Causal Inference for Intervention & Service Evaluations: A Practical Handbook

Jamie Wong

03 Feb 2026

Table of contents

Welcome

Please note that the following book is still a work in progress, with a first draft expected to be completed by the end of Jan 2026, followed by continued revisions over the next year

This is a practical handbook on causal inference methods for evaluating policy interventions and service changes within the NHS, with a focus on leveraging observational real-world data to generate robust evidence.

The handbook was written initially as part of an NHS England PhD Internship done jointly with the NHS North Central London Integrated Care Board and University College London, and will continue to be periodically updated with revisions following feedback from academics, data scientists, data analysts, clinicians, and public health practitioners.

NHS
North Central London
Integrated Care Board

UCL

Part I

Introduction

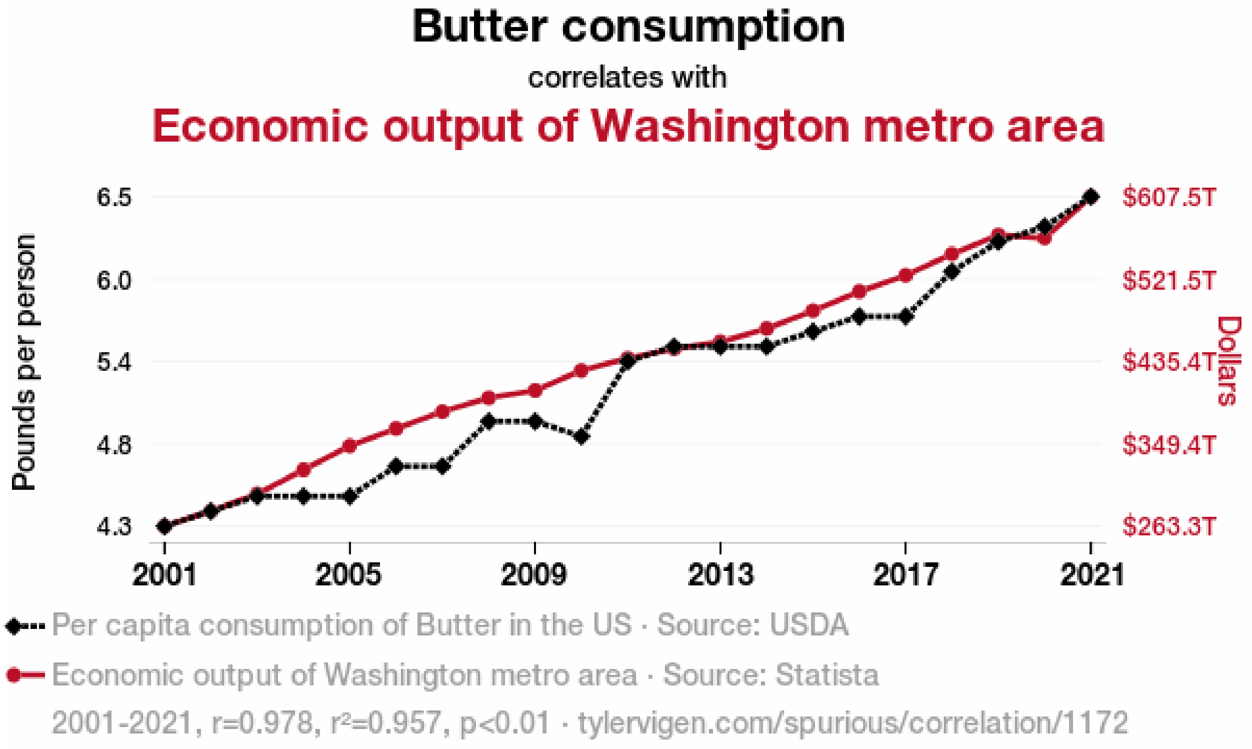
1 Background & Motivation for Causal Inference in Service Evaluations

When performing evaluations of healthcare services or interventions in the NHS, our key evaluation question of interest is almost always causal in nature:

- Has our new breast cancer screening programme reduced mortality?
- Did implementing a community-based model of care divert patients away from A&E services?
- Did implementing a hub and spoke model reduce clinical harm and lower patient waiting times?

Yet many of the methods currently used to evaluate interventions within the NHS are often descriptive in nature. Dashboards are frequently used in this manner, where any changes in an outcome of interest are fully attributed to a single change in implementation in services, without considering whether there are alternative explanations and/or other confounding factors are involved that could instead explain the change instead.

To better illustrate this point, Tyler Vigen publishes great examples of spurious correlations on his webpage (https://tylervigen.com/spurious-correlations), from which I've attached an example of below (Vigen 2024):



As you can see, butter consumption is positively correlated with the economic output of the Washington metro area. Needless to say, I highly doubt butter consumption was actually "greasing the wheels" of the local economy in the Washington metro area! It is therefore imperative that we don't conflate association with causation when we're conducting service evaluation as well.

Many books, papers, and reports have been published providing guidance on how one can estimate the causal effect of different interventions have on specific outcomes, in the absence of a randomised trial. Some resources have even been developed to help narrow down which types of methods would be most appropriate depending on the context. However, much of the guidance has been limited to methods originating from a specific field (e.g., econometrics, social sciences, epidemiology), rather than encompassing all approaches relevant to service and intervention evaluation within the realm of health and social care.

This handbook aims to bridge this gap in the literature, and establish new unifying guidance for selecting appropriate study designs and analytical methodologies for conducting causal inference in service evaluations. You'll find:

1. A decision tree diagram to aid with method selection
2. Concise overviews of each approach
3. Pointers to more in-depth resources for technical details about each method (including technical explanations in textbooks, applied examples in the literature, and practical coding exercises)

Do note that this handbook itself will not be going into too much technical detail about each analytical method, but instead serves as a resource to guide analysts in choosing the appropriate to use for their question.

I would also like to note that this handbook will primarily focus largely on established statistical methods for causal inference. More novel and experimental machine learning based methods do exist for causal inference. However, most of these methods require additional computational power, have components that are more opaque (akin to a black box), and many of these still haven't been extensively used in the literature yet. For the purposes of keeping this guide as approachable as possible, and our methods and assumptions as transparent as we can, this is the focus we have chosen for this handbook. An extensions section will be included at the end for those who may be interested in further delving into novel machine methods as well.

2 A Brief History on Conceptualising Causality

2.1 The Potential Outcomes Framework

The field of causal inference has a long history built on insights from multiple disciplines. Multiple schools of thought have been developed, contributing different ideas on how to conceptualise and define the concept of causality.

One of the most widely adopted ways to conceptualise causal inference in the fields of applied statistics, econometrics, and epidemiology is through the idea of potential outcomes, also referred to as the counterfactual approach to causality. This is where we aim to determine if the outcome for an individual would have been different if given different hypothetical value(s) of the exposure of interest.

Such counterfactual ideas stretch back to key texts written during the early to mid 19th century:

"If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the cause of his death" (Mill 1843).

However, it wasn't until Neyman's 1923 Master's thesis where the notation for potential outcomes was first formalised, introducing the idea that within randomized experiments, if we're interested in estimating the effect of a treatment or intervention, each unit in a study for a binary outcome can only have two fixed potential outcomes, one under treatment and one under control (Splawa-Neyman, Dabrowska, and Speed 1990).

In 1925, Fisher proposed randomizing treatments to study units/participants in order to derive unbiased inferences from experiments, though without reference to potential outcomes or estimating average treatment effects (Ronald A. Fisher 1925; R. A. Fisher and Yates 1990).

This changed in 1974, where Rubin formally established what we now refer to as the potential outcomes framework, extending these ideas to non-randomized observational studies, and writing formal notation for calculating the average casual effect that is used today (Rubin 1974). Holland later named this the Rubin Causal Model, and emphasised the fundamental problem of causal inference:

"It is impossible to observe [the value of the response that would be observed if the unit was exposure to treatment] and [the value that would be observed on the same unit if it were exposed to the control], and therefore it is impossible to observe the effect of [treatment] on [the unit of interest]" (Holland, Glymour, and Granger 1985).

In short, if we're interested in evaluating a treatment, one cannot observe both hypothetical outcomes for a single unit or individual, one in a universe where the unit/individual of interest is treated, and another where the same unit/individual is not treated.

Because the potential outcomes framework mirrors the logic found in randomized trials, the school of thought has underpinned virtually every method covered in this handbook, including those with origins in econometrics and epidemiology.

2.2 Structural Causal Models and Directed Acyclic Graphs

However, alongside the counterfactual school of thought, competing theories were developed in parallel, the most prominent of which came from Pearl. Rather than framing causation in terms of counterfactual outcomes, Pearl popularised the development of Structural Causal Models (SCMs), which use Directed Acyclic Graphs (DAGs) and structural equations to present causal relationships and underlying assumptions in both a visual and mathematical way (Pearl 2009; Pearl and Mackenzie 2018).

I will be delving into the specifics of Directed Acyclic Graphs in their own section later into this handbook. As a brief overview though, DAGs are diagrams which include each of your variables (exposure, covariates, and outcomes) as nodes, connected by arrows that indicate a direct causal relationship between them. Digitale et al., from UCSF have a great tutorial on DAGs if you'd like a brief introduction on the topic (Digitale, Martin, and Glymour 2022).

Historically when these methods were introduced, the two schools of thought were completely split, with econometricians favouring the potential outcomes approach, while computer scientists favoured structural causal models. A great discussion on this is provided by Nobel laureate Imbens (an economist) more recently, who argues that although these two frameworks are complementary, the potential outcomes approach is often preferred for econometrics because it aligns naturally with study designs that exploit policy changes or other "as-if random" interventions, rather than collecting data on and adjusting for a wide range of individual variables mapped out on a DAG (Imbens 2020). This type of thinking was how the field of quasi-experimental methods originated, the specifics of which will be discussed in later sections of the handbook.

Overtime though, some fields such as Epidemiology and Public Health have found ways to combine both school of thought into a single workflow:

1. DAGs are drawn to visualise causal relationships, identify confounders of interest, present assumptions made transparently.
2. Use back door criteria on the DAG to select the minimal set of covariates that require adjustment of to make an unbiased estimate of the causal contrast of interest (e.g., average treatment effect, average treatment effect on the treated).
3. Apply appropriate statistical methods to adjust for the set of covariates, such as regression modelling, to estimate the target causal contrast, under the potential outcomes framework. This hybrid approach works particularly well in settings with clearly defined treatment contrasts (say, drug X vs. drug Y) and rich data from longitudinal studies and electronic health records, which are harder to come by in economics.

This hybrid approach works particularly well in settings with clearly defined treatment contrasts (say, drug X vs. drug Y) and rich data from longitudinal studies and electronic health records, which are harder to come by in economics.

If you are interested in reading further into the origins of the potential outcome framework and quasi-experimental designs, Cunningham's supplementary teaching material hosted on github as part of his "Causal Inference: The Mixtape" book goes into this in more depth (Cunningham 2025). His slides of the "Foundation of Causality" in particular focuses on the history of these foundational ideas. The introductory chapter of Morgan and Winships' "Counterfactuals and Causal Inference," as well as chapter 2 of Imbens and Rubins' "Causal Infernece for Statistics, Social, and Biomedical Sciences" also provide a great overview of the historical context behind how causal inference as we know it today was first formalised (Morgan and Winship 2014; Imbens and Rubin 2015).

2.3 Additional Approaches to Establishing Causality

During the mid-19th century, before Rubin had established the potential outcomes framework, Hill had notably published a seminal paper covering a set of criteria that would help establish causality within epidemiological studies, now commonly referred to as the "Bradford Hill Criteria." These included: strength, consistency, specificity, temporality, biological gradient, biological plausibility, coherence, experimental evidence, and analogy (Hill 1965). The criteria, however, did not provide a practical way for epidemiologists and social scientists more generally to design observational studies targeting causal inference though. The criteria is therefore used more so as a guide for evaluating whether associations are causal while evaluating a set of evidence, and still often used in policy evaluations. Chapter 2 of Lash et al.'s textbook "Modern Epidemiology" provides a great explanation of each criteria if you'd like to delve deeper into this.

Additional causal frameworks have been built on top of this foundation as well, namely the Target Trial Emulation framework by Hernan and Robins which will be discussed in further detail in its own section (M. A. Hernán and Robins 2016). The framework aims to overcome many of the methodological issues facing classical epidemiological analyses by framing causal questions of interest within an ideal pragmatic randomized trial. This will underpin much of the methods covered as part of the "Epidemiology Methods" section further into the handbook.

3 Common Assumptions in Causal Inference Methodology

Before delving into individual methods, here's a short reference of common assumptions that will be referred to throughout this handbook. Not every method relies on all of these assumptions, but this will serve as your go to reference point in case there is any confusion between them later on. Method specific assumptions will be detailed within each methods section as well.

3.1 Exchangeability

Markozannes et al., define exchangeability as holding when: *"Treatment assignment is independent of the potential outcomes; this roughly translates to no unmeasured confounders and no informative censoring"* (Markozannes, Vourli, and Ntzani 2021).

Exchangeability can be divided into two types:

1. Unconditional exchangeability
    - Where in a randomized trial, potential outcomes are inherently independent from treatment assignment due to random assignment of treatment.
2. Conditional exchangeability
    - Where in an observational study, potential outcomes are only independent from treatment assignment given/controlling for a set of measured covariates.

Hernan and Robins' definition is framed more as conditional exchangeability in their book, given its focus on observational studies, and define exchangeability as holding when: *"The conditional probability of receiving every value of treatment, though not decided by the investigators, depends only on measured covariates,"* i.e. no residual confounding (M. A. Hernán and Robins 2025).

For example, if we're studying the relationship between smoking and cardiovascular disease, and sex is known to be independently associated between the two variables and does not lie on the causal pathway (i.e. a known confounder), we would violate the condition by not adjusting for sex in our analysis.

Exchangeability is also referred to as ignorability (ignorable treatment assignment mechanism) or unconfoundedness in the literature.

3.2 Consistency

Hernan and Robins define consistency as holding when: "The values of treatment under comparison correspond to well-defined interventions that, in turn, correspond to the versions of treatment in the data" (M. A. Hernán and Robins 2025).

For example, in a study evaluating the effect of paracetamol on chronic pain, if some participants have one dose a week, and others have one daily, that would violate the condition, as the intervention is not uniformly defined.

3.3 Positivity

Hernan and Robins define positivity as holding when: "The probability of receiving every value of treatment conditional on relevant covariants is greater than zero, i.e., positive" (M. A. Hernán and Robins 2025).

For example, if we're studying the effect of physical activity on the incidence of obesity, participants who are disabled and bedbound have a 0% chance of receiving the exercise intervention, thereby violating the condition.

3.4 Stable Unit Treatment Value Assumption (SUTVA)

The Stable Unit Treatment Value Assumption (SUTVA), as defined by Rubin who coined the term: *"Is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive"* (Rubin 1986).

A simpler definition and explanation has been published by Markozannes et al.,: *"The stable unit treatment value assumption states that there is no interference among units, that is, the treatment status of a unit does not affect the potential outcomes of other units and it also requires that there is only a single version of the treatment (no hidden variations in treatment; no multiple versions of treatment)." "Possible violations of the SUTVA include settings where units interact (e.g., schools, group interventions) or different treatment dosages exist or different modes of administration operate which can affect the potential outcomes"* (Markozannes, Vourli, and Ntzani 2021).

In other words, we can define SUTVA has having two distinct requirements:

1. No multiple versions of treatment
2. No interference between units

A more concrete example would be in the context of evaluating the efficacy of a new flu vaccine, where the assumption would hold as long as:

1. All versions of the flu vaccines are the same, and are administered in the exact same manner between study participants (i.e., have the same dosage, formulation, and schedule)
2. Each study participant is isolated in separate facilities, with limited contact with other trial participants and non-participants

Of course, in practice SUTVA would not hold in an evaluation of a flu vaccine, as study participants would come into contact with other trial participants and non-participants, introducing herd level spillover effects into the population. Do note though that defining the unit of interest differently could change this depending on conditions (e.g., going from an individual level analysis to a city or country level analysis).

SUTVA is also referred to as a non-macro-effect or the partial equilibrium assumption in economics (Morgan and Winship 2014).

Assumptions made when using some of the most common statistical techniques discussed in this handbook can be found in this paper.

4 Existing Guidance on Study Design Selection

Prior to presenting my unifying guidance in my next section, I've attached a few existing reports and guides for selecting the appropriate observational causal study design relevant to your key evaluation question of interest. This was done to acknowledge the work that has come before, while also noting key limitations that motivated the development of this unifying guidance.

4.1 NICE Guidance

In their report on "methods for the development of NICE public health guidelines," the National Institute for Health and Care Excellence (NICE) provides a high-level overview of the various epidemiological study designs used in public health research. Included within the report in Appendix E is a flowchart that outlines an "algorithm for classifying quantitative (experimental and observational) study designs (NICE 2012).

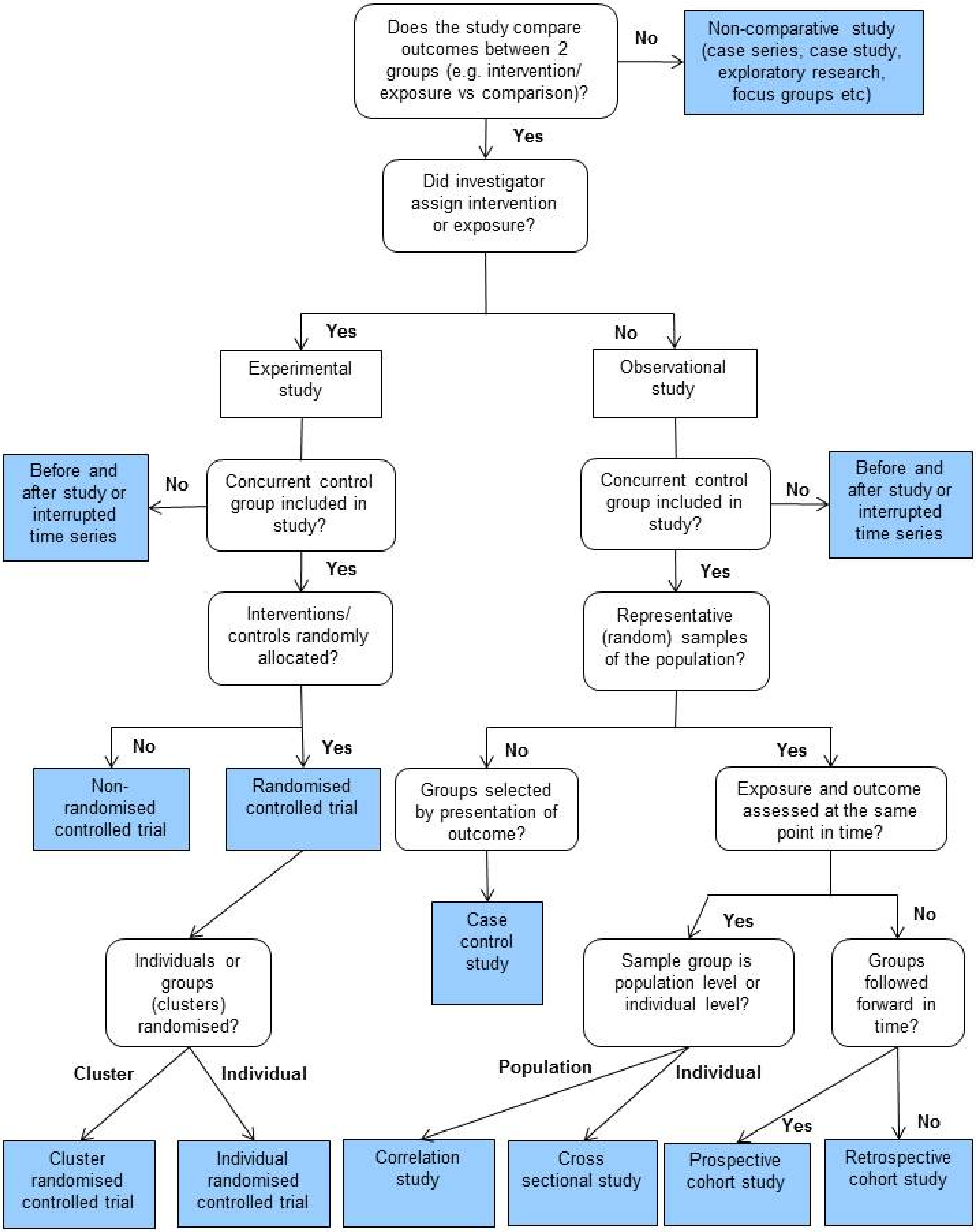

Does the study compare outcomes between 2 groups (e.g. intervention/ exposure vs comparison)?
No
Non-comparative study (case series, case study, exploratory research, focus groups etc)
Yes
Did investigator assign intervention or exposure?
Yes
No
Experimental study
Observational study
Before and after study or interrupted time series
No
Concurrent control group included in study?
Concurrent control group included in study?
No
Before and after study or interrupted time series
Yes
Yes
Interventions/ controls randomly allocated?
Representative (random) samples of the population?
No
Yes
No
Yes
Non-randomised controlled trial
Randomised controlled trial
Groups selected by presentation of outcome?
Exposure and outcome assessed at the same point in time?
Case control study
Yes
No
Individuals or groups (clusters) randomised?
Sample group is population level or individual level?
Groups followed forward in time?
Cluster
Individual
Population
Individual
Yes
No
Cluster randomised controlled trial
Individual randomised controlled trial
Correlation study
Cross sectional study
Prospective cohort study
Retrospective cohort study

Figure 4.1: Algorithm for classifying quantitative (experimental and observational) study designs (NICE, 2012)

As you can see, the types of study designs covered are very general, and do not focus on specific types of analytical methods you can use to deal with confounding factors within your study to derive causal inferences. As such, this diagram is most useful for those who are new to the field of epidemiology and public health, and wish to refer to a general overview of experimental and observational study designs.

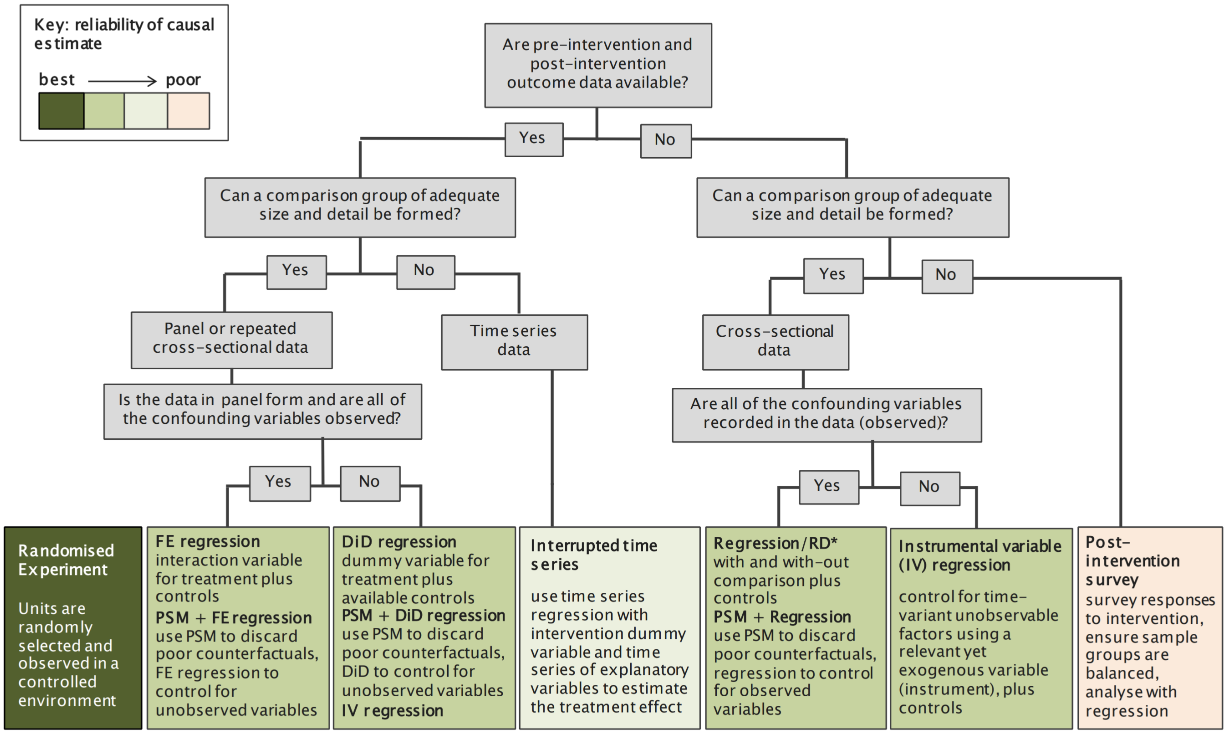
4.2 OHID Guidance

The Office for Health Improvement and Disparities (OHID) has similarly published some high-level guidance on the use of various evaluation methods, including quasi-experimental methods, to evaluate digital health products (OHID 2021a, 2021b). This guidance is aimed at analysts that are completely new to these methods though, and as such does not go into the specifics of how to perform these analyses in a programming language and/or statistical package such as R, Stata, and Python.

4.3 New Zealand Transport Agency Guidance

The New Zealand Transport Agency Waka Kotahi has published a report focused on the use of causal inference methods for the evaluation of transport interventions, and includes a causal inference method selection flowchart (Schiff, Wright, and Denne 2017).

The methods included are limited to those most often used in the social sciences and econometrics though. These methods largely exploit natural experiments that have occurred in the real world at one point in time to derive causal inferences. There is hence room to expand upon this flowchart further in a more comprehensive manner.


Key: reliability of causal estimate
best
poor
Are pre-intervention and post-intervention outcome data available?
Yes
No
Can a comparison group of adequate size and detail be formed?
Yes
No
Can a comparison group of adequate size and detail be formed?
Yes
No
Panel or repeated cross-sectional data
Time series data
Cross-sectional data
Is the data in panel form and are all of the confounding variables observed?
Are all of the confounding variables recorded in the data (observed)?
Yes
No
Yes
No
Randomised Experiment
Units are randomly selected and observed in a controlled environment
FE regression
interaction variable for treatment plus controls
PSM + FE regression
use PSM to discard poor counterfactuals, FE regression to control for unobserved variables
DiD regression
dummy variable for treatment plus available controls
PSM + DiD regression
use PSM to discard poor counterfactuals, DiD to control for unobserved variables
IV regression
Interrupted time series
use time series regression with intervention dummy variable and time series of explanatory variables to estimate the treatment effect
Regression/RD*
with and with-out comparison plus controls
PSM + Regression
use PSM to discard poor counterfactuals, regression to control for observed variables
Instrumental variable (IV) regression
control for time-variant unobservable factors using a relevant yet exogenous variable (instrument), plus controls
Post-intervention survey
survey responses to intervention, ensure sample groups are balanced, analyse with regression

* Regression Discontinuity (RD) is a special type of cross section regression used when there is no overlap between treatment and control groups.

Figure 4.2: Causal inference method selection flowchart (Schiff, Wright and Denne, 2017)

4.4 HM Treasury Guidance

The HM Treasury's Magenta Book, specifically Figure 3.1 on p.47 (PDF p.54) provides a great diagram providing guidance on when using different experimental and quasi-experimental methods for service evaluation would be the most appropriate (HM Treasury 2020). Unlike the New Zealand Transport Agency's guidance, the Treasury's diagram is more descriptive in nature, describing specific contexts where each approach would be most applicable.

However, similar to the New Zealand Transport Agency's guidance, the quasi-experimental methods presented only encompass methods used more often in econometrics and social sciences, such as difference in differences and regression discontinuity.

These methods once again largely exploit natural experiments that have occurred in the real world to derive causal inferences, rather than adjusting for individual covariates which is more popular in epidemiology. Methods focusing on the latter are absent, but this is no surprise due to the Treasury's area of focus. The guidance presented in this handbook therefore expands upon this diagram to also include such methods.

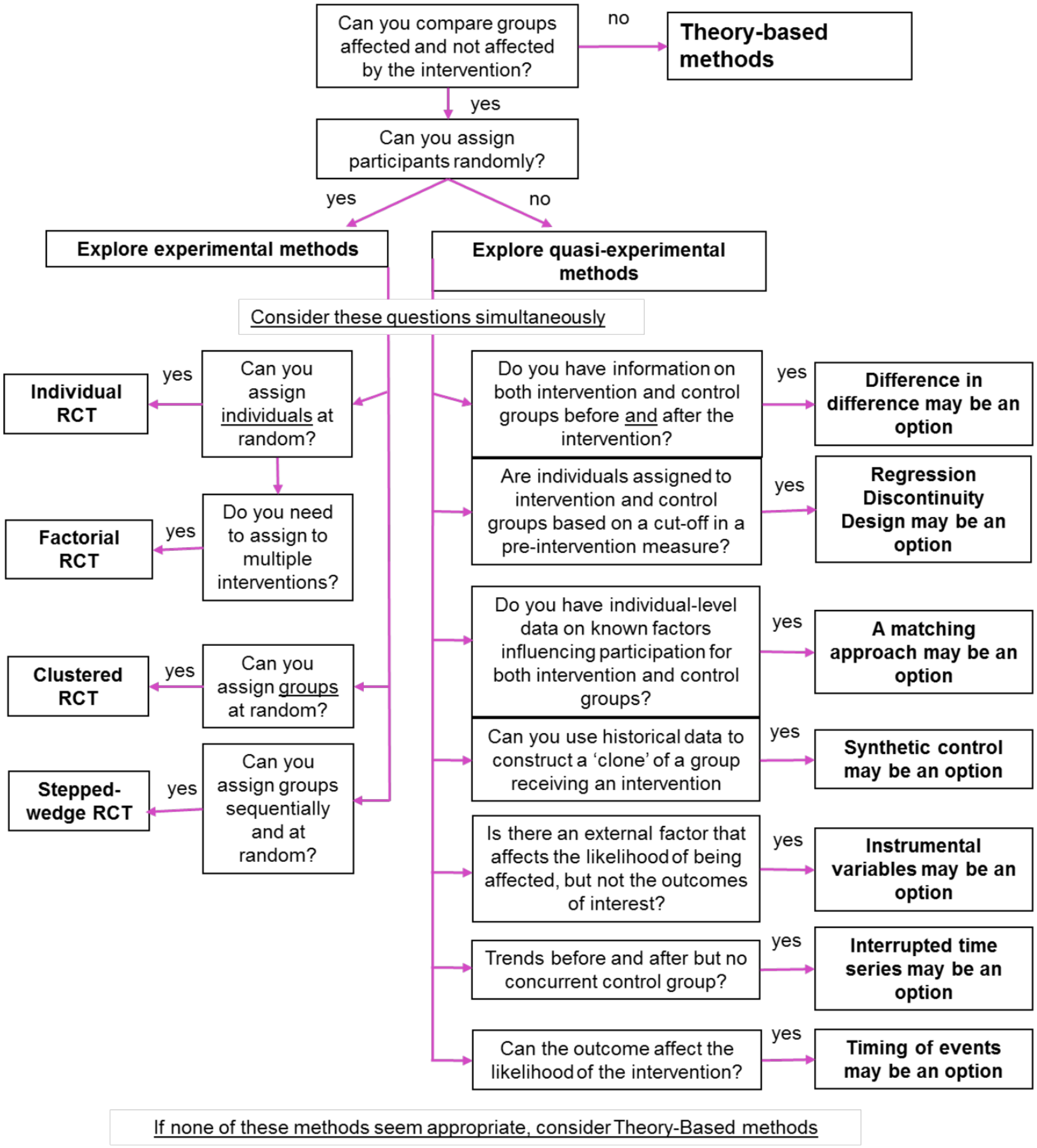

Can you compare groups affected and not affected by the intervention?
no
Theory-based methods
yes
Can you assign participants randomly?
yes
no
Explore experimental methods
Explore quasi-experimental methods
Consider these questions simultaneously
Can you assign individuals at random?
yes
Individual RCT
Do you need to assign to multiple interventions?
yes
Factorial RCT
Can you assign groups at random?
yes
Clustered RCT
Can you assign groups sequentially and at random?
yes
Stepped-wedge RCT
Do you have information on both intervention and control groups before and after the intervention?
yes
Difference in difference may be an option
Are individuals assigned to intervention and control groups based on a cut-off in a pre-intervention measure?
yes
Regression Discontinuity Design may be an option
Do you have individual-level data on known factors influencing participation for both intervention and control groups?
yes
A matching approach may be an option
Can you use historical data to construct a 'clone' of a group receiving an intervention
yes
Synthetic control may be an option
Is there an external factor that affects the likelihood of being affected, but not the outcomes of interest?
yes
Instrumental variables may be an option
Trends before and after but no concurrent control group?
yes
Interrupted time series may be an option
Can the outcome affect the likelihood of the intervention?
yes
Timing of events may be an option
If none of these methods seem appropriate, consider Theory-Based methods

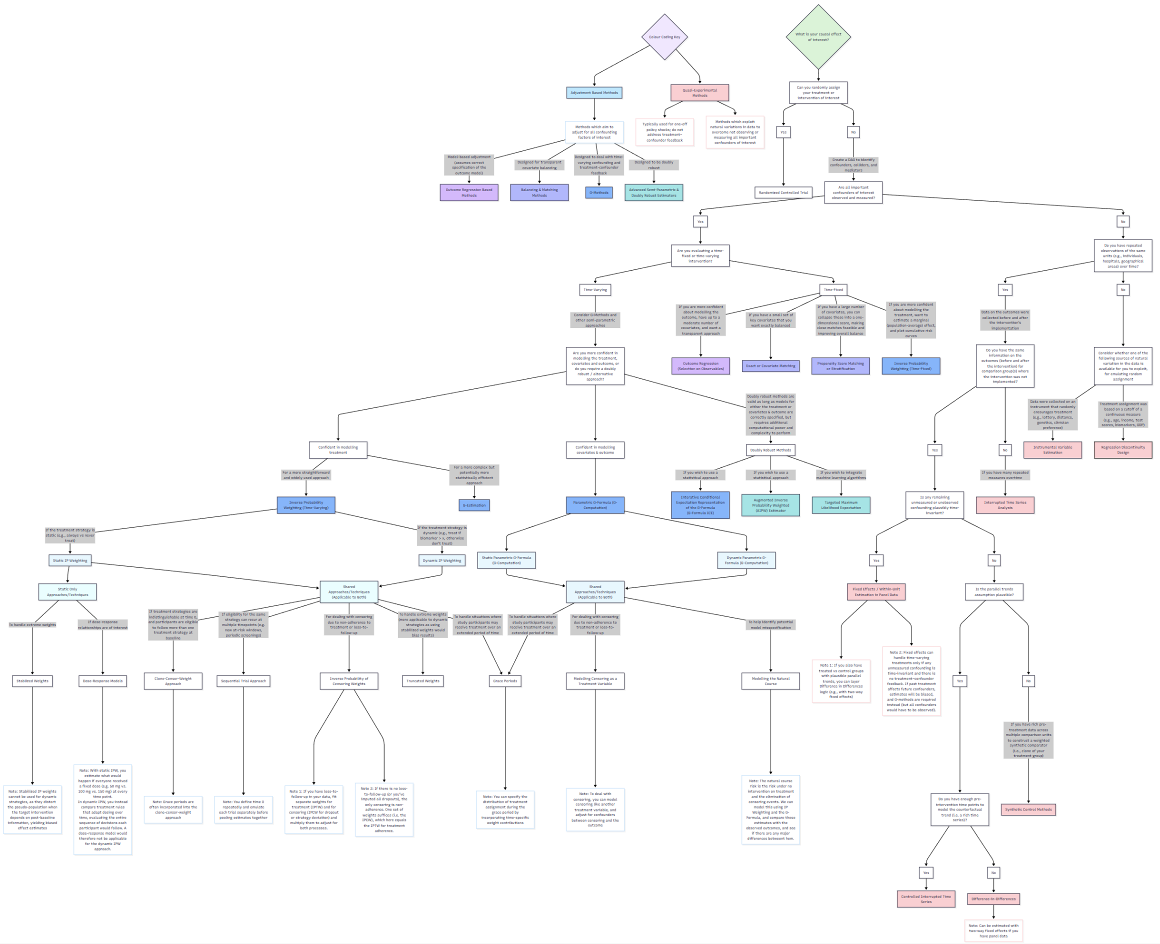
Figure 4.3: Diagram for selecting experimental and quasi-experimental methods (HM Treasury, 2020)

5 Unifying Guidance on Choosing Causal Inference Methods (Decision Tree Diagram)

Given the narrow scope of all prior guidance, especially as most of the methods selection diagrams are limited to methods used more within other fields of study, I have opted to create a new method selection tree diagram to help advise with the selection of causal methods going forward.

A flowchart of all methods mentioned in this handbook is presented below, with descriptors indicating which variables are most appropriate given your evaluation question of interest, and the data available to you. Corresponding chapter numbers of each method are also attached within the flowchart.

Please note that the diagram is still a work in progress, and will continue to be updated with input from external experts.



Colour Coding Key
Adjustment Based Methods
Quasi-Experimental Methods
Methods which aim to adjust for all confounding factors of interest
Typically used for one-off policy shocks; do not address treatment–confounder feedback
Methods which exploit natural variations in data to overcome not observing or measuring all important confounders of interest
Model-based adjustment (assumes correct specification of the outcome model)
Designed for transparent covariate balancing
Designed to deal with time-varying confounding and treatment-confounder feedback
Designed to be doubly robust
Outcome Regression Based Methods
Balancing & Matching Methods
G-Methods
Advanced Semi-Parametric & Doubly Robust Estimators
What is your causal effect of interest?
Can you randomly assign your treatment or intervention of interest
Yes
No
Create a DAG to identify confounders, colliders, and mediators
Randomized Controlled Trial
Are all important confounders of interest observed and measured?
Yes
No
Are you evaluating a time-fixed or time-varying intervention?
Time-Varying
Time-Fixed
Consider G-Methods and other semi-parametric approaches
If you are more confident about modelling the outcome, have up to a moderate number of covariates, and want a transparent approach
If you have a small set of key covariates that you want exactly balanced
If you have a large number of covariates, you can collapse these into a one-dimensional score, making close matches feasible and improving overall balance
If you are more confident about modelling the treatment, want to estimate a marginal (population-average) effect, and plot cumulative risk curves
Outcome Regression (Selection on Observables)
Exact or Covariate Matching
Propensity Score Matching or Stratification
Inverse Probability Weighting (Time-Fixed)
Do you have repeated observations of the same units (e.g., individuals, hospitals, geographical areas) over time?
Yes
No
Data on the outcomes were collected before and after the intervention's implementation
Do you have the same information on the outcomes (before and after the intervention) for comparison group(s) where the intervention was not implemented?
Consider whether one of the following sources of natural variation in the data is available for you to exploit, for emulating random assignment
Data were collected on an instrument that randomly encourages treatment (e.g., lottery, distance, genetics, clinician preference)
Treatment assignment was based on a cutoff of a continuous measure (e.g., age, income, test scores, biomarkers, GDP)
Instrumental Variable Estimation
Regression Discontinuity Design
Are you more confident in modelling the treatment, covariates and outcome, or do you require a doubly robust / alternative approach?
Doubly robust methods are valid as long as models for either the treatment or covariates & outcome are correctly specified, but requires additional computational power and complexity to perform
Doubly Robust Methods
If you wish to use a statistical approach
If you wish to use a statistical approach
If you wish to integrate machine learning algorithms
Interative Conditional Expectation Representation of the G-Formula (G-Formula ICE)
Augmented Inverse Probability Weighted (AIPW) Estimator
Targeted Maximum Likelihood Expectation
Yes
No
If you have many repeated measures overtime
Interrupted Time Series Analysis
Is any remaining unmeasured or unobserved confounding plausibly time-invariant?
Confident in modelling treatment
For a more straightforward and widely used approach
For a more complex but potentially more statistically efficient approach
Inverse Probability Weighting (Time-Varying)
G-Estimation
Confident in modelling covariates & outcome
Parametric G-Formula (G-Computation)
If the treatment strategy is static (e.g., always vs never treat)
If the treatment strategy is dynamic (e.g., treat if biomarker &gt; x, otherwise don't treat)
Static IP Weighting
Dynamic IP Weighting
Static Parametric G-Formula (G-Computation)
Dynamic Parametric G-Formula (G-Computation)
Static Only Approaches/Techniques
Shared Approaches/Techniques (Applicable to Both)
Shared Approaches/Techniques (Applicable to Both)
Yes
No
Fixed Effects / Within-Unit Estimation in Panel Data
Is the parallel trends assumption plausible?
To handle extreme weights
If dose-response relationships are of interest
If treatment strategies are indistinguishable at time 0, and participants are eligible to follow more than one treatment strategy at baseline
If eligiblity for the same strategy can recur at multiple timepoints (e.g. new at-risk windows, periodic screenings)
For dealing with censoring due to non-adherence to treatment or loss-to-follow-up
To handle extreme weights (more applicable to dynamic strategies as using stabilized weights would bias results)
To handle situations where study participants may receive treatment over an extended period of time
To handle situations where study participants may receive treatment over an extended period of time
For dealing with censoring due to non-adherence to treatment or loss-to-follow-up
To help identify potential model misspecification
Note 1: If you also have treated vs control groups with plausible parallel trends, you can layer Difference in Differences logic (e.g., with two-way fixed effects)
Note 2: Fixed effects can handle time-varying treatments only if any unmeasured confounding is time-invariant and there is no treatment-confounder feedback. If past treatment affects future confounders, estimates will be biased, and G-methods are required instead (but all confounders would have to be observed).
Yes
No
If you have rich pre-treatment data across multiple comparison units to construct a weighted synthetic comparator (i.e., clone of your treatment group)
Stablized Weights
Dose-Response Models
Clone-Censor-Weight Approach
Sequential Trial Approach
Inverse Probability of Censoring Weights
Truncated Weights
Grace Periods
Modelling Censoring as a Treatment Variable
Modelling the Natural Course
Note: Stabilized IP weights cannot be used for dynamic strategies, as they distort the pseudo-population when the target intervention depends on post-baseline information, yielding biased effect estimates
Note: With static IPW, you estimate what would happen if everyone received a fixed dose (e.g. 50 mg vs. 100 mg vs. 150 mg) at every time point. In dynamic IPW, you instead compare treatment rules that adapt dosing over time, evaluating the entire sequence of decisions each participant would follow. A dose-response model would therefore not be applicable for the dynamic IPW approach.
Note: Grace periods are often incorporated into the clone-censor-weight approach
Note: You define time 0 repeatedly and emulate each trial separately before pooling estimates together
Note 1: If you have loss-to-follow-up in your data, fit separate weights for treatment (IPTW) and for censoring (IPCW for dropout or strategy deviation) and multiply them to adjust for both processes.
Note 2: If there is no loss-to-follow-up (or you've imputed all dropouts), the only censoring is non-adherence. One set of weights suffices (i.e. the IPCW), which here equals the IPTW for treatment adherence.
Note: You can specify the distribution of treatment assignment during the grace period by incorporating time-specific weight contributions
Note: To deal with censoring, you can model censoring like another treatment variable, and adjust for confounders between censoring and the outcome
Note: The natural course risk is the risk under no intervention on treatment and the elimination of censoring events. We can model this using IP Weighting and the G-Formula, and compare these estimates with the observed outcomes, and see if there are any major differences betweent hem.
Do you have enough pre-intervention time points to model the counterfactual trend (i.e. a rich time series)?
Synthetic Control Methods
Yes
No
Controlled Interrupted Time Series
Difference-in-Differences
Note: Can be estimated with two-way fixed effects if you have panel data

6 Key Resources

Throughout the handbook, there are a few key influential texts that have consolidated contemporary methods used to derive causal inferences from observational data. These texts are widely regarded as being foundational to anyone attempting to dip their toes into the field, and are often used as required reading and/or reference material within undergraduate and postgraduate courses in the econometrics, epidemiology, statistics, and social sciences more broadly.

I've listed some the most frequently cited texts in this handbook below, along with a brief description of their areas of focus. Additional texts that are beyond the scope of this handbook have also been included for reference. A summary table is also included showing which methods are covered by each book in Section 1.7.

Brady Neal has posted a very useful flowchart on his personal website outlining "which causal inference book you should read," and covers many of the texts mentioned in this handbook (Neal 2019). If you're completely new to the field of causal inference, his diagram serves as a great place to begin your journey.

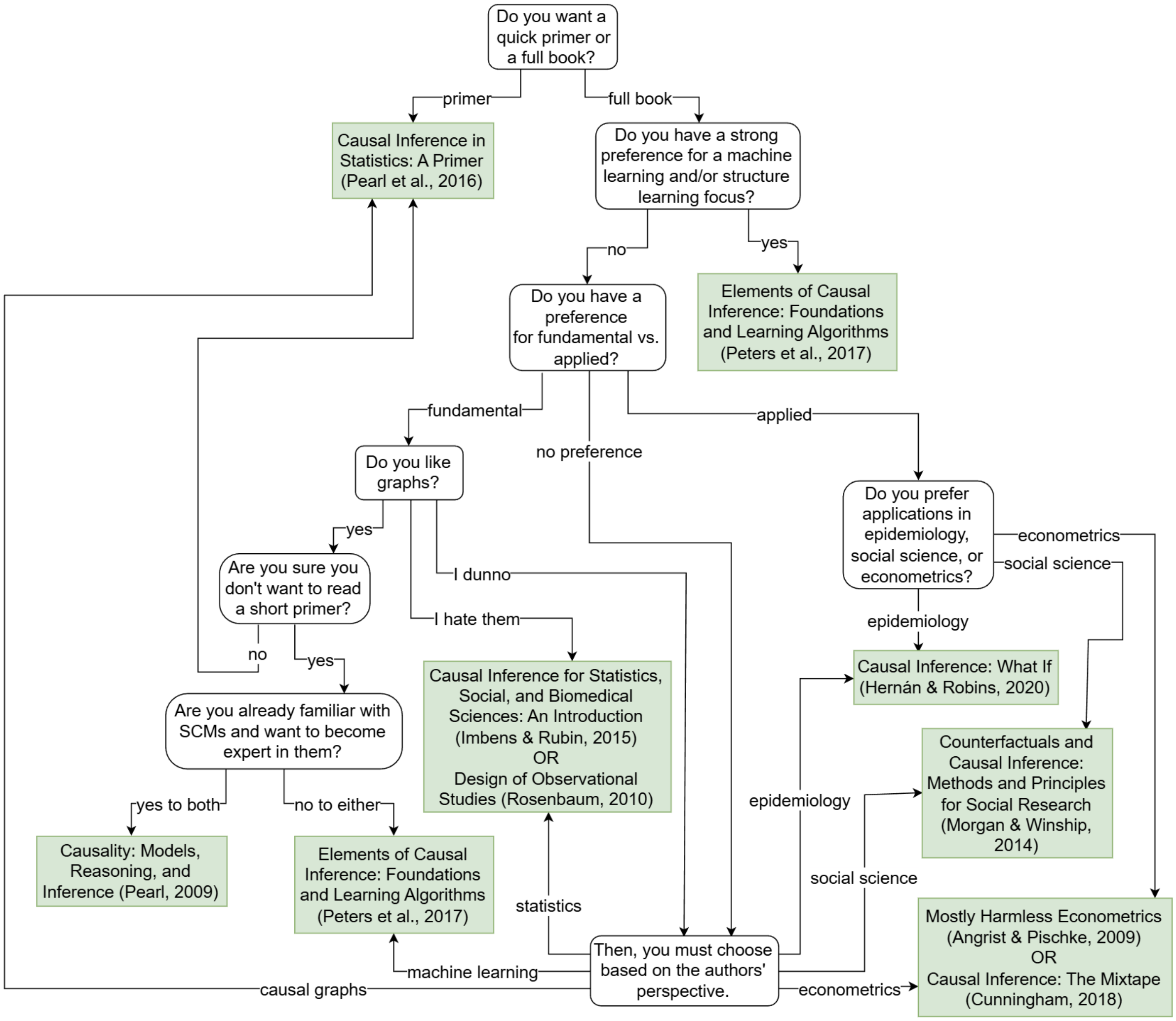

Do you want a quick primer or a full book?
primer
full book
Causal Inference in Statistics: A Primer (Pearl et al., 2016)
Do you have a strong preference for a machine learning and/or structure learning focus?
no
yes
Elements of Causal Inference: Foundations and Learning Algorithms (Peters et al., 2017)
Do you have a preference for fundamental vs. applied?
fundamental
no preference
applied
Do you like graphs?
Do you prefer applications in epidemiology, social science, or econometrics?
econometrics
social science
epidemiology
yes
I dunno
I hate them
Are you sure you don't want to read a short primer?
no
yes
Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction (Imbens & Rubin, 2015) OR Design of Observational Studies (Rosenbaum, 2010)
Causal Inference: What If (Hernán & Robins, 2020)
Counterfactuals and Causal Inference: Methods and Principles for Social Research (Morgan & Winship, 2014)
Are you already familiar with SCMs and want to become expert in them?
yes to both
no to either
Causality: Models, Reasoning, and Inference (Pearl, 2009)
Elements of Causal Inference: Foundations and Learning Algorithms (Peters et al., 2017)
epidemiology
social science
statistics
Mostly Harmless Econometrics (Angrist & Pischke, 2009) OR Causal Inference: The Mixtape (Cunningham, 2018)
Then, you must choose based on the authors' perspective.
machine learning
causal graphs
econometrics

Figure 6.1: Causal Books Flowchart (Neal, 2019)

6.1 Causal Inference: What If (M. A. Hernán and Robins 2025)

Authors
Miguel Hernan – *Kolokotrones Professor of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health & Director of Harvard CAUSALab*

James M. Robins – *Mitchell L. and Robin LaFoley Dong Professor of Epidemiology, Harvard T.H. Chan School of Public Health*

Content
Hernan and Robins' book was first published in 2020, and has been continuously updated over the past few years. The most up to date version of the book can be found on Hernan's personal webpage.

The book is considered by many in the field of epidemiology to be the definitive guide to the application of causal inference methods, framed within the target trial emulation framework, especially when it comes to estimating causal effects using complex longitudinal data (with a strong focus on G-methods). This is a result of the book being openly available online for the past two decades while it was still being drafted, with many researchers in the field contributing to the book over this period.

The book is particularly useful in providing guidance for evaluating sustained and time-varying treatment strategies, and overcoming issues such as immortal time bias within classical methods in epidemiology, which are often issues within the field of pharmacoepidemiology.

The book also (very helpfully) includes accompanying R and Stata code hosted on GitHub, which provides practical examples of how the methods covered can be applied to your own data.

6.2 Causal Inference: The Mixtape (Cunningham 2021)

Author
Scott Cunningham – *Ben H. Williams Professor of Economics at Baylor University*

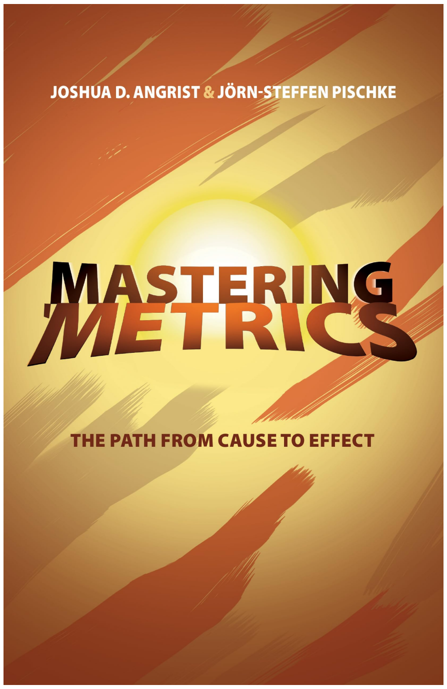
Content
Cunningham's book aimed to consolidate many of the causal inference methods from econometrics (e.g., regression discontinuity, instrumental variables, difference-in-differences) that were scattered across various other textbooks, and in addition provides worked examples of Stata, R, and Python code for each section.

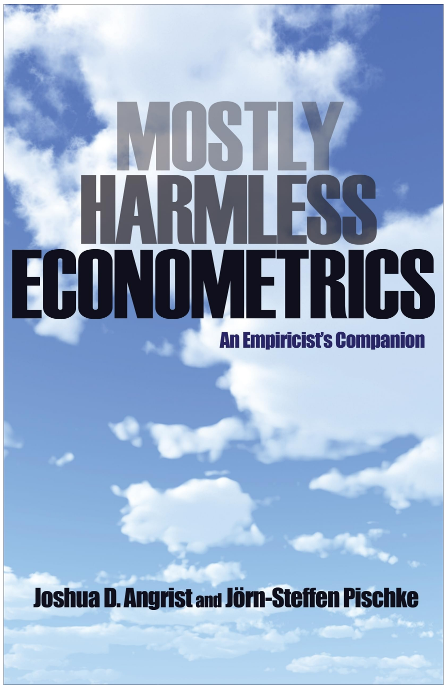
More specifically, Cunningham was mainly inspired by the work of Morgan and Winship, Angrist and Pischke, as well as Imbens and Rubin, which are all cited in this handbook as well. However, unlike these other texts, Cunningham's book is written in a very approachable manner for beginners, and is a great starting point to get to develop an understanding of various types of causal inference methods from an economist's perspective.

The book and accompanying code is available openly online.

Cunningham has also begun hosting a course based on the contents of the book, and has posted all the accompanying content (both slides and code) onto his GitHub page for those who wish to delve deeper into the history and application of these methods.

6.3 Mostly Harmless Econometrics: An Empiricist's Companion (Angrist and Pischke 2009) & Mastering 'Metrics: The Path from Cause to Effect (Angrist and Pischke 2015)


JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE
MASTERING
'METRICS
THE PATH FROM CAUSE TO EFFECT


MOSTLY
HARMLESS
ECONOMETRICS
An Empiricist's Companion
Joshua D. Angrist and Jörn-Steffen Pischke

Authors

Joshua D. Angrist – *Ford Professor of Economics at MIT and Nobel Prize Laureate (Economics)*

Jörn-Steffen Pischke – *Professor of Economics at LSE*

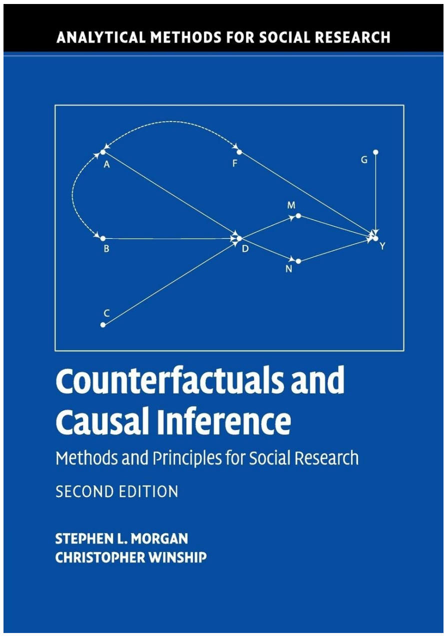
Content
Angrist and Pischke have written two textbooks on causal methods in econometrics, covering largely the same topics and methods across both: randomized trials, regression modelling, instrumental variables, regression discontinuity designs, and differences-in-differences.

However, Mastering 'Metrics is a much less technical book and easier to digest, especially for those new to causal inference methodology and/or do not have a background in mathematics. It therefore serves as a better starting point for newcomers, alongside Cumming's book, for those who want a book tackling causal inference form an economist's perspective, and is often included as accompanying reading for undergraduate and postgraduate level applied causal inference courses.

Stata code for the empirical examples presented in the book has also been shared publicly online. The examples have also been translated into R by Jeffrey Arnold on his GitHub page.

Mostly Harmless Econometrics provides an expanded overview of each of the methods covered within Mastering 'Metrics, but is written in a much more mathematical manner. There is no accompanying code with this book either. It may therefore be worth using as a resource to delve deeper in to specific methods covered after first reading through some of the other textbooks mentioned in this handbook first.

6.4 Counterfactuals and Causal Inference: Methods and Principles for Social Research (Morgan and Winship 2014)



ANALYTICAL METHODS FOR SOCIAL RESEARCH

Counterfactuals and Causal Inference

Methods and Principles for Social Research

SECOND EDITION

STEPHEN L. MORGAN
CHRISTOPHER WINSHIP

Authors
Stephen L. Morgan – *Bloomberg Distinguished Professor of Sociology and Education at Johns Hopkins University*

Christopher Winship – *Diker-Tishman Research Professor of Sociology*

Content
Morgan and Winships' book provides a deep dive into the potential outcomes framework, directed acyclic graphs (DAGs), and instrumental variables, with substantial mathematical detail. If you are particularly interested in the mathematical intuition behind these specific topics, this is a great book to refer to.

However, its sections on other econometrics/social science methods such as interrupted time series models and regression discontinuity designs are less detailed than that of other texts. It also does not provide any accompanying code. It is therefore best suited in helping analyst develop a stronger understanding of the intuition behind the potential outcomes framework and causal graphs.

6.5 Other Notable Texts

Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction (Imbens and Rubin 2015)

Building off of Rubin's potential outcomes framework, Imbens and Rubins' book is focused on methods for analysing randomized experiments, the potential outcomes framework, and specifically matching and instrumental variables. The topics covered are described in great mathematical detail. However, the text lacks content regarding some of the other methods covered as part of this handbook.

Causality: Models, Reasoning, and Inference (Pearl 2009)
Causal Inference in Statistics: A Primer (Pearl, Glymour, and Jewell 2016)
The Book of Why: The New Science of Cause and Effect (Pearl and Mackenzie 2018)

Pearl's collection of books provides options for learning more about structural causal models (SCMs) depending on the target audience, and is particularly useful for those looking to apply machine learning technical for modelling causal relationships, given Pearl's background in computer science:

- Causality (2009) serves as the most detailed and technical reference text for advanced readers, providing a comprehensive guide to DAGs and do-calculus while assuming that the reader has a strong mathematical background.
- The Primer (2016) distils all key concepts in causality in a manner that is more accessible to students and researchers.
- The Book of Why (2018) was written with a general audience in mind, introducing the DAG-focused way of causal thinking Pearl is most famous for in an approachable manner.

Observation and Experiment: An Introduction to Causal Inference (Rosenbaum 2017)
Design of Observational Studies (Rosenbaum 2020)
Causal Inference (Rosenbaum 2023)

Rosenbaum's collection of books provide an alternative source for getting a good overview of causal inference methods from a statistical perspective, given that he is a statistician by training. His books, similar to Pearl's have been written with different audiences in mind:

- Observation and Experiment (2017) serves as the most accessible introduction among the three. It is intended for a wide audience, including applied researchers and students, and focuses on the logic of design and inference in observational studies, without focusing too heavily on mathematical details. It introduces key concepts like matching, instrumental variables, and sensitivity analysis.
- Design of Observational Studies (2020) is a more technical text that dives deeply into matching methods, instrumental variables, and sensitivity analyses. It also puts strong emphasis on the fundamental design of observational studies, and how they can mimic randomized experiments as closely as possible.
- Causal Inference (2023) is Rosenbaum's most comprehensive and mathematically detailed text. It consolidates and extends many of the ideas from his earlier books while introducing more advanced content. It is best suited for researchers or methodologists with strong statistical backgrounds who want a complete overview of causal inference from Rosenbaum's design-based perspective.

Experimental and Quasi-Experimental Designs for Generalized Causal Inference (Shadish, Cook, and Campbell 2002)

Shadish, Cook, and Campbells' text, although slightly older than the rest of these, still provides a great overview of quasi-experiments, with a key focus on interrupted time series analysis and regression discontinuity designs. It has been used over the past two decades as a key tool for policy evaluation, and is a great reference text, especially as Campbell was the person who first coined and laid the groundwork for quasi-experimental study designs in 1963.

Targeted Learning: Causal Inference for Observational and Experimental Data (Van Der Laan and Rose 2011)
Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies (Van Der Laan and Rose 2018)

Van Der Laan and Roses' texts formalise the Targeted Maximum Likelihood Estimation (TMLE) framework, a semi-parametric approach to causal inference that integrates machine learning for outcome and propensity score estimation. They are best suited for readers with a background in statistics or computer science, who wish to delve deeper into combining machine learning models with causal inference when dealing with high-dimensional and longitudinal data.

Explanation in Causal Inference: Methods for Mediation and Interaction (Vanderweele 2015)

This text is the seminal book of Vanderweele, and serves as the most comprehensive text on mediation and interaction in causal inference. It is particularly relevant for individuals wishing to describe mechanisms and pathways between specific variables of interest. While not directly covered in this handbook due to it being beyond the scope of service and intervention evaluation, it's an excellent reference for those interested in incorporating effect decomposition and/or mediation analysis into their work.

Modern Epidemiology (Chapters 2-3) (Vanderweele 2015)

Chapters 2-3 of Lash et al.s' book provide great introductory perspective on causal inference from an epidemiological point of view, covering the Bradford Hill criteria, the potential outcomes framework, and DAGs. It is a particularly good resource for public health practitioners who are looking for a text that bridges the gap between traditional epidemiological criteria for causal inference and more modern causal inference frameworks.

7 Coverage of Causal Inference Methods in Key Texts

Coming Soon

Part II

Key Concepts in Causal Inference

8 Randomized Controlled Trials

Before considering specific methods for conducting causal inference, it is worth first considering how randomized experiments, commonly referred to as randomized controlled trials (RCTs), address many of the challenges and issues encountered in observational evaluations of interventions and services, and why it's considered to be the gold standard for causal inference.

To understand why this is, it's worth considering the common assumptions present in causal inference methodology covered in section 1.3, and how the fundamental design of RCTs overcome each of these.

8.1 Exchangeability

Consider once again the fundamental problem of causal inference: we cannot physically observe both counterfactual outcomes in real life for an individual, one under treatment and the other without. However, this specific difference is inherently what we're trying to measure, which Hernan and Robins illustrate well in the diagram presented below.

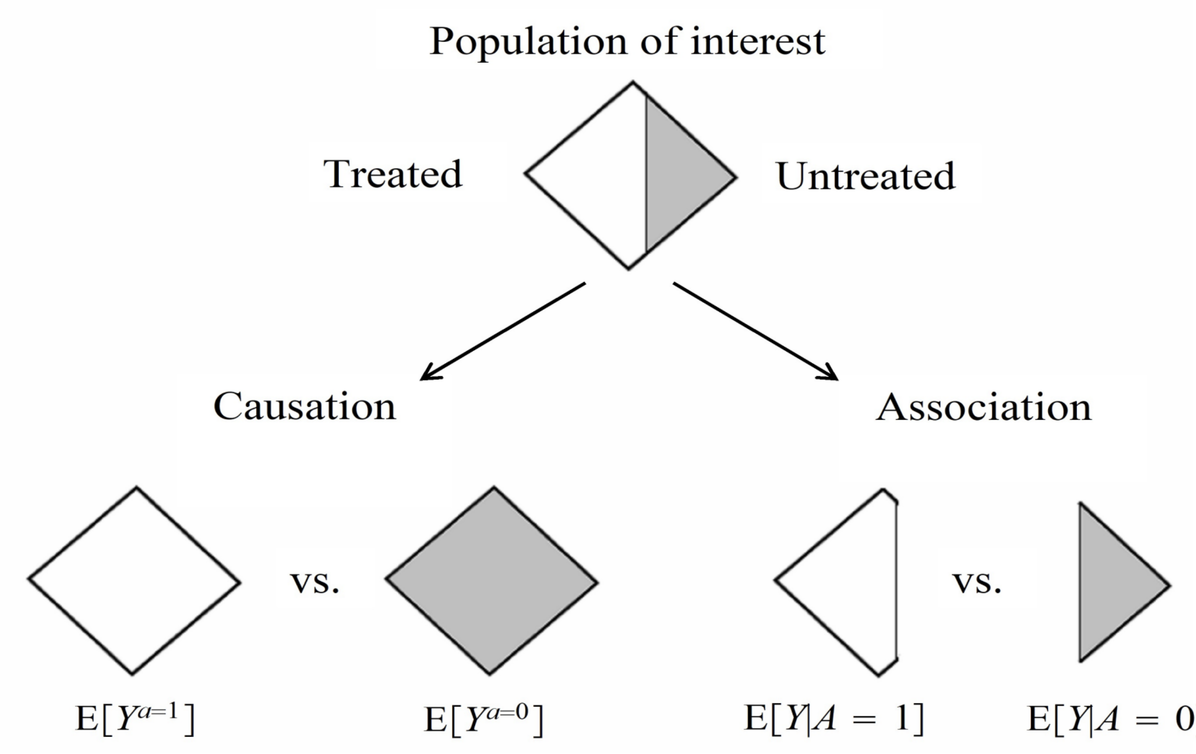


Figure 8.1: Defining Causation | p.12 of (Hernán and Robins, 2025)

As such, our best option is to establish two groups of individuals that share the same distributions of all relevant characteristics, ideally so close to the point where they are essentially clones of each other. The ideal way to do this is through randomizing study participants to each treatment group, as given a sufficient sample size, participant characteristics should in theory be equally distributed between both groups. Therefore, estimated average treatment effects should be the same regardless of which of the randomized groups actually receives treatment, hence becoming "exchangeable". Angrist and Pischke summarise this concept quite well:

"Random assignment works not by eliminating individual differences but rather by ensuring that the mix of individuals being compared is the same" (Angrist and Pischke 2015)

Of course, given random chance, there is still a possibility that participant characteristics do not end up being completely equally distributed between treatment groups. This can, however, be confirmed through descriptive statistics, and individual covariates can be adjusted for post-hoc as well through regression modelling to ensure exchangeability between treatment groups.

8.2 Consistency

In the protocol of an RCT written before patients are recruited, treatment strategies are clearly outlined within trial protocols. Almost all contemporary trials adhere to the CONSORT and SPIRIT checklists, first published in 1996 and 2013 respectively, which provide guidance on recommended items to address within trial protocols (Begg et al. 1996; Chan, Tetzlaff, Altman, et al. 2013; Chan, Tetzlaff, Gøtzsche, et al. 2013). In particular, the section on interventions in the SPIRIT checklist explicitly states to include information on:

- *"Interventions for each group with sufficient detail to allow replication, including how and when they will be administered."*
- *"Criteria for discontinuing or modifying allocation intervention for given trial participant (e.g., drug dose change in response to harms, participant request, or improving/worsening disease)."*
- *"Relevant concomitant care and interventions that are permitted or prohibited during the trial."* (Chan, Tetzlaff, Altman, et al. 2013)

By explicitly detailing the intervention strategy in advance of carrying out the trial, the treatment strategy each study participant receives is uniformly defined. In other words, this rules out ambiguity in treatment definition: for example, avoiding situations where one participant receives a drug daily while another receives it weekly, or where participants receive different doses of the same drug, despite both being labelled as "treated." Such variation would undermine the validity of the treatment contrast.

8.3 Positivity

Within the SPIRIT checklist, it is also advised that all trials report on the:

"Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centers and individuals who will perform the interventions (e.g., surgeons, psychotherapists)" (Chan, Tetzlaff, Altman, et al. 2013).

By explicitly defining what conditions participants would have to meet before becoming eligible for the trial, we ensure that all participants are physically capable of, and hence have a probability greater than zero to adhere to the treatment strategy of interest.

For example, in a trial evaluating the effect of anticoagulation therapies (e.g., warfarin or a DOAC) in preventing stroke in patients with atrial fibrillation, participants with a history of gastrointestinal bleeding or a known bleeding disorder (e.g., haemophilia) would typically be excluded. Including them would violate the positivity assumption, as it would be medically inadvisable, and therefore impossible to assign them to the treatment arm in the trial.

8.4 Stable Unit Treatment Value Assumption (SUTVA)

As a reminder, SUTVA has two distinct requirements:

1. No multiple versions of treatment

2. No interference between units

The first of these is already covered by meeting the consistency assumption. However, achieving the no interference requirement is not automatically guaranteed, and requires additional considerations in the design of a trial.

A common method of strengthening the no interference assumption is through blinding, and ideally double-blinding, where neither study participants or investigators are aware of the treatment assigned to each individual. Blinding can help reduce the chances of behavioural changes or spillover effects in study participants compared to if they were aware of the treatment they were assigned to. For example, A study participant who is aware of being on a novel therapy may choose to share their treatment with family members and friends who might be in the control group if they experience an improvement. However, blinding is not always feasible, especially in trials where a specific procedure or physical intervention is visibly administered.

Physically separating treatment groups may also be an option to address SUTVA, especially for social and public health interventions. Cluster-randomized designs, where entire clusters of individuals (e.g., entire households, schools, or communities) are randomized as a whole to receive an intervention or not. This approach allows the trial to account for interactions between individuals within the same cluster, preserving internal validity by focusing estimation on between-cluster differences in treatment effects.

8.5 Other Trial Designs and Further Reading

The picture I've painted here is quite simplified though. Aside from standard two-armed trials that I've mainly described in this section, many other more intricate designs exist to overcome issues present in such a simple design. I've included brief overviews of a selection of alternative designs that would be useful to consider, especially as many of these designs can also be emulated using observational data. Please note that this list is not exhaustive, and is presented more for inspiration on the types of trials you may choose to emulate in a causal inference study of observational data:

Core Designs

1. ***Parallel Group RCT***
   This is the conventional two-arm trial that is most commonly considered, where *"patients are randomized to the new treatment or to the standard treatment and followed-up to determine the effect of each treatment in parallel groups"* (Wang and Bakhai 2006).

   For further details, please refer to the following texts: (Wang and Bakhai 2006; Pocock 2013; Piantadosi and Meinert 2022)

2. ***Crossover Trials***
   "Crossover trials randomize patients to different sequences of treatments, but all patients eventually get all treatments in varying order, ie, the patient is his/her own control" (Wang and Bakhai 2006).

   For further details, please refer to the following texts: (Wang and Bakhai 2006; Pocock 2013; Piantadosi and Meinert 2022)

3. ***Factorial Trials***
   "Factorial trials assign patients to more than one treatment-comparison group. These are randomized in one trial at the same time, ie, while drug A is being tested against placebo, patients are re-randomized to drug B or placebo, making four possible treatment combinations in total" (Wang and Bakhai 2006).

   "A factorial structure is the only design that can assess treatment interactions, so this type of trial is required for those important therapeutic questions. When interactions between treatments are absent, which is not a trivial requirement, a factorial design can estimate each of several treatment effects from the same data. For example, two treatments can sometimes be evaluated using the same number of subjects ordinarily used to test a single therapy" (Piantadosi and Meinert 2022).

   For further details, please refer to the following texts: (Wang and Bakhai 2006; Pocock 2013; Piantadosi and Meinert 2022)

4. ***Equivalence (Clinical or Bioequivalence), Superiority, & Non-Inferiority Trials***
   "A [clinical] equivalence study is designed to prove that two drugs have the same clinical benefit. Hence, the trial should demonstrate that the effect of the new drug differs from the effect of the current treatment by a margin that is clinically unimportant."

   A bioequivalence study *"compares the pharmacokinetic (PK) parameters derived from plasma or blood concentrations of the compound,"* in order to understand whether *"the same number of drug compound molecules occupying the same number of receptors will have similar clinical effects."*

   "A superiority study aims to show that a new drug is more effective than the comparative treatment (placebo or current best treatment). Most clinical trials belong to this category."

   "A noninferiority study aims to show that the effect of a new treatment cannot be said to be significantly weaker than that of the current treatment" (Wang and Bakhai 2006).

   For further details, please refer to the following texts: (Wang and Bakhai 2006; Piantadosi and Meinert 2022)

5. ***Multicenter Trials***
   "A multicenter trial is a trial that is performed simultaneously at many centers following the same protocol." "[It] has several advantages over a single-center study, namely: it allows a large number of patients to be recruited in a shorter time; the results are more generalizable and contemporary to a broader population at large; and such studies are critical in trials involving patients with rare presentations or diseases" (Wang and Bakhai 2006).

   For further details, please refer to the following texts: (Wang and Bakhai 2006; Piantadosi and Meinert 2022)

6. ***Cluster Randomized Trials***

   "Cluster randomized trials are performed when larger groups (eg, patients of a single practitioner or hospital) are randomized instead of individual patients."* *"When individual randomization proves inappropriate, CRTs can be used to reduce the potential for contamination within treatment groups"* (Wang and Bakhai 2006).

   For further details, please refer to the following texts: (Wang and Bakhai 2006; Piantadosi and Meinert 2022)

Advanced Designs (A Non-Exhaustive List of Examples)

7. ***Adaptive Trials***

   "In adaptive trials, patient outcomes are observed and analysed at predefined interim points and predetermined modifications to study design can be implemented based on these observations" (Bothwell et al. 2018).

   For further details, please refer to the following texts: (Jennison and Turnbull 1999; Bothwell et al. 2018; Piantadosi and Meinert 2022). In particular, Bothwell et al.'s paper has very handy definitions of different types of adaptive designs within Table 1.

8. ***Stepped Wedge Cluster Randomized Trials***

   "A stepped wedge design is a type of crossover design in which different clusters cross over (switch treatments) at different time points. In addition, the clusters cross over in one direction only—typically, from control to intervention."

   "In a parallel or traditional crossover design, the intervention must be implemented in half of the total clusters simultaneously. However, limited resources or geographical constraints may make this logistically impossible. The stepped wedge design allows the researcher to implement the intervention in a smaller fraction of the clusters at each time point" (Hussey and Hughes 2007).

   For further details, please refer to the following texts: (Hussey and Hughes 2007; Hemming et al. 2015; Piantadosi and Meinert 2022).

9. ***Master Protocols: Umbrella Trials, Basket Trials, & Platform Trials***

   "Master protocols coordinate several closely linked investigations into a single trial, enabling efficient use of resources" (Piantadosi and Meinert 2022). *"Master protocols are often classified into "basket trials", "umbrella trials", and "platform trials""* (Park et al. 2019).

   "Umbrella trials select patients from a certain disease site (e.g., lung cancer), perform genetic testing on patients, and assign them to multiple treatments according to the matched drug targets."

   "Basket trials take patients from multiple disease sites but with a certain mutation (e.g., BRAF mutation) and assign them to the corresponding target therapy (e.g., BRAF inhibitors)."

   "Platform trials provide efficient screening of multiple treatments in a certain disease in which a steady flow of patients is available. A common control group such as the standard of care can be incorporated as the reference groups. New treatments can be added to the platform and evaluated. If a treatment is promising, it can "graduate" and, if a treatment is not promising,

it can be dropped from the platform. The trial can run perpetually to efficiently screen for effective treatments" (Piantadosi and Meinert 2022).

For further details, please refer to the following texts: (Park et al. 2019; Piantadosi and Meinert 2022).

10. ***Multi-arm Multi-stage Platform Trials***
    Multi-arm Multi-stage (MAMS) trials are a type of platform trial. *"The MAMS design aims to speed up the evaluation of new therapies and improve success rates in identifying effective ones." "In this framework, multiple experimental treatments are compared against a common control arm in several stages. This approach has several advantages over the more traditional designs since it obviates the need for multiple two-arm studies, and allows poorly performing experimental treatments to be discontinued during the study"* (Piantadosi and Meinert 2022).

    For further details, please refer to the following texts: (Royston et al. 2011; J. M. S. Wason and Jaki 2012; J. Wason et al. 2012; Piantadosi and Meinert 2022).

11. ***Sequential, Multiple Assignment, Randomized Trials (SMART)***
    "A SMART is a type of multi-stage, factorial, randomized trial, in which some or all participants are randomized at two or more decision points. Whether a patient is randomized at the second or a later decision point, and the available treatment options, may depend on the patient's response to prior treatment" (Kidwell and Almirall 2023).

    For further details, please refer to the following texts: (Piantadosi and Meinert 2022; Kidwell and Almirall 2023).

9 Directed Acyclic Graphs

9.1 Graphically Representing Causal Relationships

Unlike randomized experiments, when conducting observational studies, we are fundamentally unable to randomize individuals into distinct treatment and control groups, as data was collected retrospectively. Therefore, we have to consider ways to emulate randomized treatment assignment, of which there are various options depending on the evaluation question of interest.

Before attempting to emulate randomized treatment assignment though, it is strong recommended, especially in the field of epidemiology, to first map out the assumed causal network that links together your treatments, outcomes, as well as any other variables of interest. This process allows us to clarify which confounding variables need to be adjusted for, which can be safely ignored, and which could introduce bias if handled improperly. We can thereby be explicit about the assumptions underlying any statistical models we develop afterwards. The most popular tool for this purpose is the Directed Acyclic Graph, often shortened as DAG.

While graphical representations of causal relationships have a long history, with DAGs originating from the field of mathematics, specifically graph theory, it wasn't until one of Pearl's landmark papers "Causal Diagrams for Empirical Research" where the idea of incorporating causal diagrams in empirical research was first consolidated (Pearl 1995). Greenland, Pearl and Robins' similarly notable paper "Causal Diagrams for Epidemiologic Research" published a few years later further extended and popularized these ideas to the field of epidemiology, establishing how DAGs can be used for confounder selection and identifying when incorrectly adjusting for certain variables, known as colliders, may induce biased results (Greenland, Pearl, and Robins 1999).

9.2 An Introduction to DAGs

DAGs are structured with nodes indicating individual variables of interest, connected by one way arrows (referred to as directed edges in mathematics). Unique properties of DAGs are found in their name. They are "directed" given that arrows within them only point in one direction, and "acyclic" due to variables measured at a single time point cannot influence themselves in a feedback loop. If the same variable is measured at multiple time points (e.g., BMI at baseline, at 6 months, at 12 months…), these can be treated as separate nodes to reflect temporal ordering.

Nodes are often described using terminology for a family tree. A variable that causally influences another along a directed path is referred to as an ancestor (or a parent if the influence is direct). Variables that lie downstream (i.e., those that are affected by the ancestor) are called descendants (or children, if directly influenced). These relationships help define how information and causal effects can flow through the DAG.

9.3 Blocking Causal Pathways: Frontdoor and Backdoor Adjustment

Pearl's framework for causal diagrams revolves around the idea of blocking or closing paths through which spurious associations might flow, and is described in great detail in their book "Causality" (Pearl 2009). These associations may result from random chance or from unmeasured variables. When estimating causal effects, the goal is to block all non-causal pathways between the treatment and the outcome to eliminate these biases. These pathways fall into two main categories: backdoor and frontdoor paths.

9.4 D-Separation

Before going over frontdoor and backdoor paths in detail, it's worth understanding how paths can be considered open or blocked in a DAG. This is governed by a set of criteria called D-separation, where "D" stands for directional. D-separation provides formal rules for deciding whether one set of variables is statistically independent of another, once we condition on a third set. In other words, it tells us when information can or cannot flow through a path in the graph.

Hernán provides the following succinct definitions of D-separation rules in his free online course about causal diagrams on HarvardX (M. Hernán 2017).

1. *"If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable on the path."*
    - The path between A and Y here is blocked by collider L:
      A → L ← Y [Blocked]
2. *"Any path that contains a noncollider that has been conditioned on is blocked."*
    - The path between A to Y through B is open here:
      A → B → Y [Open]
    - However, if we condition on / adjust for B, signified by drawing a box around B, we block the path between A to Y:
      A → B → Y [Blocked]
3. *"A collider that has been conditioned on does not block a path."*
    - Previously our path between A and Y here was blocked by collider L:
      A → L ← Y [Blocked]

- However, if we condition on / adjust for L, signified by drawing a box around L, we open up the path between A to Y:
  A → L ← Y [Open]

4. "A collider that has a descendant that has been condition on does not block a path."
   - Previously our path between A and Y here was blocked by collider L:
     A → L ← Y [Blocked]
   - However, if we condition on / adjust for a descendant of L, S, signified by drawing a box around S, we open up the path between A to Y:
     A → L ← Y AND L → S,   i.e., A → S AND Y → L → S [Open]

(M. Hernán 2017)

With these rules in mind, we can now continue to how backdoor and frontdoor paths use D-separation to determine which variables to adjust for in our analyses.

9.5 Backdoor Paths

A backdoor path is any path from the treatment to the outcome that starts with an arrow pointing into the treatment. These paths generally reflect confounding and can bias the estimate of the treatment effect if not properly addressed. Confounding is a situation where additional variables, known as confounders, distort the causal relationship between the treatment and outcome of interest.

Pearl's backdoor criterion specifies that we can identify a causal effect of a treatment [A] on an outcome [Y] by adjusting for a set of covariates [Z] if:

1. No variables in the set of covariates [Z] are a descendant of the treatment [A].
   - i.e., they cannot lie on the causal pathway between the
     Treatment [A] → Outcome [Y]
   - Adjusting for a descendant of A (referred to as a mediator) would block part of the causal effect itself, leading to an underestimate of the total effect of treatment [A] on outcome [Y].
2. The set of covariates [Z] blocks every path between the treatment [A] and outcome [Y] that starts with an arrow pointing into the treatment [A] (i.e., all backdoor paths).

When these criteria is met, we can estimate the causal effect by conditioning on the set of covariates [Z], which restores the conditional exchangeability assumption discussed in previous chapters. This is conceptually equivalent to the idea that, once we account for the set of covariates [Z], the assignment of treatment [A] can be considered as good as randomised (i.e., participant characteristics are equally distributed between the treatment groups).

Pearl represented the backdoor criterion in a diagram in his book "Causality" attached below. Adjusting for $X_3 + X_4$ or $X_4 + X_5$ (all of which would be considered confounders in this scenario) would correctly adjust to all backdoor paths, as the remaining variables that are not descendants of Xi would have all their paths connecting X_i and X_j blocked.

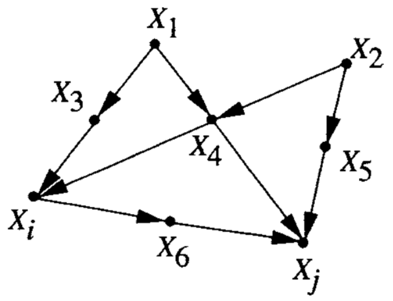


Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 9.1: A diagram representing the backdoor criterion | p.80 (Pearl, 2009)

For example, if $X_3 + X_4$ were adjusted for:

1. X_1's path to X_i and X_j would be blocked by X_3 and X_4
2. X_2 would only have a path to X_j through X_5, but not to X_i as this is blocked by X_4
3. X_5 only has a path to X_j and not X_i

If only X4 was adjusted for, a new path between Xi and Xj would be created:
$\mathbf{X_i \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \leftarrow X_2 \rightarrow X_5 \rightarrow X_j}$

This is due to rule #3 of D-separation defined by Hernán, as conditioning on a blocked pathway opens up the path.

If only X_6 was adjusted for, we would violate Pearl's criterion, as the variable is a descendant of X_i (i.e., a mediator between X_i and X_j). This would block part of the causal effect between X_i and X_j, biasing results.

A Practical Example

To provide a practical example of backdoor adjustment, imagine a situation where we wish to evaluate the impact of a community exercise programme on reducing cardiovascular-related hospitalisations:

- *Age and underlying health conditions may affect both the likelihood of participating in the programme and the risk of hospitalisation. These are confounders lying on backdoor paths, which must be blocked by adjusting for them in our analysis to obtain an unbiased estimate of the exercise programme's effect on cardiovascular-related hospitalisations.*

However, there are some variables which should not be adjusted for:

1. ***Incorrectly adjusting for a collider***
   Suppose individuals are more likely to receive the programme if they are either highly motivated or referred by their GP due to elevated blood pressure. If we adjust for referral status, which we know to be collider through the following path:
   Exercise ← Motivation → Referral ← Blood Pressure → Hospitalisation

If we adjust for referral in a regression model, we're comparing people with the same referral status. But since referral happens for different reasons, this creates a spurious and false path between motivation and blood pressure.

For example, among referred individuals, those who are more motivated are more likely to have lower blood pressure and be healthier overall, thereby inherently having a lower risk of hospitalisation, and vice versa. This leads to selection bias, as the people being compared may already differ in their risk of hospitalisation, even before the programme. As a result, the programme's effect may be overestimated, not because it's more effective, but because adjusting for referral has distorted the comparison.

2. ***Incorrectly adjusting for a descendant of the treatment (a mediator)***
   Suppose we also have access to data on post-programme weight loss. Weight loss lies on the causal pathway between the treatment (participation in the exercise programme) to the outcome (hospitalisations).

   Adjusting for this mediator (a descendant of the treatment) would block part of the causal effect we are trying to estimate, resulting in an underestimate of the total effect of the programme. This violates the backdoor criterion, which explicitly states that descendants of the treatment should not be included in the adjustment set when estimating the total causal effect.

Therefore, DAGs are not only useful in helping identify confounders to adjust for, but also in helping avoid adjustments for variables that would induce bias, either through introducing spurious associations by opening backdoor paths, or through incorrectly adjusting for a mediator.

9.6 Frontdoor Paths

A frontdoor path is a path where the causal effect flows from the treatment [A] to a mediator or set of mediators [M], which then influences the outcome [Y]. Under certain assumptions, frontdoor adjustment can be used when backdoor paths cannot be directly blocked.

To use frontdoor adjustment for estimating a causal effect, we must satisfy three criteria:

1. The mediator(s) [M] fully intercepts all directed paths from the treatment [A] to outcome [Y]
   - i.e., there are no direct paths from the treatment [A] to the outcome [Y] that bypasses the mediator(s) [M]
     Treatment [A] → Mediator(s) [M] → Outcome [Y]
2. There are no unblocked backdoor paths from the treatment [A] to the mediator(s) [M]
   - i.e., there must be no unmeasured confounders between the treatment [A] and the mediator(s) [M]
3. All backdoor paths from mediator(s) [M] to the outcome [Y] are blocked by conditioning on the treatment [A]

- i.e., there must be no unmeasured confounders between the mediator(s) [M] and the outcome [Y]

Pearl similarly represented the frontdoor criterion in another diagram in his book "Causality" attached below, outlining how it can be used to estimate a causal effect.

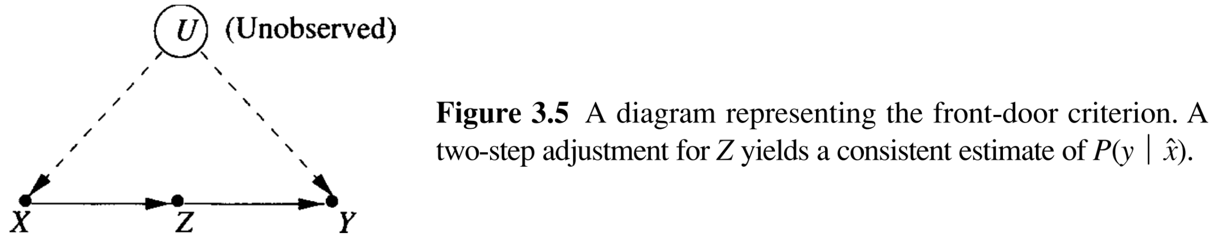


Figure 3.5 A diagram representing the front-door criterion. A two-step adjustment for *Z* yields a consistent estimate of $P(y \mid \hat{x})$.

Figure 9.2: A diagram representing the frontdoor criterion | p.81 (Pearl, 2009)

In short, the process described in the diagram can be performed as follows:

1. Model the relationship between the treatment [A] and the mediator(s) [M]:
   - Estimate the distribution of the mediator(s) [M] given treatment [A]
   - e.g., fit a regression model where mediator(s) [M] is the outcome and treatment [A] is the predictor
2. Model the relationship between the mediator(s) [M] and the outcome [Y], adjusting for the treatment:
   - Estimate the conditional distribution of the outcome [Y] given mediator(s) [M] and treatment [A]
   - i.e., fit a regression model regressing the outcome [Y] on both mediator(s) [M] and treatment [A]
3. Combine the two models to simulate the total causal effect of the treatment [A] on the outcome [Y]:
   - Use the first model to predict the distribution of the mediator(s) [M] under different values of treatment [A]
   - Then use the second model to estimate the expected outcome [Y] for each predicted value of the mediator(s) [M]
   - Finally, average these predictions to estimate the causal effect of the treatment [A] on outcome [Y]

In essence, this two-step procedure reconstructs the causal effect by tracing how the treatment [A] affects the mediator(s) [M], and how the mediator(s) [M] affects the outcome [Y], bypassing the backdoor pathways (i.e., confounding) that may exist between the treatment [A] and outcome [Y] directly.

Do note, however, that in the field of epidemiology, backdoor adjustment is typically preferred over frontdoor adjustment, as it is often unrealistic to assume that all mediators lying between the treatment and the outcome have been correctly identified and measured. This assumption is essential for valid frontdoor adjustment, but rarely holds when answering causal epidemiological questions using real world data.

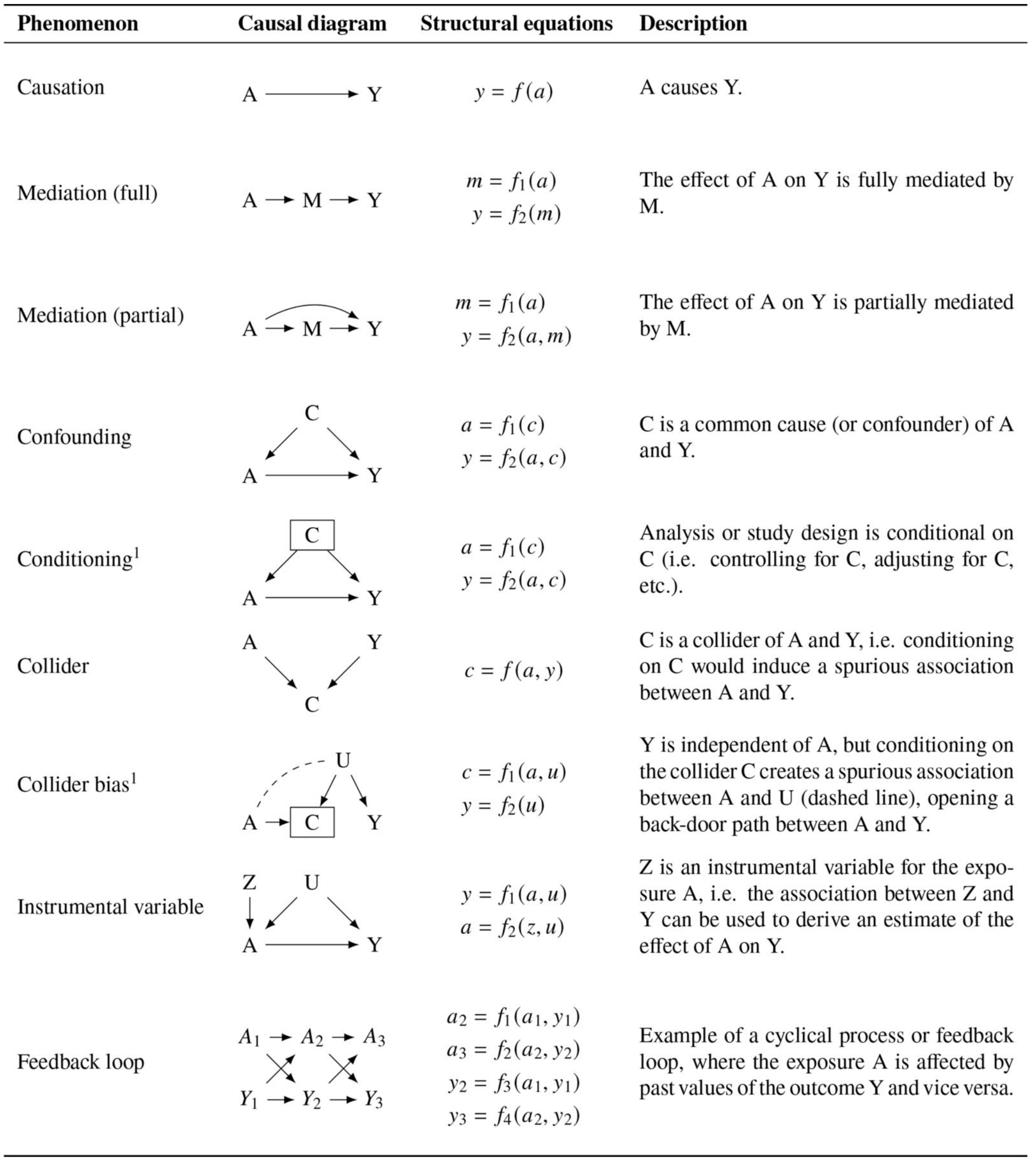
A Practical Example

To provide a practical example of frontdoor adjustment, imagine a situation where we wish to evaluate the effect of an alcohol awareness focused public health campaign on liver disease hospitalisations:

- Say we do not have a measure individual alcohol consumption, which is a major confounder. However, we do observe whether individuals attended alcohol counselling services, which is a mediator between exposure to the public health campaign and liver disease hospitalisation.
- If counselling attendance causally affects liver disease risk, and there are no unmeasured confounders on the paths mentioned above, a frontdoor approach could recover the causal effect.
- Do note, however, that this approach relies on the strong assumption that alcohol counselling is the only mediator of the campaign's effect on liver disease. Such an assumption is often implausible in real world settings, where multiple other mechanisms may also mediate the relationship.

9.7 Common Causal Phenomena Illustrated by DAGs

To further consolidate some of the causal phenomena mentioned above which can be illustrated in DAGs, Igelström et al. provide a helpful visual summary of these in their 2022 paper where they present a table of key concepts with visual and text descriptions attached to each (Igelström et al. 2022). This is a great point of reference to refer back to while considering whether the use of methods covered in later chapters are appropriate for evaluating your key evaluation question of interest.



Phenomenon	Causal diagram	Structural equations	Description
Causation	A → Y	$y = f(a)$	A causes Y.
Mediation (full)	A → M → Y	$m = f_1(a)$ $y = f_2(m)$	The effect of A on Y is fully mediated by M.
Mediation (partial)	A → M → Y	$m = f_1(a)$ $y = f_2(a, m)$	The effect of A on Y is partially mediated by M.
Confounding	C A → Y	$a = f_1(c)$ $y = f_2(a, c)$	C is a common cause (or confounder) of A and Y.
Conditioning[1]	C A → Y	$a = f_1(c)$ $y = f_2(a, c)$	Analysis or study design is conditional on C (i.e. controlling for C, adjusting for C, etc.).
Collider	A Y C	$c = f(a, y)$	C is a collider of A and Y, i.e. conditioning on C would induce a spurious association between A and Y.
Collider bias[1]	U A → C Y	$c = f_1(a, u)$ $y = f_2(u)$	Y is independent of A, but conditioning on the collider C creates a spurious association between A and U (dashed line), opening a back-door path between A and Y.
Instrumental variable	Z U A → Y	$y = f_1(a, u)$ $a = f_2(z, u)$	Z is an instrumental variable for the exposure A, i.e. the association between Z and Y can be used to derive an estimate of the effect of A on Y.
Feedback loop	A_1 → A_2 → A_3 Y_1 → Y_2 → Y_3	$a_2 = f_1(a_1, y_1)$ $a_3 = f_2(a_2, y_2)$ $y_2 = f_3(a_1, y_1)$ $y_3 = f_4(a_2, y_2)$	Example of a cyclical process or feedback loop, where the exposure A is affected by past values of the outcome Y and vice versa.

Figure 9.3: Common causal phenomena represented in DAGs (Igelström et al., 2022)

9.8 Additional Resources

DAGitty is an excellent tool for making your own DAGs, with both an online tool that you can use in your browser, as well as an R package named “dagitty.” The tool also, very usefully, provides information on the minimal adjustment set of variables needed to block all backdoor pathways and estimate an unbiased causal effect after designing a DAG. Both tools can be access through their website.

If you are interested in getting a more comprehensive overview of DAGs, I would highly recommend auditing the free online HarvardX course "Causal Diagrams: Draw Your Assumptions Before Your Conclusions" taught by Hernan as well.

9.9 Statistically Testing the Validity of DAGs

Further details coming soon

DAGs can be tested statistically against both continuous and categorical data that you already have within your dataset (Ankan, Wortel, and Textor 2021)…

Note that although this can help verify the validity of a DAG based on existing data, these methods are unable to help identify whether there are additional variables that are not included in the DAG but should have been present.

9.10 Alternative Methods for Covariate Selection

Further details coming soon

There are alternatively automated ways that are used to select variables of interest in situations where you have high dimensional data (e.g., 90k+ vars), this is where LASSO regression might be useful.

However, this comes at the expense of potentially incorrectly adjusting for specific covariates in your models (Lau et al. 2022)…

10 Target Trial Emulation

Coming Soon

11 Defining Treatment Strategies

Coming Soon

Part III

Quasi-Experimental Methods

Coming Soon

12 Difference-in-Differences

Coming Soon

13 Interrupted Time Series

Coming Soon

13.1 Standard Interrupted Time Series Analysis

Coming Soon

13.2 Controlled Interrupted Time Series Analysis

Coming Soon

14 Synthetic Controls

Coming Soon

15 Regression Discontinuity

Coming Soon

16 Instrumental Variables

Coming Soon

17 Extensions of Quasi-Experimental Designs

Coming Soon

17.1 Negative Controls

Coming Soon

17.2 Sensitivity Analysis

Coming Soon

Part IV

Adjustment-Based Methods

Coming Soon

18 Outcome Regression (Selection on Observables)

Coming Soon

19 Matching (Exact & Covariate)

Coming Soon

20 Propensity Scores

Coming Soon

20.1 Estimating Propensity Scores

Coming Soon

20.2 Propensity Score Matching

Coming Soon

21 G-Methods: An Introduction

Coming Soon

22 G-Methods 1: Inverse Probability Weighting

Coming Soon

23 G-Methods 2: Parametric G-Formula

23.1 Parametric G-Formula (Non-Iterative Expectation)

Coming Soon

23.2 Iterative Conditional Expectation (ICE) G-Formula

Coming Soon

24 Extensions of Adjustment-Based Methods

Coming Soon

Part V

Applying Causal Methods: Worked Examples in the NHS

Coming Soon

25 Case Study 1a: DOAC Head-to-Head Emulated Trial

Coming Soon

26 Case Study 1b: DOAC Policy Prescription Change Interrupted Time Series Analyses

Coming Soon

27 # Case Study 2: Ophthalmology Hub of Care Model Interrupted Time Series Analyses

Coming Soon

28 Practical Barriers to Implementation: Common Challenges in Causal Inference

Coming Soon

29 Conclusion

Coming Soon

30 About the Author & Acknowledgements

30.1 About Me

Hi, my name is Jamie and I'm an MRC-funded PhD student at UCL, focusing on the application of causal inference methods within Epidemiology, Data Science, and Health Economics.

Before my PhD, I previously graduated with a BSc in Population Health Sciences (Data Science) from UCL, and an NIHR funded MSc in Epidemiology from the London School of Hygiene and Tropical Medicine.

Prior to my PhD, my interest in researching the application of causal inference methods in observational data stemmed from my varied experience working across multiple areas of public health research, including supporting the delivery of cancer trials, evaluating obesity policy decisions, identifying barriers to STI testing, and investigating the long-term effects of cancer survivorship.

Having worked with both randomized trials and observational data, focusing my research on causal inference in real-world data felt like a natural next step.

If you have any questions do feel free to reach out to me via the following channels: UCL Email: jamie.wong [at] ucl.ac.uk NHS Email: laichunjamie.wong [at] nhs.net LinkedIn: https://www.linkedin.com/in/jamiewlc/ GitHub: https://github.com/jamiewlc

30.2 Acknowledgements

This handbook was developed during my PhD placement with NHS England and the North Central London Integrated Care Board (NCL ICB) as part of the NHS England Data Science PhD Internship Programme. I am grateful to both organisations for providing the opportunity, data access, and collaborative environment that enabled this work.

I would like to express my sincere thanks to my internship supervisors, Jonathan Pearson (NHS England) and Emily Baldwin (NCL ICB), whose guidance, support, and encouragement were invaluable throughout the project.

I am also deeply grateful to my academic supervisors at UCL, Dr Michalis Katsoulis, Prof Manuel Gomes, and Dr Sophie Eastwood, for their continued support in shaping the causal analyses and the writing of this handbook.

I would also like to thank colleagues across the NHS England Applied Data Science and AI team and the NCL ICB Analytics team for their advice and feedback during the development of this work, especially Jo Lam and Adam Hollings at NHS England, as well as Tom Shallcross, Felix Peckitt, Anita Garrick, and Enrico Panizzo at the NCL ICB.

References

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. "Mastering 'Metrics: The Path from Cause to Effect." 2015. https://press.princeton.edu/books/paperback/9780691152844/mastering-metrics.

Ankan, Ankur, Inge M. N. Wortel, and Johannes Textor. 2021. "Testing Graphical Causal Models Using the R Package 'Dagitty'." *Current Protocols* 1 (2): e45. https://doi.org/10.1002/cpz1.45.

Begg, Colin, Mildred Cho, Susan Eastwood, Richard Horton, David Moher, Ingram Olkin, Roy Pitkin, et al. 1996. "Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement." *JAMA* 276 (8): 637–39. https://doi.org/10.1001/jama.1996.03540080059030.

Bothwell, Laura E., Jerry Avorn, Nazleen F. Khan, and Aaron S. Kesselheim. 2018. "Adaptive Design Clinical Trials: A Review of the Literature and ClinicalTrials.gov." *BMJ Open* 8 (2): e018320. https://doi.org/10.1136/bmjopen-2017-018320.

Chan, An-Wen, Jennifer M. Tetzlaff, Douglas G. Altman, Andreas Laupacis, Peter C. Gøtzsche, Karmela Krleža-Jerić, Asbjørn Hróbjartsson, et al. 2013. "SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials." *Annals of Internal Medicine* 158 (3): 200–207. https://doi.org/10.7326/0003-4819-158-3-201302050-00583.

Chan, An-Wen, Jennifer M. Tetzlaff, Peter C. Gøtzsche, Douglas G. Altman, Howard Mann, Jesse A. Berlin, Kay Dickersin, et al. 2013. "SPIRIT 2013 Explanation and Elaboration: Guidance for Protocols of Clinical Trials." *BMJ* 346 (January): e7586. https://doi.org/10.1136/bmj.e7586.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press. https://doi.org/10.2307/j.ctv1c29t27.

———. 2025. "Mixtape Sessions (Github)." 2025. https://github.com/Mixtape-Sessions.

Digitale, Jean C., Jeffrey N. Martin, and Medellena Maria Glymour. 2022. "Tutorial on Directed Acyclic Graphs." *Journal of Clinical Epidemiology* 142 (February): 264–67. https://doi.org/10.1016/j.jclinepi.2021.08.001.

Fisher, R A, and F Yates. 1990. *Statistical Methods, Experimental Design, and Scientific Inference: A Re-issue of Statistical Methods for Research Workers, the Design of Experiments and Statistical Methods and Scientific Inference*. Edited by J H Bennett. Oxford University Press. https://doi.org/10.1093/oso/9780198522294.001.0001.

Fisher, Ronald A. 1925. "Statistical Methods for Research Workers (Scan of 1st Edition)." Classics in the History of Psychology. 1925. https://psychclassics.yorku.ca/Fisher/Methods/.

Greenland, Sander, Judea Pearl, and James M. Robins. 1999. "Causal Diagrams for Epidemiologic Research." *Epidemiology* 10 (1): 37. https://journals.lww.com/epidem/abstract/1999/01000/causal_diagrams_for_epidemiologic_research.8.aspx?casa_token=t1bYrLdrL5EAAAAA:jek7bcCsZNWdAWZSBuH4WUKpadC7BXnvKae9-OTeDK57f6KworoLmwgGS6PNX3AQRq8JEB4WL12pwWt-l9cMm5g_JwByHr78.

Hemming, K., T. P. Haines, P. J. Chilton, A. J. Girling, and R. J. Lilford. 2015. "The Stepped Wedge Cluster Randomised Trial: Rationale, Design, Analysis, and Reporting." *BMJ* 350

(February): h391. https://doi.org/10.1136/bmj.h391.

Hernán, Miguel. 2017. "Causal Diagrams: Draw Your Assumptions Before Your Conclusions (Online Course)." Harvard University. August 16, 2017. https://pll.harvard.edu/course/causal-diagrams-draw-your-assumptions-your-conclusions.

Hernán, Miguel A., and James M. Robins. 2016. "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available." *American Journal of Epidemiology* 183 (8): 758–64. https://doi.org/10.1093/aje/kwv254.

———. 2025. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC. https://miguelhernan.org/whatifbook.

Hill, Austin Bradford. 1965. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58 (5): 295–300. https://doi.org/10.1177/003591576505800503.

HM Treasury. 2020. *Magenta Book: Central Government Guidance on Evaluation*. United Kingdom: HM Treasury. https://assets.publishing.service.gov.uk/media/5e96cab9d3bf7f412b2264b1/HMT_Magenta_Book.pdf.

Holland, Paul W., Clark Glymour, and Clive Granger. 1985. "Statistics and Causal Inference." *ETS Research Report Series* 1985 (2): i–72. https://doi.org/10.1002/j.2330-8516.1985.tb00125.x.

Hussey, Michael A., and James P. Hughes. 2007. "Design and Analysis of Stepped Wedge Cluster Randomized Trials." *Contemporary Clinical Trials* 28 (2): 182–91. https://doi.org/10.1016/j.cct.2006.05.007.

Igelström, Erik, Peter Craig, Jim Lewsey, John Lynch, Anna Pearce, and Srinivasa Vittal Katikireddi. 2022. "Causal Inference and Effect Estimation Using Observational Data." *J Epidemiol Community Health* 76 (11): 960–66. https://doi.org/10.1136/jech-2022-219267.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79. https://doi.org/10.1257/jel.20191597.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139025751.

Jennison, Christopher, and Bruce W. Turnbull. 1999. *Group Sequential Methods with Applications to Clinical Trials*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9780367805326.

Kidwell, Kelley M., and Daniel Almirall. 2023. "Sequential, Multiple Assignment, Randomized Trial Designs." *JAMA* 329 (4): 336–37. https://doi.org/10.1001/jama.2022.24324.

Lau, Wallis C. Y., Carmen Olga Torre, Kenneth K. C. Man, Henry Morgan Stewart, Sarah Seager, Mui Van Zandt, Christian Reich, et al. 2022. "Comparative Effectiveness and Safety Between Apixaban, Dabigatran, Edoxaban, and Rivaroxaban Among Patients With Atrial Fibrillation." *Annals of Internal Medicine* 175 (11): 1515–24. https://doi.org/10.7326/M22-0511.

Markozannes, Georgios, Georgia Vourli, and Evangelia Ntzani. 2021. "A Survey of Methodologies on Causal Inference Methods in Meta-Analyses of Randomized Controlled Trials." *Systematic Reviews* 10 (1): 170. https://doi.org/10.1186/s13643-021-01726-1.

Mill, John Stuart. 1843. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. Vol. 1. Cambridge Library Collection - Philosophy. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139149839.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Analytical Methods for Social Research. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107587991.

Neal, Brady. 2019. "Which Causal Inference Book You Should Read." November 23, 2019. https://www.bradyneal.com/which-causal-inference-book.

NICE. 2012. "Appendix E Algorithm for Classifying Quantitative (Experimental and Observational) Study Designs | Methods for the Development of NICE Public Health Guidance (Third Edition)." NICE. September 26, 2012. https://www.nice.org.uk/process/pmg4/chapter/appendix-e-algorithm-for-classifying-quantitative-experimental-and-observational-study-designs.

OHID. 2021a. "Guidance: Choose Evaluation Methods: Evaluating Digital Health Products." GOV.UK. May 19, 2021. https://www.gov.uk/guidance/choose-evaluation-methods-evaluating-digital-health-products.

———. 2021b. "Guidance: Quasi-experimental Study: Comparative Studies." GOV.UK. September 8, 2021. https://www.gov.uk/guidance/quasi-experimental-study-comparative-studies.

Park, Jay J. H., Ellie Siden, Michael J. Zoratti, Louis Dron, Ofir Harari, Joel Singer, Richard T. Lester, Kristian Thorlund, and Edward J. Mills. 2019. "Systematic Review of Basket Trials, Umbrella Trials, and Platform Trials: A Landscape Analysis of Master Protocols." *Trials* 20 (1): 572. https://doi.org/10.1186/s13063-019-3664-1.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82 (4): 669–88. https://doi.org/10.1093/biomet/82.4.669.

———. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley. https://www.wiley.com/en-us/Causal+Inference+in+Statistics%3A+A+Primer-p-9781119186861.

Pearl, Judea, and Dana Mackenzie. 2018. "The Book of Why: The New Science of Cause and Effect." Guide books. 2018. https://dl.acm.org/doi/book/10.5555/3238230.

Piantadosi, Steven, and Curtis L. Meinert, eds. 2022. *Principles and Practice of Clinical Trials*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-52636-2_196.

Pocock, Stuart J. 2013. *Clinical Trials: A Practical Approach*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118793916.fmatter.

Rosenbaum, Paul R. 2017. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press. https://www.jstor.org/stable/j.ctv253f7k2.

———. 2020. *Design of Observational Studies*. Springer Series in Statistics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46405-9.

———. 2023. *Causal Inference*. The MIT Press. https://doi.org/10.7551/mitpress/14244.001.0001.

Royston, Patrick, Friederike M-S Barthel, Mahesh KB Parmar, Babak Choodari-Oskooei, and Valerie Isham. 2011. "Designs for Clinical Trials with Time-to-Event Outcomes Based on Stopping Guidelines for Lack of Benefit." *Trials* 12 (1): 81. https://doi.org/10.1186/1745-6215-12-81.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350.

———. 1986. "Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81 (396): 961–62. https://doi.org/10.2307/2289065.

Schiff, A, L Wright, and T Denne. 2017. "Ex-Post Evaluation of Transport Interventions Using Causal Inference Methods." NZ Transport Agency Waka Kotahi. https://nzta.govt.nz/resources/research/reports/630/.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-*

Experimental Designs for Generalized Causal Inference. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA, US: Houghton, Mifflin and Company.

Splawa-Neyman, Jerzy, D. M. Dabrowska, and T. P. Speed. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4): 465–72. https://doi.org/10.1214/ss/1177012031.

Van Der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-9782-1.

———. 2018. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Series in Statistics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65304-4.

Vanderweele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Explanation in Causal Inference: Methods for Mediation and Interaction. New York, NY, US: Oxford University Press.

Vigen, Tyler. 2024. "Tyler Vigen Spurious Correlation #1,172: Butter Consumption Correlates with Economic Output of Washington Metro Area." TylerVigen.com. January 2024. https://tylervigen.com/spurious/correlation/1172_butter-consumption_correlates-with_economic-output-of-washington-metro-area.

Wang, Duolao, and Ameet Bakhai. 2006. *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*. Remedica.

Wason, James M. S., and Thomas Jaki. 2012. "Optimal Design of Multi-Arm Multi-Stage Trials." *Statistics in Medicine* 31 (30): 4269–79. https://doi.org/10.1002/sim.5513.

Wason, James, Dominic Magirr, Martin Law, and Thomas Jaki. 2012. "Some Recommendations for Multi-Arm Multi-Stage Trials." *Statistical Methods in Medical Research*, December. https://doi.org/10.1177/0962280212465498.