s41569-023-00941-8

Baratz, D. (1983). *How Justified Is the 'Obvious' Reaction* [Ph.D., Stanford University]. https://www.proquest.com/docview/303182481/abstract/152AD577CB914075PQ/1

Barboza Solís, C., Fantin, R., Castagné, R., Lang, T., Delpierre, C., & Kelly-Irving, M. (2016). Mediating pathways between parental socio-economic position and allostatic load in mid-life: Findings from the 1958 British birth cohort. *Social Science & Medicine*, *165*, 19–27. https://doi.org/10.1016/j.socscimed.2016.07.031

Barboza Solís, C., Kelly-Irving, M., Fantin, R., Darnaudéry, M., Torrisani, J., Lang, T., & Delpierre, C. (2015). Adverse childhood experiences and physiological wear-and-tear in midlife: Findings from the 1958 British birth cohort. *Proceedings of the National Academy of Sciences*, *112*(7), E738–E746. https://doi.org/10.1073/pnas.1417325112

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bartlett, J. W., & Hughes, R. A. (2020). Bootstrap inference for multiple imputation under uncongeniality and misspecification. *Statistical Methods in Medical Research*, 096228022093218. https://doi.org/10.1177/0962280220932189

Basta, N. E., & Halloran, M. E. (2019). Evaluating the Effectiveness of Vaccines Using a Regression Discontinuity Design. *American Journal of Epidemiology*, *188*(6), 987–990. https://doi.org/10.1093/aje/kwz043

Bates, T. C., Maher, B. S., Medland, S. E., McAloney, K., Wright, M. J., Hansell, N. K., Kendler, K. S., Martin, N. G., & Gillespie, N. A. (2018). The Nature of Nurture: Using a Virtual-Parent Design to Test Parenting Effects on Children's Educational Attainment in Genotyped Families. *Twin Research and Human Genetics*, *21*(2), 73–83. https://doi.org/10.1017/thg.2018.11

Baxter, A. J., Dundas, R., Popham, F., & Craig, P. (2021). How effective was England's teenage pregnancy strategy? A comparative analysis of high-income countries. *Social Science & Medicine*, *270*, 113685. https://doi.org/10.1016/j.socscimed.2021.113685

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*(2), 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, *31*(2), 285–293. https://doi.org/10.1093/intjepid/31.2.285

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435–435. https://doi.org/10.1038/d41586-019-01307-2

Blum, M. R., Tan, Y. J., & Ioannidis, J. P. A. (2020). Use of E-values for addressing confounding in observational studies—An empirical assessment of the literature. *International Journal of Epidemiology*, *49*(5), 1482–1494. https://doi.org/10.1093/ije/dyz261

Booth, C., Wong, E., Brown, M., & Fitzsimons, E. (2024). Evaluating the Effect of Monetary Incentives on Web Survey Response Rates in the UK Millennium Cohort Study. *Survey Research Methods*, *18*(1), Article 1. https://doi.org/10.18148/srm/2024.v18i1.8210

Bradford Hill, A. (1965). The Environment and Disease: Association Or Causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300.

Branwen, G. (2019, December 9). *Why Correlation Usually ≠ Causation*. Https://Gwern.Net/Correlation. https://gwern.net/causality

Branwen, G. (2022, June 14). *How Often Does Correlation = Causality?* Https://Gwern.Net/Correlation. https://gwern.net/correlation

Breum, M. S., Didelez, V., Gabriel, E. E., & Sachs, M. C. (2025). *Bounds for causal mediation effects* (arXiv:2512.11549). arXiv. https://doi.org/10.48550/arXiv.2512.11549

Brion, M.-J. A., Lawlor, D. A., Matijasevich, A., Horta, B., Anselmi, L., Araújo, C. L., Menezes, A. M. B., Victora, C. G., & Smith, G. (2011). What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts. *International Journal of Epidemiology*, *40*(3), 670–680.

Brion, M.-J. A., Leary, S. D., DaveySmith, G., & Ness, A. R. (2007). Similar Associations of Parental Prenatal Smoking Suggest Child Blood Pressure Is Not Influenced by Intrauterine Effects. *Hypertension*, *49*(6), 1422–1428. https://doi.org/10.1161/HYPERTENSIONAHA.106.085316

Bröckelmann, N., Balduzzi, S., Harms, L., Beyerbach, J., Petropoulou, M., Kubiak, C., Wolkewitz, M., Meerpohl, J. J., & Schwingshackl, L. (2022). Evaluating agreement between bodies of evidence from randomized controlled trials and cohort studies in medical research: A meta-epidemiological study. *BMC Medicine*, *20*(1), 174. https://doi.org/10.1186/s12916-022-02369-2

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., Havdahl, A., Hopper, J., Neale, M.,

Nivard, M. G., Pedersen, N. L., Reynolds, C. A., Tucker-Drob, E. M., Grotzinger, A., Howe, L., ... Davies, N. M. (2020). Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*, *11*(1), 3519. https://doi.org/10.1038/s41467-020-17117-4

Calderwood, L. (2014). Improving between-wave mailings on longitudinal surveys: A randomised experiment on the UK Millennium Cohort Study. *Survey Research Methods*, *8*(2), Article 2. https://doi.org/10.18148/srm/2014.v8i2.5391

Caniglia, E. C., & Murray, E. J. (2020). Difference-in-Difference in the Time of Cholera: A Gentle Introduction for Epidemiologists. *Current Epidemiology Reports*, *7*(4), 203–211. https://doi.org/10.1007/s40471-020-00245-2

Card, D., Chetty, R., & Weber, A. (2007). The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job? *American Economic Review*, *97*(2), 113–118. https://doi.org/10.1257/aer.97.2.113

Catalogue of Mental Health Measures. (2026). *Catalogue of Mental Health Measures*. https://www.cataloguementalhealth.ac.uk/

Centre for Longitudinal Studies. (2026). *Centre for Longitudinal Studies GitHub*. GitHub. https://github.com/CLS-Data

Cesarini, D., Lindqvist, E., Östling, R., & Wallace, B. (2016). Wealth, Health, and Child Development: Evidence from Administrative Data on Swedish Lottery Players. *The Quarterly Journal of Economics*, *131*(2), 687–738.

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., & Laibson, D. I. (2015). The Fourth Law of Behavior Genetics. *Current Directions in Psychological Science*, *24*(4), 304–312. https://doi.org/10.1177/0963721415580430

Chen, L., Davey Smith, G., Harbord, R. M., & Lewis, S. J. (2008). Alcohol Intake and Blood Pressure: A Systematic Review Implementing a Mendelian Randomization Approach. *PLOS Medicine*, *5*(3), e52. https://doi.org/10.1371/journal.pmed.0050052

Chen, M., & Lacey, R. E. (2018). Adverse childhood experiences and adult inflammation: Findings from the 1958 British birth cohort. *Brain, Behavior, and Immunity*, *69*, 582–590. https://doi.org/10.1016/j.bbi.2018.02.007

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols*, *15*(9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, *357*(4), 370–379. https://doi.org/10.1056/NEJMsa066082

Cinelli, C., Ferwerda, J., Hazlett, C., Tsao, D., Rudkin, A., & Ljubownikow, G. (2024). *sensemakr: Sensitivity Analysis Tools for Regression Models* (Version 0.1.6) [Computer software]. https://cran.r-project.org/web/packages/sensemakr/index.html

Cinelli, C., Forney, A., & Pearl, J. (2024). A Crash Course in Good and Bad Controls. *Sociological Methods & Research*, *53*(3), 1071–1104. https://doi.org/10.1177/00491241221099552

Cinelli, C., & Hazlett, C. (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cinelli, C., & Hazlett, C. (2025). An omitted variable bias framework for sensitivity analysis of instrumental variables. *Biometrika*, *112*(2), asaf004. https://doi.org/10.1093/biomet/asaf004

CLOSER. (2023). *CLOSER* (Version 8th Release). UK Data Service. https://doi.org/10.5255/UKDA-SERIES-2000111

Clouston, S. A. P., Rubin, M. S., Phelan, J. C., & Link, B. G. (2016). A Social History of Disease: Contextualizing the Rise and Fall of Social Inequalities in Cause-Specific Mortality. *Demography*, *53*(5), 1631–1656. https://doi.org/10.1007/s13524-016-0495-5

Cochran, W. G. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General)*, *128*(2), 234–266. https://doi.org/10.2307/2344179

Cohen-Cole, E., & Fletcher, J. M. (2008). Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis. *BMJ*, *337*(dec04 2), a2533–a2533. https://doi.org/10.1136/bmj.a2533

Coleman, T. (2019). *Causality in the Time of Cholera: John Snow As a Prototype for Causal Inference* (SSRN Scholarly Paper No. 3262234). Social Science Research Network. https://doi.org/10.2139/ssrn.3262234

Conley, D., & Heerwig, J. (2012). The Long-Term Effects of Military Conscription on Mortality: Estimates From the Vietnam-Era Draft Lottery. *Demography*, *49*(3), 841–855. https://doi.org/10.1007/s13524-012-0103-2

Conley, D., Laidley, T. M., Boardman, J. D., & Domingue, B. W. (2016). Changing Polygenic Penetrance on Phenotypes in the 20th Century Among Adults in the US Population. *Scientific Reports*, *6*(1), 30348. https://doi.org/10.1038/srep30348

Cook, J. R., & Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, *89*(428), 1314–1328. https://doi.org/10.2307/2290994

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *JNCI: Journal of the National Cancer Institute*, *22*(1), 173–203. https://doi.org/10.1093/jnci/22.1.173

Cornish, R., & Brennan, I. (2025). Exclusion from School and Risk of Serious Violence: A Target Trial Emulation Study. *The British Journal of Criminology*, *65*(6), 1221–1240. https://doi.org/10.1093/bjc/azaf015

Courtin, E., Kim, S., Song, S., Yu, W., & Muennig, P. (2020). Can Social Policies Improve Health? A Systematic Review and Meta-Analysis of 38 Randomized Trials. *The Milbank Quarterly*, *98*(2), 297–371. https://doi.org/10.1111/1468-0009.12451

Craig, P., Dundas, R., Leyland, A., & Popham, F. (2016). How successful was the English teenage pregnancy strategy? *The Lancet*, *388*(10060), 2604. https://doi.org/10.1016/S0140-6736(16)32223-1

Crawford, C., Dearden, L., & Greaves, Ellen. (2011). *Does when you are born matter? The impact of month of birth on children's cognitive and non-cognitive skills in England* (p. 53). Institue for Fiscal Studies. https://ifs.org.uk/sites/default/files/output_url_files/bn122.pdf

Crüwell, S., Van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science: An Annotated Reading List. *Zeitschrift Für Psychologie*, *227*(4), 237–248. https://doi.org/10.1027/2151-2604/a000387

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. https://doi.org/10.2307/j.ctv1c29t27

DAGitty. (2025). *Learn about DAGs and DAGitty*. https://dagitty.net/learn/

Dahabreh, I. J., & Bibbins-Domingo, K. (2024). Causal inference about the effects of interventions from observational studies in medical journals. *JAMA : The Journal of the American Medical Association*, *331*(21), 1845. https://doi.org/10.1001/jama.2024.7741

Dang, L. E., & Balzer, L. B. (2023). Start with the Target Trial Protocol, Then Follow the Roadmap for Causal Inference. *Epidemiology (Cambridge, Mass.)*, *34*(5), 619–623. https://doi.org/10.1097/EDE.0000000000001637

Dang, L. E., Gruber, S., Lee, H., Dahabreh, I. J., Stuart, E. A., Williamson, B. D., Wyss, R., Díaz, I., Ghosh, D., Kıcıman, E., Alemayehu, D., Hoffman, K. L., Vossen, C. Y., Huml, R. A., Ravn, H., Kvist, K., Pratley, R., Shih, M.-C., Pennello, G., ... Petersen, M. (2023). A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science*, *7*(1), e212. https://doi.org/10.1017/cts.2023.635

Daniel, R. M., & De Stavola, B. L. (2019). Mediation Analysis for Life Course Studies. In G. B. Ploubidis, B. Pongiglione, B. De Stavola, R. Daniel, L. Benova, E. Grundy, & S. Read (Eds), *Pathways to Health* (pp. 1–40). Springer Netherlands. https://doi.org/10.1007/978-94-024-1707-4_1

Daniel, R. M., De Stavola, B. L., & Cousens, S. N. (2011). Gformula: Estimating Causal Effects in the Presence of Time-Varying Confounding or Mediation using the G-Computation Formula. *The Stata Journal*, *11*(4), 479–517. https://doi.org/10.1177/1536867X1201100401

Daniel, R. M., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal Mediation Analysis with Multiple Mediators. *Biometrics*, *71*(1), 1–14. https://doi.org/10.1111/biom.12248

Davey Smith, G. (2008). Assessing Intrauterine Influences on Offspring Health Outcomes: Can Epidemiological Studies Yield Robust Findings? *Basic & Clinical Pharmacology & Toxicology*, *102*(2), 245–256. https://doi.org/10.1111/j.1742-7843.2007.00191.x

Davey Smith, G. (2011). Epidemiology, epigenetics and the 'Gloomy Prospect': Embracing randomness in population health research and practice. *International Journal of Epidemiology*, *40*(3), 537–562. https://doi.org/10.1093/ije/dyr117

Davey Smith, G., & Ebrahim, S. (2001). Epidemiology—Is it time to call it a day? *International Journal of Epidemiology*, *30*(1), 1–11. https://doi.org/10.1093/ije/30.1.1

Davey Smith, G., & Phillips, A. N. (1996). Inflation in epidemiology: 'the proof and measurement of association between two things' revisited. *British Medical Journal*, *312*(7047), 1659–1661. https://doi.org/10.1136/bmj.312.7047.1659

Davey Smith, G., Phillips, A. N., & Neaton, J. D. (1992). Smoking as 'independent' risk factor for suicide: Illustration of an artifact from observational epidemiology? *The Lancet*, *340*(8821), 709–712. https://doi.org/10.1016/0140-6736(92)92242-8

Davey Smith, G., Sterne, J. A., Fraser, A., Tynelius, P., Lawlor, D. A., & Rasmussen, F. (2009). The association between BMI and mortality using offspring BMI as an

indicator of own BMI: Large intergenerational mortality study. *BMJ*, *339*, b5043. https://doi.org/10.1136/bmj.b5043

Davies, N. M., Dickson, M., Davey Smith, G., Windmeijer, F., & van den Berg, G. J. (2023). The causal effects of education on adult health, mortality and income: Evidence from Mendelian randomization and the raising of the school leaving age. *International Journal of Epidemiology*, *52*(6), 1878–1886. https://doi.org/10.1093/ije/dyad104

Davies, N. M., Hill, W. D., Anderson, E. L., Sanderson, E., Deary, I. J., & Davey Smith, G. (2019). Multivariable two-sample Mendelian randomization estimates of the effects of intelligence and education on health. *eLife*, *8*, e43990. https://doi.org/10.7554/eLife.43990

de Roo, M., Hartman, C., Wiertsema, M., & Kretschmer, T. (2024). Gene-environment interplay explaining individual variation in weight-related outcomes: A systematic review and meta-analysis of studies using polygenic indices. OSF. https://doi.org/10.31234/osf.io/sfykx_v1

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., & Micali, N. (2015). Mediation Analysis With Intermediate Confounding: Structural Equation Modeling Viewed Through the Causal Inference Lens. *American Journal of Epidemiology*, *181*(1), 64–80. https://doi.org/10.1093/aje/kwu239

Dearden, L., Machin, S., & Reed, H. (1997). Intergenerational Mobility in Britain. *The Economic Journal*, *107*(440), 47–66.

Deary, I. J., Hill, W. D., & Gale, C. R. (2021). Intelligence, health and death. *Nature Human Behaviour*, *5*(4), 416–430. https://doi.org/10.1038/s41562-021-01078-9

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, Randomized Controlled Trials and Evidence-Based Policy: A Multidisciplinary Dialogue*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Del Bono, E., & Galindo-Rueda, F. (2007). *The Long Term Impacts of Compulsory Schooling: Evidence from a Natural Experiment in School Leaving Dates* (No. 0074;

CEE Discussion Papers). Centre for the Economics of Education, LSE. https://ideas.repec.org//p/cep/ceedps/0074.html

Delpierre, C., Barbosa-Solis, C., Torrisani, J., Darnaudery, M., Bartley, M., Blane, D., Kelly-Irving, M., Getz, L., Tomasdottir, M. O., Roberston, T., & Gustafsson, P. E. (2016). Origins of heath inequalities: The case for Allostatic Load. *Longitudinal and Life Course Studies*, *7*(1). https://doi.org/10.14301/llcs.v7i1.325

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *The Review of Economics and Statistics*, *95*(3), 932–945. https://doi.org/10.1162/REST_a_00318

Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S., & Hernán, M. A. (2019). Avoidable flaws in observational analyses: An application to statins and cancer. *Nature Medicine*, *25*(10), 1601–1606. https://doi.org/10.1038/s41591-019-0597-x

Didelez, V., Dawid, A. P., & Geneletti, S. (2006). Direct and indirect effects of sequential treatments. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI'06*, 138–146.

Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, *142*, 264–267. https://doi.org/10.1016/j.jclinepi.2021.08.001

Ding, P., & Li, F. (2019). A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. *Political Analysis*, *27*(4), 605–615. https://doi.org/10.1017/pan.2019.25

Ding, P., VanderWeele, T. J., & Robins, J. M. (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, *104*(2), 291–302. https://doi.org/10.1093/biomet/asx009

Dinsa, G. D., Goryakin, Y., Fumagalli, E., & Suhrcke, M. (2012). Obesity and socioeconomic status in developing countries: A systematic review. *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity*, *13*(11), 1067–1079. https://doi.org/10.1111/j.1467-789X.2012.01017.x

Doll, R., & Hill, A. B. (1954). The Mortality of Doctors in Relation to Their Smoking Habits. *British Medical Journal*, *1*(4877), 1451–1455. https://doi.org/10.1136/bmj.1.4877.1451

Donohue, J. J., III, & Levitt, S. D. (2001). The Impact of Legalized Abortion on Crime*. *The Quarterly Journal of Economics*, *116*(2), 379–420. https://doi.org/10.1162/00335530151144050

Donovan, S. J., & Susser, E. (2011). Commentary: Advent of sibling designs. *International Journal of Epidemiology*, *40*(2), 345–349. https://doi.org/10.1093/ije/dyr057

Dräger, J., Klein, M., & Sosu, E. M. (2025). Does the impact of pupil absences on achievement depend on their timing? *American Educational Research Journal*, *62*(5), 872–908. https://doi.org/10.3102/00028312251347666

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias—An Updated Review. *PLOS ONE*, *8*(7), e66844. https://doi.org/10.1371/journal.pone.0066844

Eggers, A. C., & Hainmueller, J. (2009). MPs for Sale? Returns to Office in Postwar British Politics. *American Political Science Review*, *103*(4), 513–533. https://doi.org/10.1017/S0003055409990190

Eggers, A. C., Tuñón, G., & Dafoe, A. (2024). Placebo Tests for Causal Inference. *American Journal of Political Science*, *68*(3), 1106–1121. https://doi.org/10.1111/ajps.12818

Erikson, R., & Goldthorpe, J. H. (2010). Has social mobility in Britain decreased? Reconciling divergent findings on income and class mobility. *The British Journal of Sociology*, *61*(2), 211–230. https://doi.org/10.1111/j.1468-4446.2010.01310.x

Eriksson, S., & Rooth, D. O. (2014). Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment. *American Economic Review*, *104*(3), 1014–1039. https://doi.org/10.1257/aer.104.3.1014

Ermisch, J., & Pevalin, D. J. (2005). Early motherhood and later partnerships. *Journal of Population Economics*, *18*(3), 469–489. https://doi.org/10.1007/s00148-004-0216-z

Evans, L. (1986). The effectiveness of safety belts in preventing fatalities. *Accident Analysis & Prevention*, *18*(3), 229–241. https://doi.org/10.1016/0001-4575(86)90007-2

Ferri, E., Bynner, J., & Wadsworth, M. (Eds). (2003). *Changing Britain, changing lives: Three generations at the turn of the century*. Inst. of Education, Univ. of London.

Feynman, R. (1974). *Cargo Cult Science*. Commencement Address, Caltech. https://calteches.library.caltech.edu/51/2/CargoCult.htm

Fitzsimons, E., & Pongiglione, B. (2019). The impact of maternal employment on children's weight: Evidence from the UK. *SSM - Population Health*, *7*, 100333. https://doi.org/10.1016/j.ssmph.2018.100333

Fitzsimons, E., & Vera-Hernández, M. (2022). Breastfeeding and Child Development. *American Economic Journal: Applied Economics*, *14*(3), 329–366. https://doi.org/10.1257/app.20180385

Flanders, W. D., Klein, M., Darrow, L. A., Strickland, M. J., Sarnat, S. E., Sarnat, J. A., Waller, L. A., Winquist, A., & Tolbert, P. E. (2011). A Method for Detection of Residual Confounding in Time-Series and Other Observational Studies. *Epidemiology (Cambridge, Mass.)*, *22*(1), 59–67. https://doi.org/10.1097/EDE.0b013e3181fdcabe

Fox, M. P., MacLehose, R. F., & Lash, T. L. (2021). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-82673-4

Fumagalli, E., & Fumagalli, L. (2019). Neighbourhood Ethnic Composition and Social Participation of Young People in England. *The Economic Journal*, *129*(622), 2459–2521. https://doi.org/10.1093/ej/uez003

Gabadinho, A., & Studer, M. (2010). *Sequence Analysis for Social Scientists*. http://www.bristol.ac.uk/cmm/software/support/workshops/materials/sequence.html

Gage, S. H., Munafò, M. R., & Davey Smith, G. (2016). Causal Inference in Developmental Origins of Health and Disease (DOHaD) Research. *Annual Review of Psychology*, *67*(1), 567–585. https://doi.org/10.1146/annurev-psych-122414-033352

Gallen, T., & Raymond, B. (2020). *Broken Instruments* (No. 3671850). Social Science Research Network. https://doi.org/10.2139/ssrn.3671850

Gallop, M., & Weschle, S. (2019). Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach. *Political Science Research and Methods*, *7*(2), 367–384. https://doi.org/10.1017/psrm.2016.53

Garicano, L., Lelarge, C., & Van Reenen, J. (2016). Firm Size Distortions and the Productivity Distribution: Evidence from France. *American Economic Review*, *106*(11), 3439–3479. https://doi.org/10.1257/aer.20130232

Gelman, A., & Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456. https://doi.org/10.1080/07350015.2017.1366909

Gerring, J. (2010). Causal Mechanisms: Yes, But.... *Comparative Political Studies*, *43*(11), 1499–1526. https://doi.org/10.1177/0010414010376911

Gervais, J., Lefebvre, G., & Moodie, E. E. M. (2025). Natural Effects in the Presence of an Intermediate Confounder: Evaluation of Pragmatic Estimation Strategies With an Emphasis on the Relationship Between Natural and Interventional Effects. *Statistics in Medicine*, *44*(15–17), e70038. https://doi.org/10.1002/sim.70038

Gilsanz, P., Young, J. G., Glymour, M. M., Tchetgen Tchetgen, E. J., Eng, C. W., Koenen, K. C., & Kubzansky, L. D. (2022). Marginal Structural Models for Life-Course Theories and Social Epidemiology: Definitions, Sources of Bias, and Simulated Illustrations. *American Journal of Epidemiology*, *191*(2), 349–359. https://doi.org/10.1093/aje/kwab253

Girma, S., & Paton, D. (2011). The impact of emergency birth control on teen pregnancy and STIs. *Journal of Health Economics*, *30*(2), 373–380. https://doi.org/10.1016/j.jhealeco.2010.12.004

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental Versus Experimental Estimates of Earnings Impacts. *The ANNALS of the American Academy of Political and Social Science*, *589*(1), 63–93. https://doi.org/10.1177/0002716203254879

Gneezy, U., & Rustichini, A. (2000). A Fine is a Price. *The Journal of Legal Studies*, *29*(1), 1–17. https://doi.org/10.1086/468061

Goetghebeur, E., Cessie, S. le, Stavola, B. D., Moodie, E. E., & Waernbaum, I. (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, *39*(30), 4922–4948. https://doi.org/10.1002/sim.8741

Golder, S., Loke, Y. K., & Bland, M. (2011). Meta-analyses of Adverse Effects Data Derived from Randomised Controlled Trials as Compared to Observational Studies: Methodological Overview. *PLOS Medicine*, *8*(5), e1001026. https://doi.org/10.1371/journal.pmed.1001026

Gondek, D., Bann, D., Brown, M., Hamer, M., Sullivan, A., & Ploubidis, G. B. (2021). Prevalence and early-life determinants of mid-life multimorbidity: Evidence from the 1970 British birth cohort. *BMC Public Health*, *21*(1), 1319. https://doi.org/10.1186/s12889-021-11291-w

Gondek, D., Bann, D., Patalay, P., Goodman, A., McElroy, E., Richards, M., & Ploubidis, G. B. (2021). Psychological distress from early adulthood to early old age: Evidence from the 1946, 1958 and 1970 British birth cohorts. *Psychological Medicine*, 1–10.

Goodman, A., Brown, M., Silverwood, R. J., Sakshaug, J. W., Calderwood, L., Williams, J., & Ploubidis, G. B. (2022). The Impact of Using the Web in a Mixed-Mode Follow-up of a Longitudinal Birth Cohort Study: Evidence from the National Child Development Study. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 822–850. https://doi.org/10.1111/rssa.12786

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, Themed Issue: Treatment Effect 1*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive" fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology*, *86*(1), 174.

Greenland, S. (2020). Commentary: An argument against E-values for assessing the plausibility that an association could be explained away by residual confounding. *International Journal of Epidemiology*, *49*(5), 1501–1503. https://doi.org/10.1093/ije/dyaa095

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Gutierrez, S., Glymour, M. M., & Davey Smith, G. (2025). Evidence triangulation in health research. *European Journal of Epidemiology*, *40*(7), 743–757. https://doi.org/10.1007/s10654-024-01194-6

Haack, S. (2009). Evidence and inquiry: A pragmatist reconstruction of epistemology (2nd, expanded ed edns). Prometheus Books.

Haber, N. A., Wieten, S. E., Rohrer, J. M., Arah, O. A., Tennant, P. W. G., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., Howcutt, S. J., Simmons, A. E., Leyrat, C., Schoenegger, P., Booman, A., Dufour, M.-S. K., O'Donoghue, A. L., Baglini, R., Do, S., … Fox, M. P. (2022). Causal and Associational Language in Observational Health Research: A Systematic Evaluation. *American Journal of Epidemiology*, *191*(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

Hadley, A. (2017). Teenage Pregnancy and Young Parenthood: Effective Policy and Practice (1st edn). Routledge.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Hamer, D., & Sirota, L. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, *5*(1), 11–13. https://doi.org/10.1038/sj.mp.4000662

Hamer, M., & Batty, G. D. (2020). Markers of Early Life Infection in Relation to Adult Diabetes: Prospective Evidence From a National Birth Cohort Study Over Four Decades. *Diabetes Care*, *43*(5), e61–e62. https://doi.org/10.2337/dc19-2471

Harmony. (2023). *Harmony*. https://harmonydata.ac.uk/

Hasegawa, R. B., Webster, D. W., & Small, D. S. (2019). Evaluating Missouri's Handgun Purchaser Law: A Bracketing Method for Addressing Concerns About History Interacting with Group. *Epidemiology*, *30*(3), 371–379. https://doi.org/10.1097/EDE.0000000000000989

Hawkins, S. S., Cole, T. J., & Law, C. (2011). Examining smoking behaviours among parents from the UK Millennium Cohort Study after the smoke-free legislation in Scotland. *Tobacco Control*, *20*(2), 112–118. https://doi.org/10.1136/tc.2010.037028

Herd, P., Freese, J., Sicinski, K., Domingue, B. W., Mullan Harris, K., Wei, C., & Hauser, R. M. (2019). Genes, Gender Inequality, and Educational Attainment. *American Sociological Review*, *84*(6), 1069–1098. https://doi.org/10.1177/0003122419886550

Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, *26*(10), 674–680. https://doi.org/10.1016/j.annepidem.2016.08.016

Hernán, M. A. (2018a). *HarvardX: Causal Diagrams: Draw Your Assumptions Before Your Conclusions*. edX. https://www.edx.org/learn/data-analysis/harvard-university-causal-diagrams-draw-your-assumptions-before-your-conclusions

Hernán, M. A. (2018b). The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American Journal of Public Health*, *108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary

heart disease. *Epidemiology (Cambridge, Mass.)*, *19*(6), 766–779. https://doi.org/10.1097/EDE.0b013e3181875e61

Hernán, M. A., Dahabreh, I. J., Dickerman, B. A., & Swanson, S. A. (2025). The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful? *Annals of Internal Medicine*, *178*(3), 402–407. https://doi.org/10.7326/ANNALS-24-01871

Hernán, M. A., Hsu, J., & Healy, B. (2019). A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*, *32*(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578

Hernán, M. A., & Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *American Journal of Epidemiology*, *183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hernán, M. A., & Robins, J. M. (2023). *Causal inference: What if* (First edition). Taylor and Francis.

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, *79*, 70–75. https://doi.org/10.1016/j.jclinepi.2016.04.014

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Hole, A. R., & Ratcliffe, A. (2020). The Impact of the London Bombings on the Well-Being of Adolescent Muslims. *The Scandinavian Journal of Economics*, *122*(4), 1606–1639. https://doi.org/10.1111/sjoe.12379

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hope, S., Pearce, A., Chittleborough, C., Deighton, J., Maika, A., Micali, N., Mittinty, M., Law, C., & Lynch, J. (2019). Temporal effects of maternal psychological distress on child mental health problems at ages 3, 5, 7 and 11: Analysis from the UK Millennium Cohort Study. *Psychological Medicine*, *49*(4), 664–674. https://doi.org/10.1017/S0033291718001368

Hope, S., Pearce, A., Cortina-Borja, M., Chittleborough, C., Barlow, J., & Law, C. (2021). Modelling the potential for parenting skills interventions to reduce inequalities and population prevalence of children's mental health problems: Evidence from the Millennium Cohort Study. *SSM - Population Health*, *14*, 100817. https://doi.org/10.1016/j.ssmph.2021.100817

Huffman, C., & van Gameren, E. (2018). Covariate Balancing Inverse Probability Weights for Time-Varying Continuous Interventions. *Journal of Causal Inference*, *6*(2), 20170002. https://doi.org/10.1515/jci-2017-0002

Hughes, A. M., Torvik, F. A., van Bergen, E., Hannigan, L. J., Corfield, E. C., Andreassen, O. A., Ystrom, E., Ask, H., Davey Smith, G., Davies, N. M., & Havdahl, A. (2024). Parental education and children's depression, anxiety, and ADHD traits, a within-family study in MoBa. *Npj Science of Learning*, *9*(1), 46. https://doi.org/10.1038/s41539-024-00260-8

Huntington-Klein, N. (2022). *The Effect: An Introduction To Research Design And Causality* (1st edn). CRC Press, Taylor & Francis Group.

Huntington-Klein, N. (2026). The Effect: An Introduction To Research Design And Causality (2nd edn). CRC Press.

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *BMJ*, *340*, c2289. https://doi.org/10.1136/bmj.c2289

Hyppönen, E., Carslake, D., Berry, D. J., Power, C., & Davey Smith, G. (2022). Estimating the influence of body mass index (BMI) on mortality using offspring BMI as an instrumental variable. *International Journal of Obesity*, *46*(1), 77–84. https://doi.org/10.1038/s41366-021-00962-8

Ioannidis, J. P. A. (2001). Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies. *JAMA*, *286*(7), 821. https://doi.org/10.1001/jama.286.7.821

Ioannidis, J. P. A., Tan, Y. J., & Blum, M. R. (2019). Limitations and Misinterpretations of E-Values for Sensitivity Analyses of Observational Studies. *Annals of Internal Medicine*, *170*(2), 108–111. https://doi.org/10.7326/M18-2159

Jackson, J. W., & Swanson, S. A. (2015). Toward a Clearer Portrayal of Confounding Bias in Instrumental Variable Applications: *Epidemiology*, *26*(4), 498–504. https://doi.org/10.1097/EDE.0000000000000287

Jackson, L. A., Jackson, M. L., Nelson, J. C., Neuzil, K. M., & Weiss, N. S. (2006). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology*, *35*(2), 337–344. https://doi.org/10.1093/ije/dyi274

James, J. (2024). Smoking, information, and education: The Royal College of Physicians and the new public health movement. *Journal of Policy Analysis and Management*, *43*(2), 446–471. https://doi.org/10.1002/pam.22508

Janiaud, P., Agarwal, A., Tzoulaki, I., Theodoratou, E., Tsilidis, K. K., Evangelou, E., & Ioannidis, J. P. A. (2021). Validity of observational evidence on putative risk and protective factors: Appraisal of 3744 meta-analyses on 57 topics. *BMC Medicine*, *19*(1), 157. https://doi.org/10.1186/s12916-021-02020-6

Jerrim, J., & Sims, S. (2019). Why do so few low- and middle-income children attend a grammar school? New evidence from the Millennium Cohort Study. *British Educational Research Journal*, *45*(3), 425–457. https://doi.org/10.1002/berj.3502

Jerrim, J., & Sims, S. (2020). The Association Between Attending a Grammar School and Children's Socio-Emotional Outcomes. New Evidence from the Millennium Cohort Study. *British Journal of Educational Studies*, *68*(1), 25–42. https://doi.org/10.1080/00071005.2018.1518513

Johnson, W., Li, L., Kuh, D., & Hardy, R. (2015). How Has the Age-Related Process of Overweight or Obesity Development Changed over Time? Co-ordinated Analyses of

Individual Participant Data from Five United Kingdom Birth Cohorts. *PLOS Medicine*, *12*(5), e1001828. https://doi.org/10.1371/journal.pmed.1001828

Joshi, H., & Fitzsimons, E. (2016). The Millennium Cohort Study: The making of a multi-purpose resource for social science and policy. *Longitudinal and Life Course Studies*, *7*(4), 409–430. https://doi.org/10.14301/llcs.v7i4.410

Karim, M. E., Platt, R. W., & Group, T. B. study. (2017). Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural Cox model context. *Statistics in Medicine*, *36*(13), 2032–2047. https://doi.org/10.1002/sim.7266

Katsoulis, M., Narayanan, M., Dodgeon, B., Ploubidis, G., & Silverwood, R. (2024). *A data driven approach to address missing data in the 1970 British birth cohort* (p. 2024.02.01.24302101). medRxiv. https://doi.org/10.1101/2024.02.01.24302101

Kelly, Y., Xue, B., Booker, C., Sacker, A., Lacey, R., Ploubidis, G., & Patalay, P. (2022). What (if anything) is going on? Examining longitudinal associations between social media use and mental ill-health among young people. https://doi.org/10.1101/2022.03.31.22273198

Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V., & Freedman, L. S. (2020). STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment. *Statistics in Medicine*, *39*(16), 2197–2231. https://doi.org/10.1002/sim.8532

Keogh, R. H., & White, I. R. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in Medicine*, *33*(12), 2137–2155. https://doi.org/10.1002/sim.6095

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., Benonisdottir, S., Oddsson, A., Halldorsson, B. V., Masson, G., Gudbjartsson, D.

F., Helgason, A., Bjornsdottir, G., Thorsteinsdottir, U., & Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. *Science*, *359*(6374), 424–428. https://doi.org/10.1126/science.aan6877

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics*, *62*(1), 85–96. https://doi.org/10.1111/j.1541-0420.2005.00396.x

Kuh, D., Pierce, M., Adams, J., Deanfield, J., Ekelund, U., Friberg, P., Ghosh, A. K., Harwood, N., Hughes, A., Macfarlane, P. W., Mishra, G., Pellerin, D., Wong, A., Stephen, A. M., Richards, M., Hardy, R., & on behalf of the NSHD scientific and data collection team. (2011). Cohort Profile: Updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research. *International Journal of Epidemiology*, *40*(1), e1–e9. https://doi.org/10.1093/ije/dyq231

Kunz, R., & Oxman, A. D. (1998). The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*, *317*(7167), 1185–1190. https://doi.org/10.1136/bmj.317.7167.1185

Labrecque, J. A. (2023, August 21). *So you want to estimate an association…* Jeremy Labrecque. https://www.jeremylabrecque.org/post/so_association/

Labrecque, J. A., & Kezios, K. L. (2026). How and when to use causal and associational language. *BMJ*, *392*, e085749. https://doi.org/10.1136/bmj-2025-085749

Lash, T. L., Fox, M. P., MacLehose, R. F., Maldonado, G., McCandless, L. C., & Greenland, S. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology*, *43*(6), 1969–1985. https://doi.org/10.1093/ije/dyu149

Lawlor, D. A., Davey Smith, G., Bruckdorfer, K. R., Kundu, D., & Ebrahim, S. (2004). Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *The Lancet*, *363*(9422), 1724–1727. https://doi.org/10.1016/S0140-6736(04)16260-0

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, *45*(6), 1866–1886. https://doi.org/10.1093/ije/dyw314

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., & Carpenter, J. R. (2018). Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. *PharmacoEconomics*, *36*(8), 889–901. https://doi.org/10.1007/s40273-018-0650-5

Li, L., & Avendano, M. (2023). Lone parents' employment policy and adolescents' socioemotional development: Quasi-experimental evidence from a UK reform. *Social Science & Medicine*, *320*, 115754. https://doi.org/10.1016/j.socscimed.2023.115754

Libuy, N., Church, D., Ploubidis, G., & Fitzsimons, E. (2024). Fast food proximity and weight gain in childhood and adolescence: Evidence from Great Britain. *Health Economics*, *33*(3), 449–465. https://doi.org/10.1002/hec.4770

Link, B. G., & Phelan, J. (1995). Social Conditions As Fundamental Causes of Disease. *Journal of Health and Social Behavior*, 80–94. https://doi.org/10.2307/2626958

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*, *21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Liu, W., Brookhart, M. A., Schneeweiss, S., Mi, X., & Setoguchi, S. (2012). Implications of M Bias in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, *176*(10), 938–948. https://doi.org/10.1093/aje/kws165

Liu, Y., Zhou, K., & Sims, K. D. (2025). Tipping Point Analysis: Assessing the Potential Impact of Missing Data. *JAMA*, *334*(3), 265–266. https://doi.org/10.1001/jama.2025.7420

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

Lu, H., Cole, S. R., Howe, C. J., & Westreich, D. (2022). Toward a Clearer Definition of Selection Bias When Estimating Causal Effects. *Epidemiology*, *33*(5), 699–706. https://doi.org/10.1097/EDE.0000000000001516

Lu, H., Cole, S. R., Platt, R. W., & Schisterman, E. F. (2021). Revisiting Overadjustment Bias. *Epidemiology*, *32*(5), e22–e23. https://doi.org/10.1097/EDE.0000000000001377

Lu, H., Gonsalves, G. S., & Westreich, D. (2024). Selection Bias Requires Selection: The Case of Collider Stratification Bias. *American Journal of Epidemiology*, *193*(3), 407–409. https://doi.org/10.1093/aje/kwad213

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

Lutfey, K., & Freese, J. (2005). Toward Some Fundamentals of Fundamental Causality: Socioeconomic Status and Health in the Routine Clinic Visit for Diabetes. *American Journal of Sociology*, *110*(5), 1326–1372. https://doi.org/10.1086/428914

Mackenbach, J. P. (2019). Health inequalities: Persistence and change in modern welfare states. Oxford University Press, USA.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Routledge. https://doi.org/10.4324/9780203809556

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, *58*(Volume 58, 2007), 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

Madia, J. E., Obsuth, I., Thompson, I., Daniels, H., & Murray, A. L. (2022). Long-term labour market and economic consequences of school exclusions in England: Evidence from two counterfactual approaches. *British Journal of Educational Psychology*, *92*(3), 801–816. https://doi.org/10.1111/bjep.12487

Marmot, M. (2015). Status syndrome: How your social standing directly affects your health (1st edn). Bloomsbury.

Marmot, M. (2016). The health gap: The challenge of an unequal world (1st edn). Bloomsbury.

Mathur, M. B., Smith, L. H., Ding, P., & VanderWeele, T. J. (2025). *EValue: Sensitivity Analyses for Unmeasured Confounding and Other Biases in Observational Studies and Meta-Analyses* (Version 4.1.4) [Computer software]. https://cran.r-project.org/web/packages/EValue/index.html

Matilla-Santander, N., Matthews, A. A., Gunn, V., Muntaner, C., Kreshpaj, B., Wegman, D. H., Sánchez-Martínez, N., Hernando-Rodriguez, J. C., Albin, M., Balogh, R., Davis, L., & Bodin, T. (2023). Causal effect of shifting from precarious to standard employment on all-cause mortality in Sweden: An emulation of a target trial. *J Epidemiol Community Health*, *77*(11), 736–743. https://doi.org/10.1136/jech-2023-220734

McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan (2nd edn). Taylor and Francis, CRC Press.

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Mellon, J. (2025). Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable. *American Journal of Political Science*, *69*(3), 881–898. https://doi.org/10.1111/ajps.12894

Metherell, T. E., Ploubidis, G. B., & Moreno-Agostino, D. (2023). Associations between early-life mental health and abnormal sleep duration in midlife: Findings from a prospective cohort study in Great Britain (p. 2023.12.13.23299897). medRxiv. https://doi.org/10.1101/2023.12.13.23299897

Miles, C. H. (2023). On the causal interpretation of randomised interventional indirect effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(4), 1154–1172. https://doi.org/10.1093/jrsssb/qkad066

Moreno-Betancur, M., & Carlin, J. B. (2018). Understanding Interventional Effects: A More Natural Approach to Mediation Analysis? *Epidemiology*, *29*(5), 614. https://doi.org/10.1097/EDE.0000000000000866

Morris, T., Wright, L., Shireby, G., & Bann, D. (2026). Genetic associations with education have increased and are patterned by socioeconomic context: Evidence from 3 studies born 1946-1970. *Proceedings of the National Academy of Sciences*. https://discovery.ucl.ac.uk/id/eprint/10219068/

Mostafa, T., Narayanan, M., Pongiglione, B., Dodgeon, B., Goodman, A., Silverwood, R. J., & Ploubidis, G. B. (2021). Missing at random assumption made more plausible: Evidence from the 1958 British birth cohort. *Journal of Clinical Epidemiology*, *136*, 44–54. https://doi.org/10.1016/j.jclinepi.2021.02.019

Moulton, V., Sullivan, A., Patalay, P., Fitzsimons, E., Henderson, M., Bann, D., & Ploubidis, G. B. (2023). Association between psychological distress trajectories from adolescence to midlife and mental health during the pandemic: Evidence from two British birth cohorts. *Psychological Medicine*, *53*(14), 6547–6559. https://doi.org/10.1017/S0033291722003877

Mulder, J. D., & Hamaker, E. L. (2021). Three Extensions of the Random Intercept Cross-Lagged Panel Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 638–648. https://doi.org/10.1080/10705511.2020.1784738

Mumford, S., & Anjum, R. L. (2013). *Causation: A very short introduction* (1. ed). Oxford Univ. Press. https://doi.org/10.1093/actrade/9780199684434.001.0001

Muñoz, J., Falcó-Gimeno, A., & Hernández, E. (2020). Unexpected Event during Survey Design: Promise and Pitfalls for Causal Inference. *Political Analysis*, *28*(2), 186–206. https://doi.org/10.1017/pan.2019.27

Newman, S. J. (2024). *The global pattern of centenarians highlights deep problems in demography* (p. 2024.09.06.24313170). medRxiv. https://doi.org/10.1101/2024.09.06.24313170

Newman, S. J. (2025). Supercentenarian and remarkable age records exhibit patterns indicative of clerical errors and pension fraud (p. 704080). bioRxiv. https://doi.org/10.1101/704080

Nguyen, C. D., Lee, K. J., White, I. R., Buuren, S. van, & Moreno-Betancur, M. (2025). *Sensitivity analysis for multivariable missing data using multiple imputation: A tutorial* (arXiv:2502.02892). arXiv. https://doi.org/10.48550/arXiv.2502.02892

Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*, *26*(2), 255–271. https://doi.org/10.1037/met0000299

Ning, K., Gondek, D., Pereira, S. M. P., & Lacey, R. E. (2024). Mediating mechanisms of the relationship between exposure to deprivation and threat during childhood and adolescent psychopathology: Evidence from the Millennium Cohort Study. *European Child & Adolescent Psychiatry*, *33*(6), 1907–1920. https://doi.org/10.1007/s00787-023-02289-3

Ning, K., Patalay, P., L. Maggs, J., & Ploubidis, G. B. (2021). Early life mental health and problematic drinking in mid-adulthood: Evidence from two British birth cohorts. *Social Psychiatry and Psychiatric Epidemiology*, *56*(10), 1847–1858. https://doi.org/10.1007/s00127-021-02063-3

Orben, A., & Przybylski, A. K. (2019). Screens, Teens, and Psychological Well-Being: Evidence From Three Time-Use-Diary Studies. *Psychological Science*, *30*(5), 682–696. https://doi.org/10.1177/0956797619830329

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204. https://doi.org/10.1080/07350015.2016.1227711

Papanikolaou, P. N., Christidi, G. D., & Ioannidis, J. P. A. (2006). Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ : Canadian Medical Association Journal*, *174*(5), 635–641. https://doi.org/10.1503/cmaj.050873

Parra, C. O., Daniel, R. M., & Bartlett, J. W. (2021). *Hypothetical estimands in clinical trials: A unification of causal inference and missing data methods* (arXiv:2107.04392). arXiv. https://doi.org/10.48550/arXiv.2107.04392

Pastore, C., & Jones, A. M. (2023). Human capital consequences of missing out on a grammar school education. *Economic Modelling*, *126*, 106414. https://doi.org/10.1016/j.econmod.2023.106414

Patacchini, E., & Zenou, Y. (2011). Neighborhood Effects and Parental Involvement in the Intergenerational Transmission of Education. *Journal of Regional Science*, *51*(5), 987–1013. https://doi.org/10.1111/j.1467-9787.2011.00722.x

Patel, A., & Sissons, R. (2024, November 27). *'The IVF postcode lottery is not fair'*. BBC. BBC News. https://www.bbc.co.uk/news/articles/ckglq7kn11jo

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI'01*, 411–420. https://dl.acm.org/doi/10.5555/2074022.2074073

Pearl, J. (2022). *Causality: Models, reasoning, and inference* (Second edition, reprinted with corrections). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (First edition). Basic Books.

Pearson, H. W. (2016). *The life project: The extraordinary story of our ordinary lives*. Allen Lane, an imprint of Penguin Books.

Pereira, R. D., Biroli, P., Galama, T., Hinke, S. von, Kippersluis, H. van, Rietveld, C. A., & Thom, K. (2022). *Gene-Environment Interplay in the Social Sciences*. https://doi.org/10.1093/acrefore/9780190625979.013.804

Persson, M. (2014). Testing the Relationship Between Education and Political Participation Using the 1970 British Cohort Study. *Political Behavior*, *36*(4), 877–897. https://doi.org/10.1007/s11109-013-9254-0

Petersen, M. L., Sinisi, S. E., & van der Laan, M. J. (2006). Estimation of Direct Causal Effects. *Epidemiology*, *17*(3), 276. https://doi.org/10.1097/01.ede.0000208475.99429.2d

Petersen, M. L., & van der Laan, M. J. (2014). Causal Models and Learning from Data. *Epidemiology (Cambridge, Mass.)*, *25*(3), 418–426. https://doi.org/10.1097/EDE.0000000000000078

Petitti, D. B., Perlman, J. A., & Sidney, S. (1986). Postmenopausal Estrogen Use and Heart Disease. *New England Journal of Medicine*, *315*(2), 131–132. https://doi.org/10.1056/NEJM198607103150213

Phillips, A. N., & Davey Smith, G. (1991). How independent are 'independent' effects? Relative risk estimation when correlated exposures are measured imprecisely. *Journal of Clinical Epidemiology*, *44*(11), 1223–1231. https://doi.org/10.1016/0895-4356(91)90155-3

Piccininni, M., & Stensrud, M. J. (2024). Using Negative Control Populations to Assess Unmeasured Confounding and Direct Effects. *Epidemiology (Cambridge, Mass.)*, *35*(3), 313–319. https://doi.org/10.1097/EDE.0000000000001724

Pina-Sánchez, J., Brunton-Smith, I., Buil-Gil, D., & Cernat, A. (2023). Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis. *Crime Science*, *12*(1), 14. https://doi.org/10.1186/s40163-023-00192-5

Pinto Pereira, S. M., De Stavola, B. L., Rogers, N. T., Hardy, R., Cooper, R., & Power, C. (2020). Adult obesity and mid-life physical functioning in two British birth cohorts: Investigating the mediating role of physical inactivity. *International Journal of Epidemiology*, *49*(3), 845–856. https://doi.org/10.1093/ije/dyaa014

Plomin, R. (2018). Blueprint: How DNA Makes Us Who We Are. Penguin.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences*, *10*(1), 1–16. https://doi.org/10.1017/S0140525X00055941

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*(7), 702–709. https://doi.org/10.1038/ng.3285

Poole, C. (2020). Commentary: Continuing the E-value's post-publication peer review. *International Journal of Epidemiology*, *49*(5), 1497–1500. https://doi.org/10.1093/ije/dyaa097

Purba, A. K., Henderson, M., Baxter, A., Pearce, A., & Katikireddi, S. V. (2025). The Relationship Between Time Spent on Social Media and Adolescent Cigarette, E-cigarette, and Dual Use: A Longitudinal Analysis of the UK Millennium Cohort Study. *Nicotine & Tobacco Research*, *27*(4), 693–704. https://doi.org/10.1093/ntr/ntae057

Rajah, N., Calderwood, L., De Stavola, B. L., Harron, K., Ploubidis, G. B., & Silverwood, R. J. (2023). Using linked administrative data to aid the handling of non-response and restore sample representativeness in cohort studies: The 1958 national child development study and hospital episode statistics data. *BMC Medical Research Methodology*, *23*(1), 266. https://doi.org/10.1186/s12874-023-02099-w

Rambachan, A., & Roth, J. (2023). A More Credible Approach to Parallel Trends. *Review of Economic Studies*, *90*(5), 2555–2591. https://doi.org/10.1093/restud/rdad018

Reeves, A., Clair, A., McKee, M., & Stuckler, D. (2016). Reductions in the United Kingdom's Government Housing Benefit and Symptoms of Depression in Low-Income Households. *American Journal of Epidemiology*, *184*(6), 421–429. https://doi.org/10.1093/aje/kww055

Reichardt, C. S. (2019). *Quasi-Experimentation: A Guide to Design and Analysis* (T. D. Little, Ed.). Guilford Press. https://www.jstor.org/stable/10.1521/jj.41115888

Rich Harris, J. (1999). The nurture assumption: Why children turn out the way they do. Bloomsbury.

Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K., & Smith, G. D. (2020). Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study. *Bmj*, *369*.

Richmond, R. C., Al-Amin, A., Davey Smith, G., & Relton, C. L. (2014). Approaches for drawing causal inferences from epidemiological birth cohorts: A review. *Early Human Development, Special Issue: Neonatal Update 2014*, *90*(11), 769–780. https://doi.org/10.1016/j.earlhumdev.2014.08.023

Ripollone, J. E., Huybrechts, K. F., Rothman, K. J., Ferguson, R. E., & Franklin, J. M. (2018). Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology. *American Journal of Epidemiology*, *187*(9), 1951–1961. https://doi.org/10.1093/aje/kwy078

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, *29*(8), 1358–1369.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6

Robins, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. E. Freeman, & A. H. Mulley Jr. (Eds), *Health service research methodology: A focus on AIDS* (pp. 113–159). U.S. Public Health Service.

Robins, J. M., & Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, *3*(2), 143.

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology: *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Robins, J. M., Richardson, T. S., & Shpitser, I. (2022). An Interventionist Approach to Mediation Analysis. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (1st edn, Vol. 36, pp. 713–764). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3501714.3501754

Romero, C., Leeuw, C. de, Schipper, M., Maciel, B. de A. P. C., Heuvel, M. P. van den, Brouwer, R. M., Smit, A. B., Koopmans, F., Posthuma, D., & Sluis, S. van der. (2025).

Immune-Developmental Processes Contribute to Schizophrenia Risk: Insights From a Genetic Overlap Study With Height. *Biological Psychiatry*, *98*(9), 689–697. https://doi.org/10.1016/j.biopsych.2025.02.902

Ronson, J. (2001). Them: Adventures with Extremists. Picador.

Rosenbaum, P. R. (1987). The Role of a Second Control Group in an Observational Study. *Statistical Science*, *2*(3). https://doi.org/10.1214/ss/1177013232

Rosenbaum, P. R. (1989). The Role of Known Effects in Observational Studies. *Biometrics*, *45*(2), 557–569. https://doi.org/10.2307/2531497

Rosenbaum, P. R. (2001). Replicating Effects and Biases. *The American Statistician*, *55*(3), 223–227. https://doi.org/10.1198/000313001317098220

Rosenbaum, P. R. (2005). Reasons for Effects. *CHANCE*, *18*(1), 5–10. https://doi.org/10.1080/09332480.2005.10722702

Rosenbaum, P. R. (2006). Differential Effects and Generic Biases in Observational Studies. *Biometrika*, *93*(3), 573–586.

Rosenbaum, P. R. (2013). Using Differential Comparisons in Observational Studies. *CHANCE*, *26*(3), 18–25. https://doi.org/10.1080/09332480.2013.845002

Rosenbaum, P. R. (2015). Cochran's Causal Crossword. *Observational Studies*, *1*(1), 205–211. https://doi.org/10.1353/obs.2015.0021

Rosenbaum, P. R. (2019). Observation and experiment: An introduction to causal inference. Harvard University Press.

Rosenbaum, P. R. (2020). *Design of Observational Studies* (2nd ed). Springer International Publishing AG.

Rosenbaum, P. R. (2023). *Causal Inference*. The MIT Press. https://doi.org/10.7551/mitpress/14244.001.0001

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Royal College of Physicians. (1962). *Smoking and Health*. Pitman Medical Publishing Co. LTD. https://www.rcp.ac.uk/media/rujmcoyp/smoking-and-health-1962_1.pdf

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316696

Rushton, A., Maughan, B., Grant, M., & Sehmi, R. (2020). Infant domestic adoptions followed up to adulthood: Considerations with reference to British birth cohort data. *Adoption & Fostering*, *44*(4), 334–348. https://doi.org/10.1177/0308575920968237

Sabates, R., & Duckworth, K. (2010). Maternal schooling and children's relative inequalities in developmental outcomes: Evidence from the 1947 school leaving age reform in Britain. *Oxford Review of Education*, *36*(4), 445–461. https://doi.org/10.1080/03054981003775277

Sacerdote, B. (2002). The Nature and Nurture of Economic Outcomes. *American Economic Review*, *92*(2), 344–348. https://doi.org/10.1257/000282802320191589

Sanderson, E., Richardson, T. G., Hemani, G., & Davey Smith, G. (2021). The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International Journal of Epidemiology*, *50*(4), 1350–1361. https://doi.org/10.1093/ije/dyaa288

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, *20*(4), 488.

Schmitz, H. (2011). Why are the unemployed in worse health? The causal effect of unemployment on health. *Labour Economics*, *18*(1), 71–78. https://doi.org/10.1016/j.labeco.2010.08.005

Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation: Bootstrap Inference When Using Multiple Imputation. *Statistics in Medicine*, *37*(14), 2252–2266. https://doi.org/10.1002/sim.7654

Scott, R. (2022). Does university make you more liberal? Estimating the within-individual effects of higher education on political values. *Electoral Studies*, *77*, 102471. https://doi.org/10.1016/j.electstud.2022.102471

Seaman, S. R., & Vansteelandt, S. (2018). Introduction to Double Robust Methods for Incomplete Data. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *33*(2), 184–197. https://doi.org/10.1214/18-STS647

Sellers, R., Warne, N., Rice, F., Langley, K., Maughan, B., Pickles, A., Thapar, A., & Collishaw, S. (2020). Using a cross-cohort comparison design to test the role of maternal smoking in pregnancy in child mental health and learning: Evidence from two UK cohorts born four decades apart. *International Journal of Epidemiology*, *49*(2), 390–399.

Selzam, S., McAdams, T. A., Coleman, J. R. I., Carnell, S., O'Reilly, P. F., Plomin, R., & Llewellyn, C. H. (2018). Evidence for gene-environment correlation in child feeding: Links between common genetic variation for BMI in children and parental feeding practices. *PLOS Genetics*, *14*(11), e1007757. https://doi.org/10.1371/journal.pgen.1007757

Shadish, W. R., & Cook, T. D. (1999). Comment-Design Rules: More Steps toward a Complete Theory of Quasi-Experimentation. *Statistical Science*, *14*(3), 294–300.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (Nachdr.). Wadsworth Cengage Learning.

Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Keogh, R. H., Kipnis, V., Tooze, J. A., Wallace, M. P., Küchenhoff, H., & Freedman, L. S. (2020). STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics. *Statistics in Medicine*, *39*(16), 2232–2263. https://doi.org/10.1002/sim.8531

Sheikh, M. A., Abelsen, B., & Olsen, J. A. (2016). Differential Recall Bias, Intermediate Confounding, and Mediation Analysis in Life Course Epidemiology: An Analytic

Framework with Empirical Example. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01828

Shi, B., Choirat, C., Coull, B. A., VanderWeele, T. J., & Valeri, L. (2021). CMAverse: A Suite of Functions for Reproducible Causal Mediation Analyses. *Epidemiology*, *32*(5), e20. https://doi.org/10.1097/EDE.0000000000001378

Shi, X., Miao, W., & Tchetgen, E. T. (2020). A Selective Review of Negative Control Methods in Epidemiology. *Current Epidemiology Reports*, *7*(4), 190–202. https://doi.org/10.1007/s40471-020-00243-4

Shireby, G., Morris, T. T., Wong, A., Chaturvedi, N., Ploubidis, G. B., Fitzsimons, E., Goodman, A., Sanchez-Galvez, A., Davies, N. M., Wright, L., & Bann, D. (2025). Data Resource Profile: Genomic data in multiple British birth cohorts (1946–2001)—linkage with health, social, and environmental data from birth to old age. *International Journal of Epidemiology*, *54*(5), dyaf141. https://doi.org/10.1093/ije/dyaf141

Silva, L., Tropf, F. C., & Wolfram, T. (2025). Mind the gap: The interplay between genes and neighbourhood context on educational achievement. *European Sociological Review*, *41*(5), 755–773. https://doi.org/10.1093/esr/jcaf001

Silverwood, R. J., Calderwood, L., Henderson, M., Sakshaug, J. W., & Ploubidis, G. B. (2024). A data-driven approach to understanding non-response and restoring sample representativeness in the UK Next Steps cohort. *Longitudinal and Life Course Studies*, *15*(2), 227–250. https://doi.org/10.1332/17579597Y2024D000000010

Silverwood, R. J., Narayanan, M., Dodgeon, B., Katsoulis, M., & Ploubidis, G. B. (2024). *Handling missing data in the CLS cohort studies* (p. 117) [User guide]. UCL Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2020/04/Handling-Missing-Data-User-Guide-2024.pdf

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Singh-Manoux, A., Ferrie, J. E., Lynch, J. W., & Marmot, M. (2005). The Role of Cognitive Ability (Intelligence) in Explaining the Association between Socioeconomic Position and Health: Evidence from the Whitehall II Prospective Cohort Study. *American Journal of Epidemiology*, *161*(9), 831–839. https://doi.org/10.1093/aje/kwi109

Sjölander, A., Frisell, T., Kuja-Halkola, R., Öberg, S., & Zetterqvist, J. (2016). Carryover Effects in Sibling Comparison Designs. *Epidemiology*, *27*(6), 852–858. https://doi.org/10.1097/EDE.0000000000000541

Sjölander, A., Frisell, T., & Öberg, S. (2022). Sibling Comparison Studies. *Annual Review of Statistics and Its Application*, *9*(1), null. https://doi.org/10.1146/annurev-statistics-040120-024521

Sjölander, A., & Greenland, S. (2022). Are E-values too optimistic or too pessimistic? Both and neither! *International Journal of Epidemiology*, *51*(2), 355–363. https://doi.org/10.1093/ije/dyac018

Speyer, L. G., Hang, Y., Hall, H. A., & Murray, A. L. (2022). The role of harsh parenting practices in early- to middle-childhood socioemotional development: An examination in the Millennium Cohort Study. *Child Development*, *93*(5), 1304–1317. https://doi.org/10.1111/cdev.13761

Stacey, T., Tennant, P., McCowan, L., Mitchell, E., Budd, J., Li, M., Thompson, J., Martin, B., Roberts, D., & Heazell, A. (2019). Gestational diabetes and the risk of late stillbirth: A case–control study from England, UK. *BJOG: An International Journal of Obstetrics & Gynaecology*, *126*(8), 973–982. https://doi.org/10.1111/1471-0528.15659

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., & Hernán, M. A. (2022). Separable Effects for Causal Inference in the Presence of Competing Events. *Journal of the American Statistical Association*, *117*(537), 175–183. https://doi.org/10.1080/01621459.2020.1765783

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Tabor, E., Kneale, D., & Patalay, P. (2025). Sexuality and respiratory outcomes in the UK: Disparities, development and mediators in multiple longitudinal studies. *Public Health*, *247*, 105886. https://doi.org/10.1016/j.puhe.2025.105886

Tanksley, P. T., Brislin, S. J., Wertz, J., de Vlaming, R., Courchesne-Krak, N. S., Mallard, T. T., Raffington, L. L., Karlsson Linnér, R., Koellinger, P., Palmer, A. A., Sanchez-Roige, S., Waldman, I. D., Dick, D., Moffitt, T. E., Caspi, A., & Harden, K. P. (2025). Do Polygenic Indices Capture "Direct" Effects on Child Externalizing Behavior Problems? Within-Family Analyses in Two Longitudinal Birth Cohorts. *Clinical Psychological Science*, *13*(2), 316–331. https://doi.org/10.1177/21677026241260260

Tattan-Birch, H., Marsden, J., West, R., & Gage, S. H. (2021). Assessing and addressing collider bias in addiction research: The curious case of smoking and COVID-19. *Addiction (Abingdon, England)*, *116*(5), 982–984. https://doi.org/10.1111/add.15348

Tennant, P. W. G., & Murray, E. J. (2021). The Quest for Timely Insights into COVID-19 Should not Come at the Cost of Scientific Rigor. *Epidemiology*, *32*(1), e2. https://doi.org/10.1097/EDE.0000000000001258

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2021). Use of directed acyclic graphs (DAGs) to identify confounders

in applied health research: Review and recommendations. *International Journal of Epidemiology*, *50*(2), 620–632. https://doi.org/10.1093/ije/dyaa213

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

Thistlethwaite, D. L., & Campbell, D. T. (2017). Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment. *Observational Studies*, *3*(2), 119–128. https://doi.org/10.1353/obs.2017.0000

Thoemmes, F., & Ong, A. D. (2016). A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. *Emerging Adulthood*, *4*(1), 40–59. https://doi.org/10.1177/2167696815621645

Toews, I., Anglemyer, A., Nyirenda, J. L., Alsaid, D., Balduzzi, S., Grummich, K., Schwingshackl, L., & Bero, L. (2024). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: A meta-epidemiological study. *The Cochrane Database of Systematic Reviews*, *1*(1), MR000034. https://doi.org/10.1002/14651858.MR000034.pub3

Tomova, G. D., Silverwood, R. J., Tennant, P. W. G., & Wright, L. (2026). How can the use of different modes of survey data collection introduce bias? An introduction to mode effects using directed acyclic graphs (DAGs). *American Journal of Epidemiology*, kwag017. https://doi.org/10.1093/aje/kwag017

Tomova, G. D., Silverwood, R. J., & Wright, L. (2026a). *Mode effects on survey item measurement: A systematic review of the experimental evidence* (Bc7qn_v1). SocArXiv. https://osf.io/preprints/socarxiv/bc7qn_v1/

Tomova, G. D., Silverwood, R. J., & Wright, L. (2026b). When, and how much, might effect estimates be biased when analysing mixed-mode survey data? The roles of mode effects, mode selection, and mode split. SocArXiv. https://doi.org/10.31235/osf.io/6sxbq_v1

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J., & White, I. R. (2018). On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statistics in Medicine*, *37*(15), 2338–2353. https://doi.org/10.1002/sim.7643

Trochim, W. M. K. (1985). Pattern Matching, Validity, and Conceptualization in Program Evaluation. *Evaluation Review*, *9*(5), 575–604. https://doi.org/10.1177/0193841X8500900503

Trochim, W. M. K. (1989). Outcome pattern matching and program theory. *Evaluation and Program Planning, Special Issue: The Theory-Driven Perspective*, *12*(4), 355–366. https://doi.org/10.1016/0149-7189(89)90052-9

Tuffaha, H. W., Gordon, L. G., & Scuffham, P. A. (2014). Value of information analysis in healthcare: A review of principles and applications. *Journal of Medical Economics*, *17*(6), 377–383. https://doi.org/10.3111/13696998.2014.907170

Turkheimer, E. (2000). Three Laws of Behavior Genetics and What They Mean. *Current Directions in Psychological Science*, *9*(5), 160–164. https://doi.org/10.1111/1467-8721.00084

Vagni, G. (2026). Life Course Social Mobility and Parenthood. Counterfactual Estimates of the Motherhood Class Penalty in Britain. *The British Journal of Sociology*, *77*(1), 114–129. https://doi.org/10.1111/1468-4446.70039

Vagni, G., & Breen, R. (2021). Earnings and Income Penalties for Motherhood: Estimates for British Women Using the Individual Synthetic Control Method. *European Sociological Review*, *37*(5), 834–848. https://doi.org/10.1093/esr/jcab014

van Buuren, S. (2018). *Flexible imputation of missing data* (Second edition). CRC Press, Taylor & Francis Group.

van der Laan, M. J., & Gruber, S. (2012). Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions. *The International Journal of Biostatistics*, *8*(1). https://doi.org/10.1515/1557-4679.1370

VanderWeele, T. J. (2011). Sensitivity Analysis for Contagion Effects in Social Networks. *Sociological Methods & Research*, *40*(2), 240–255. https://doi.org/10.1177/0049124111404821

VanderWeele, T. J. (2017). Outcome-wide Epidemiology: *Epidemiology*, *28*(3), 399–402. https://doi.org/10.1097/EDE.0000000000000641

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

VanderWeele, T. J., & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, *167*(4), 268. https://doi.org/10.7326/M16-2607

VanderWeele, T. J., Jackson, J. W., & Li, S. (2016). Causal inference and longitudinal data: A case study of religion and mental health. *Social Psychiatry and Psychiatric Epidemiology*, *51*(11), 1457–1466. https://doi.org/10.1007/s00127-016-1281-9

VanderWeele, T. J., & Li, Y. (2019). Simple Sensitivity Analysis for Differential Measurement Error. *American Journal of Epidemiology*, *188*(10), 1823–1829. https://doi.org/10.1093/aje/kwz133

VanderWeele, T. J., Mathur, M. B., & Chen, Y. (2020). Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies. *Statistical Science*, *35*(3). https://doi.org/10.1214/19-STS728

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, *2*(4), 457–468. https://doi.org/10.4310/SII.2009.v2.n4.a7

VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder. *Epidemiology*, *25*(2), 300. https://doi.org/10.1097/EDE.0000000000000034

Vansteelandt, S., & Daniel, R. M. (2017). Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology*, *28*(2), 258. https://doi.org/10.1097/EDE.0000000000000596

Vansteelandt, S., & Sjölander, A. (2016). Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding. *Epidemiologic Methods*, *5*(1), 37–56. https://doi.org/10.1515/em-2015-0005

Vignen, T. (2025). *Spurious Correlations*. https://tylervigen.com/spurious-correlations

Wadsworth, M., Kuh, D., Richards, M., & Hardy, R. (2006). Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology*, *35*(1), 49–54. https://doi.org/10.1093/ije/dyi201

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. https://doi.org/10.1038/nature05911

Wellings, K., Palmer, M. J., Geary, R. S., Gibson, L. J., Copas, A., Datta, J., Glasier, A., Scott, R. H., Mercer, C. H., Erens, B., Macdowall, W., French, R. S., Jones, K., Johnson, A. M., Tanton, C., & Wilkinson, P. (2016). Changes in conceptions in women younger than 18 years and the circumstances of young mothers in England in 2000–12: An observational study. *The Lancet*, *388*(10044), 586–595. https://doi.org/10.1016/S0140-6736(16)30449-4

Wertz, J., Belsky, J., Moffitt, T. E., Belsky, D. W., Harrington, H., Avinun, R., Poulton, R., Ramrakha, S., & Caspi, A. (2019). Genetics of nurture: A test of the hypothesis that parents' genetics predict their observed caregiving. *Developmental Psychology*, *55*(7), 1461–1472. https://doi.org/10.1037/dev0000709

Westerman, K. E., & Sofer, T. (2024). Many roads to a gene-environment interaction. *American Journal of Human Genetics*, *111*(4), 626–635. https://doi.org/10.1016/j.ajhg.2024.03.002

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. https://doi.org/10.1177/2515245919848684

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), Article 4. https://doi.org/10.1038/s41390-018-0071-3

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (Second edition). MIT Press.

Woolf, B., Gill, D., Sallis, H., & Munafò, M. R. (2023). The UK BiLEVE and Mendelian randomisation: Using multivariable instrumental variables to address 'damned if you, damned if you don't' adjustment problems. *BMC Research Notes*, *16*(1), 157. https://doi.org/10.1186/s13104-023-06434-8

Wright, L. (2020). Arts engagement, mortality and dementia: What can the data say? *Journal of Epidemiology and Community Health*, jech-2020-214227. https://doi.org/10.1136/jech-2020-214227

Wright, L. (2021). *Investigating the Association between Youth Unemployment and Mental Health Later in Life* [University College London]. https://discovery.ucl.ac.uk/id/eprint/10128339/

Wright, L., Davies, N. M., Shireby, G., Williams, D. M., Morris, T. T., & Bann, D. (2024). Genetic Risk for High Body Mass Index Before and Amidst the Obesity Epidemic: Cross-Cohort Analysis of Four British Birth Cohort Studies (p. 2024.10.24.24315860). medRxiv. https://doi.org/10.1101/2024.10.24.24315860

Wright, L., Head, J. A., & Jivraj, S. (2021). How robust is the association between youth unemployment and later mental health? An analysis of longitudinal data from English schoolchildren. *Occupational and Environmental Medicine*, oemed-2021-107473. https://doi.org/10.1136/oemed-2021-107473

Wright, L., Ploubidis, G. B., & Silverwood, R. J. (2024). *Handling Mode Effects in CLS's cohort Studies*. UCL Centre for Longitudinal Studies.

Wright, L., Shireby, G., Morris, T. T., Davies, N. M., & Bann, D. (2025). The association between parental BMI and offspring adiposity: A genetically informed analysis of trios. *PLOS Genetics*, *21*(8), e1011775. https://doi.org/10.1371/journal.pgen.1011775

Wright, M. A., Wintemute, G. J., & Rivara, F. P. (1999). Effectiveness of denial of handgun purchase to persons believed to be at high risk for firearm violence. *American Journal of Public Health*, *89*(1), 88–90. https://doi.org/10.2105/ajph.89.1.88

Wright, S. (1934). The Method of Path Coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215. https://doi.org/10.1214/aoms/1177732676

Yang, Q., Yang, Z., Cai, X., Zhao, H., Jia, J., & Sun, F. (2024). Advances in methodologies of negative controls: A scoping review. *Journal of Clinical Epidemiology*, *166*, 111228. https://doi.org/10.1016/j.jclinepi.2023.111228

Ye, T., Keele, L., Hasegawa, R., & Small, D. S. (2024). A Negative Correlation Strategy for Bracketing in Difference-in-Differences. *Journal of the American Statistical Association*, *119*(547), 2256–2268. https://doi.org/10.1080/01621459.2023.2252576

Yudkowsky, E. (2007). Hindsight Devalues Science. *LessWrong*. https://www.lesswrong.com/posts/WnheMGAka4fL99eae/hindsight-devalues-science

Zaadstra, B., Chorus, A., Van Buuren, S., Kalsbeek, H., & Van Noort, J. (2008). Selective association of multiple sclerosis with infectious mononucleosis. *Multiple Sclerosis Journal*, *14*(3), 307–313. https://doi.org/10.1177/1352458507084265

Zivich, P. N., Ross, R. K., Shook-Sa, B. E., Cole, S. R., & Edwards, J. K. (2024). Empirical Sandwich Variance Estimator for Iterated Conditional Expectation g-Computation. *Statistics in Medicine*, *43*(29), 5562–5572. https://doi.org/10.1002/sim.10255

Zubizarreta, J. R., Small, D. S., & Rosenbaum, P. R. (2018). A Simple Example of Isolation in Building a Natural Experiment. *CHANCE*, *31*(4), 16–23. https://doi.org/10.1080/09332480.2018.1549811

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From Data to Causes I: Building A General Cross-Lagged Panel Model (GCLM). *Organizational Research Methods*, *23*(4), 651–687. https://doi.org/10.1177/1094428119847278

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2019). From Data to Causes II: Comparing Approaches to Panel Data Analysis. *Organizational Research Methods*, 109442811984728. https://doi.org/10.1177/1094428119847280