Causal inference in four British cohort studies

Richard Silverwood, Georgia Tomova, Liam Wright

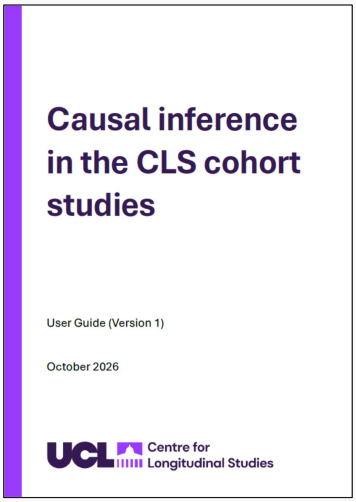
7 October 2026

UCL Centre for Longitudinal Studies

UKRI Economic and Social Research Council

1

Causal inference in the CLS cohort studies User Guide



www.cls.ucl.ac.uk/causal-inference-guide

UCL Centre for Longitudinal Studies

2

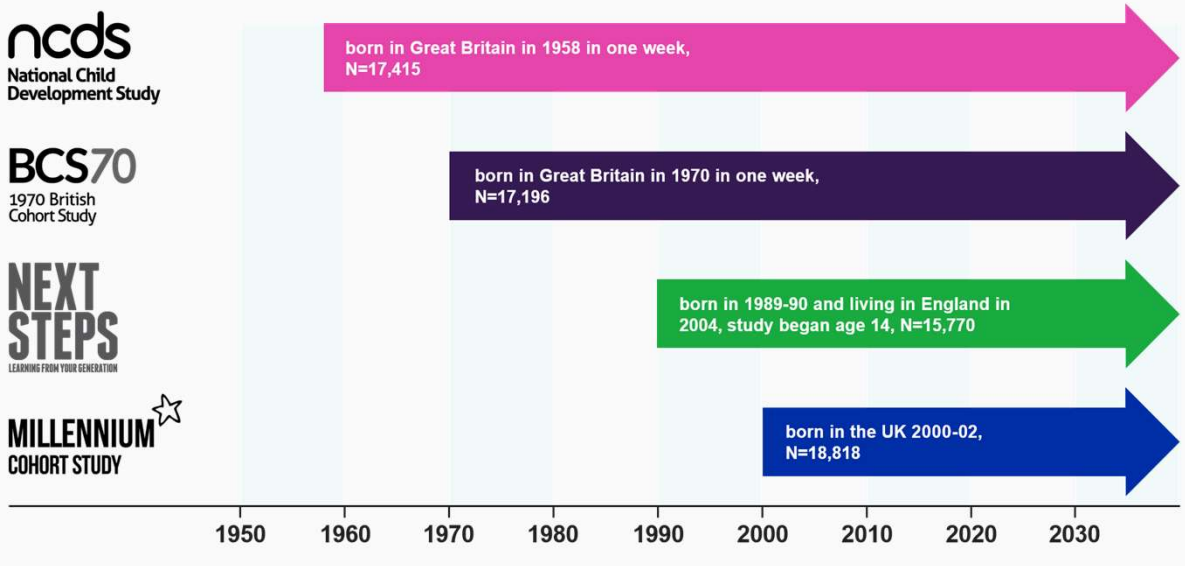
Contents of the user guide

1. Introduction 7
2. The problem of causal inference with observational data 10
3. Three approaches to causal inference 37
4. Attributes of CLS's cohort studies that support causal inference 44
5. The methods of causal inference 49
6. Discussion 182
7. Statements 185
8. Appendix 186
9. References 192

www.cls.ucl.ac.uk/causal-inference-guide

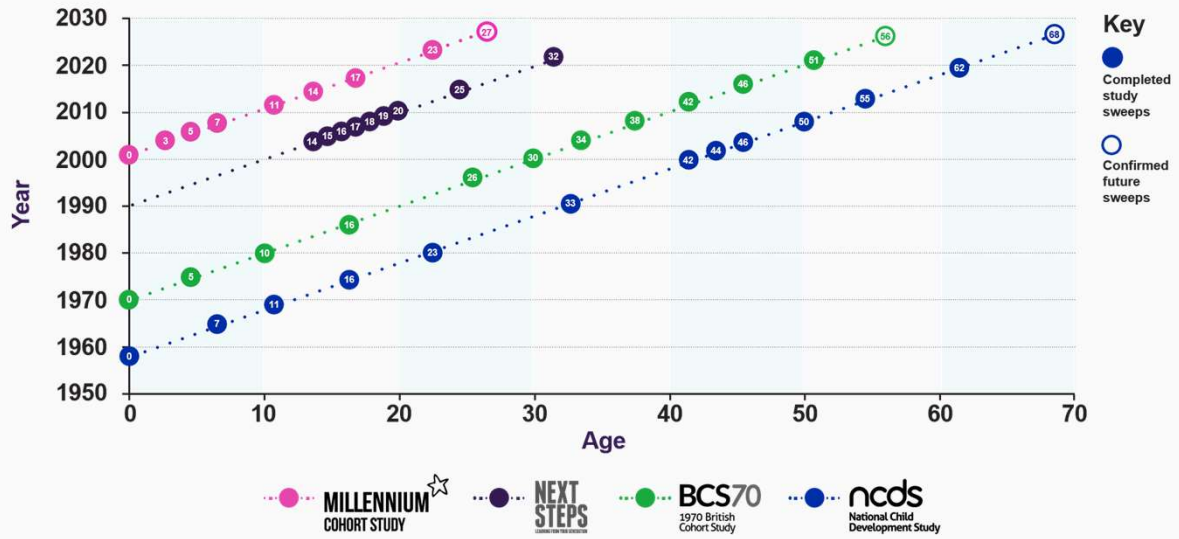
3

CLS core studies



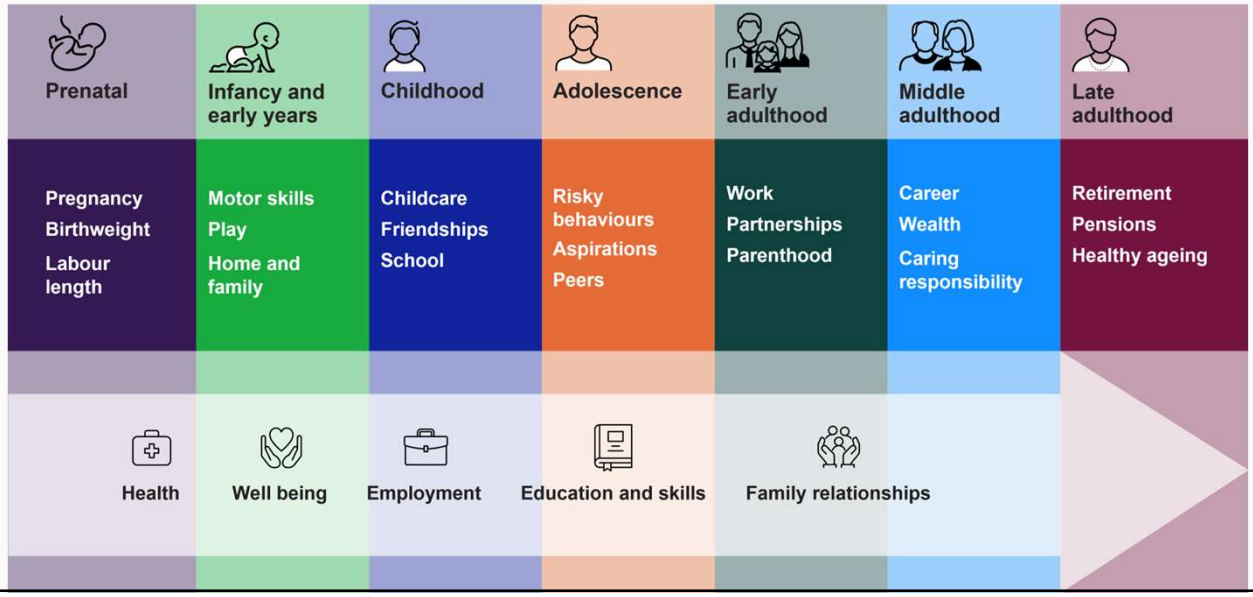
4

CLS core study timelines



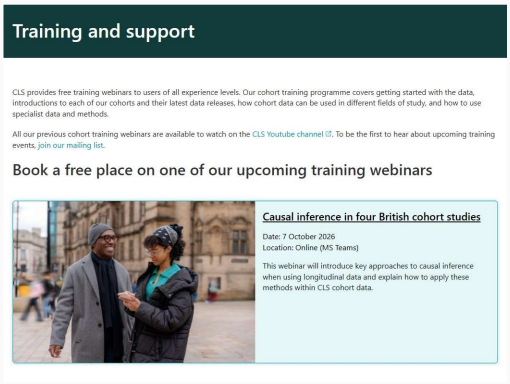
5

Types of information collected across life

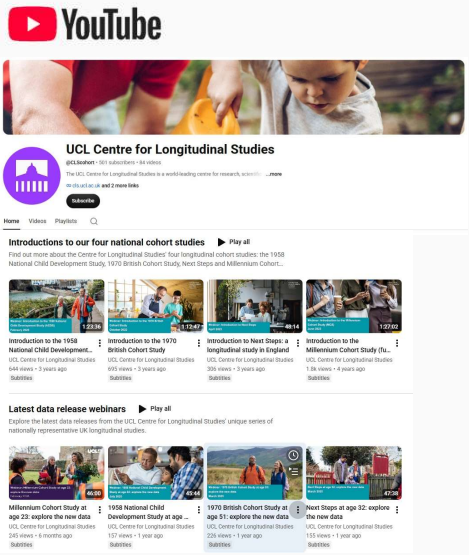


6

CLS training and support



cls.ucl.ac.uk/training



@CLScohort

7

Agenda

- The problem of causal inference with observational data
- Three approaches to causal inference:
  - Model-based approaches
  - Design-based approaches
  - Elaborate theories
- Properties of CLS cohort data supporting causal inference
- Quantitative bias analysis
- Q&A

Georgia → Liam → Richard

UCL Centre for Longitudinal Studies

8

Causal questions

Centre for Longitudinal Studies

9

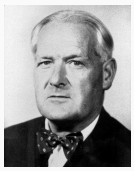
First encounters with causality

"Correlation ≠ Causation"

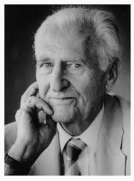
Centre for Longitudinal Studies

10

Causality: reserved for experiments



A Bradford Hill
(1897-1991)



Richard Doll
(1912-2005)



Doll R, Bradford Hill A. Lung cancer and other causes of death in relation to smoking. BMJ 1957



Ronald Fisher
(1890-1962)

British Medical J., vol. II, p. 43, 6 July 1957 and vol. II, pp. 297–298, 3 August 1957.

269–270

ALLEGED DANGERS OF CIGARETTE-SMOKING

In the *Journal* of July 20 Dr. Robert N. C. McCurdy writes: "Fisher's criticism† …would not be so unfair if he had specified what alternative explanations of the facts still await exclusion". I had hoped to be brief. A few days later the B.B.C. gave me the opportunity of putting forward examples of the two classes of alternative theories which any statistical association, observed without the predictions of a definite experiment, allows—namely, (1) that the supposed effect is really the cause, or in this case that incipient cancer, or a pre-cancerous condition with chronic inflammation, is a factor in inducing the smoking of cigarettes, or (2) that cigarette smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype.

I do not relish the prospect of this science being now discredited by a catastrophic and complacent howler. For it will be as clear in retrospect, as it is now in logic, that the data so far do not warrant the conclusions based upon them.

UCL Centre for Longitudinal Studies

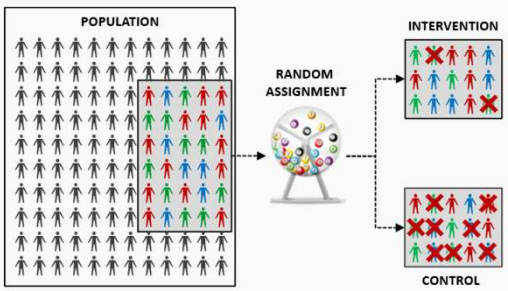
11

Randomised controlled trials

Randomly assign an '**intervention**' or '**control**'.

With large enough numbers, the two groups should be similar (in outcome probability) at assignment.

Subsequent group differences can be inferred as **average causal effect** of **intervention** vs **control** (with caveats).



UCL Centre for Longitudinal Studies

12

Randomised controlled trials

RCTs are not always possible. Interventions may be:

Unethical
- smoking
- childhood adversity
- pollution

Impractical
- school type
- house ownership
- birthplace

Impossible
- personality
- self-esteem
- birth order

But we all learn "correlation ≠ causation" and that causation is reserved for RCTs.

UCL Centre for Longitudinal Studies

13

Randomised controlled trials

RCTs are not always possible. Interventions may be:

Unethical
- smoking
- childhood adversity
- pollution

Impractical
- school type
- house ownership
- birthplace

Impossible
- personality
- self-esteem
- birth order

But we all learn "correlation ≠ causation" and that causation is reserved for RCTs.

An unfortunate **"solution":** pretend we are not interested in causal effects!

UCL Centre for Longitudinal Studies

14

Causal language: euphemisms



UCL Centre for Longitudinal Studies

15

Schrödinger’s inference

“where the authors caution against causal interpretations while themselves offering causal interpretations”

Tennant & Murray (2021) *Epidemiology.* DOI: 10.1097/EDE.0000000000001258

Typically occurs when:

- a study makes implicit causal claims or a call to policy action:

e.g. “this study found that higher levels of X reduced the risk of Y”,
e.g. “this study motivates the need for interventions in X to reduce the burden of Y”

- while also including a sentence that causal inferences cannot be made:

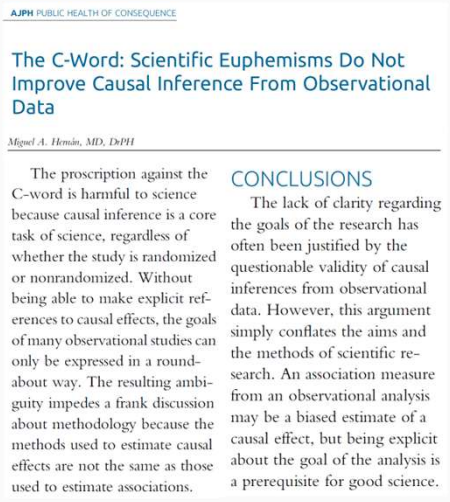
e.g. “Since our study relied on observational data, causality cannot be inferred.”

Source: Tennant (2025) LinkedIn.

UCL Centre for Longitudinal Studies

16

Causal language



AJPH PUBLIC HEALTH OF CONSEQUENCE

The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data

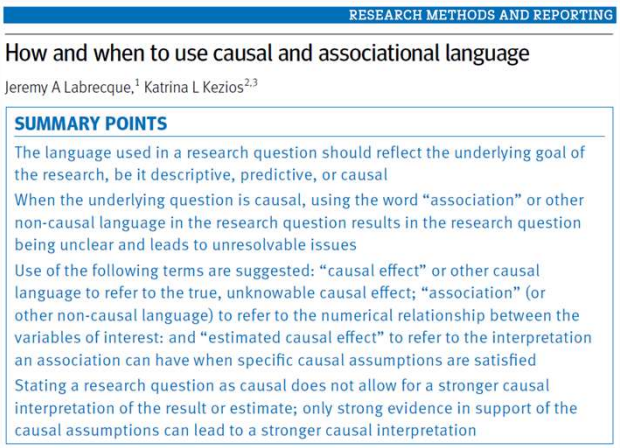
Miguel A. Hernán, MD, DrPH

The proscription against the C-word is harmful to science because causal inference is a core task of science, regardless of whether the study is randomized or nonrandomized. Without being able to make explicit references to causal effects, the goals of many observational studies can only be expressed in a roundabout way. The resulting ambiguity impedes a frank discussion about methodology because the methods used to estimate causal effects are not the same as those used to estimate associations.

CONCLUSIONS

The lack of clarity regarding the goals of the research has often been justified by the questionable validity of causal inferences from observational data. However, this argument simply conflates the aims and the methods of scientific research. An association measure from an observational analysis may be a biased estimate of a causal effect, but being explicit about the goal of the analysis is a prerequisite for good science.

Hernán (2018) *Am J Public Health*. DOI: 10.2105/AJPH.2018.304337



RESEARCH METHODS AND REPORTING

How and when to use causal and associational language

Jeremy A Labrecque,[1] Katrina L Kezios[2,3]

SUMMARY POINTS

The language used in a research question should reflect the underlying goal of the research, be it descriptive, predictive, or causal

When the underlying question is causal, using the word "association" or other non-causal language in the research question results in the research question being unclear and leads to unresolvable issues

Use of the following terms are suggested: "causal effect" or other causal language to refer to the true, unknowable causal effect; "association" (or other non-causal language) to refer to the numerical relationship between the variables of interest: and "estimated causal effect" to refer to the interpretation an association can have when specific causal assumptions are satisfied

Stating a research question as causal does not allow for a stronger causal interpretation of the result or estimate; only strong evidence in support of the causal assumptions can lead to a stronger causal interpretation

Labrecque & Kezios (2026) *BMJ*. DOI: 10.1136/bmj-2025-085749

UCL Centre for Longitudinal Studies

17

Three distinct tasks

Most research questions in health and social science can be divided into **three scientific tasks**, each with different methods and philosophies.



Shmueli (2010) To Explain or to Predict? *Stat Sci*. DOI: 10.1214/10-STS330
Hernán et al. (2019) A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*. DOI: 10.1080/09332480.2019.1579578

UCL Centre for Longitudinal Studies

18

Description



- Focussed on **summarising**, **describing**, &/or **visualising.**
- Aim is to characterise some feature(s) of the population.

Example:

What proportion of people aged 18-24 are Not in Education, Employment, or Training (NEET) across different birth cohorts?

UCL Centre for Longitudinal Studies

19

Prediction



- Focussed on **pattern recognition** and **forecasting.**
- Aim is to estimate the expected value of an unmeasured outcome from a set of predictor variables.

Example:

Who is at high risk of being Not in Education, Employment, or Training (NEET) at age 21 based on early-life and parental characteristics?

UCL Centre for Longitudinal Studies

20

Causal inference



- Focused on **understanding.**
- Aim is to estimate the effects of exposures on outcomes.
- Often framed as 'what if' scenarios.

Example:

What is the effect of ever experiencing school exclusion on being Not in Education, Employment, or Training (NEET) at age 21?

UCL Centre for Longitudinal Studies

21

Prediction vs causal inference

Example: **First week of school attendance**

- "Pupils in England who missed school during the first week back in September 2024 were more likely to be absent for large parts later in the year".
- Attendance during first week of school is a good predictor of future attendance.
- This allows us to predict future behaviour.
- But does not mean that *enforcing* attendance in the first week itself has any *effect* on future attendance.



UCL Centre for Longitudinal Studies

22

Description vs causal inference

Example: **Descriptive associations 'adjusted for variables'**

- Descriptive studies often present associations adjusted for many variables.
- A justification is rarely provided.
- Adjusted associations are often not appropriate for descriptive aims.
- Many descriptive studies even talk about 'confounding' – an entirely causal concept.

Descriptive estimands, causal estimands, and avoiding the jungle of adjusted associations 'in between'

Moritz Oberndorfer [1,2,3,4] Annie Herbert [5,6]
S Vittal Katikireddi [7] Anna Pearce [8]

Oberndorfer et al. (2026) *J Epidemiol Community Health*.
DOI: 10.1136/jech-2026-226120

JOURNAL ARTICLE EDITOR'S CHOICE

A Framework for Descriptive Epidemiology

Catherine R Lesko, Matthew P Fox, Jessie K Edwards

Lesko et al. (2022). *Am J Epidemiol*.
DOI: 10.1093/aje/kwac115

UCL Centre for Longitudinal Studies

23

Distinct tasks

- Summarising occurrence and trends
- Broad group comparisons
- Generating hypotheses



- Understanding mechanisms
- Identifying targets for intervention
- Testing interventions

- Screening / detection
- Forecasting / prognosticating

UCL Centre for Longitudinal Studies

24

‘Associational’ questions

Without clear aims and goals, we drift into unclear **associational questions.**



What are the ‘risk factors’ for being NEET at age 21?

What is the association between mental ill health during adolescence and being NEET at age 21?

UCL Centre for Longitudinal Studies

25

‘Associational’ questions

Without clear aims and goals, we drift into unclear **associational questions.**



BMJ 2016;355:i6536 doi: 10.1136/bmj.i6536 (Published 9 December 2016)

Is caviar a risk factor for being a millionaire?

Anders Huitfeldt argues that the answer depends on your definition of “risk factor” and calls for greater clarity in research

Anders Huitfeldt *postdoctoral scholar*

Meta-Research Innovation Center at Stanford, Stanford University School of Medicine, Palo Alto, CA, USA

DOI: 10.1136/bmj.i6536

UCL Centre for Longitudinal Studies

26

‘Associational’ questions

It is important to have a clearly defined research question and task (and not fall into ‘associational’ questions) because:

- The task informs the entire approach.
- Each task may use similar statistical methods (e.g. linear regression), but how this method is used (e.g. covariate choice) depends on the task.

Wrong approach

↓

Wrong results and/or interpretation

↓

Misleading evidence

UCL Centre for Longitudinal Studies

27

Causal questions and formulating estimands

Formulating a causal question involves outlining it in terms of the precise comparison that you want to make, and in what population.

Exposure	Experiencing school exclusion during Years 7-11
Comparator	Not experiencing school exclusion during Years 7-11
Outcome	Being NEET at age 21
Population	Pupils starting secondary school in England in 2015
Effect measure	Risk ratio
Intercurrent events	Custody (incorporated into a composite outcome)

UCL Centre for Longitudinal Studies

28

Causal questions and formulating estimands

Among pupils starting secondary school in England in 2015, how many times higher would the risk of being NEET (or in custody) at age 21 be if they experienced school exclusion during Years 7–11, compared with if they did not?

Centre for Longitudinal Studies

29

Causal questions and formulating estimands

- **PICOT**
  Richardson W S, Wilson M C, Nishikawa J, et al. (1995) **The well-built clinical question: a key to evidence-based decisions**. *ACP Journal Club*, 123(3) doi:10.7326/ACPJC-1995-123-3-A12
- **ICH E9(R1)**
  Kahan B C, Hindley J, Edwards M, Cro S, Morris T P. (2024) **The estimands framework: a primer on the ICH E9(R1) addendum**. *BMJ 2024*, 384:e076316 doi:10.1136/bmj-2023-076316
- **Target Trial Emulation**
  Hernán M A, Robins J M. (2016) **Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available**. *American Journal of Epidemiology*, 183(8), Pages 758–764 doi:10.1093/aje/kwv254

Centre for Longitudinal Studies

30

Causal inference frameworks

Centre for Longitudinal Studies

31

Causal inference as a structured approach

Centre for Longitudinal Studies

CAUSAL ESTIMAND → STATISTICAL ESTIMAND



Goal
e.g. True difference in Y due to changing X



Recipe
e.g. Observable difference in Y between X | Z

IDENTIFICATION

Source: Adapted with permission from Peter Tennant.

32

Causal inference as a structured approach

Centre for Longitudinal Studies

CAUSAL ESTIMAND → STATISTICAL ESTIMAND → ESTIMATOR + DATA



Goal
e.g. True difference in Y due to changing X



Recipe
e.g. Observable difference in Y between X | Z



Equipment
e.g. OLS Regression model



Ingredients
e.g. Frozen Follow-up Study (FFS)

IDENTIFICATION → ESTIMATION

Source: Adapted with permission from Peter Tennant.

33

Causal inference as a structured approach

Centre for Longitudinal Studies

CAUSAL ESTIMAND → STATISTICAL ESTIMAND → ESTIMATOR + DATA → ESTIMATE



Goal
e.g. True difference in Y due to changing X



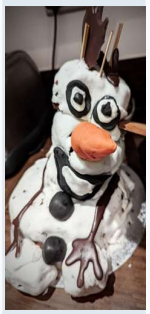
Recipe
e.g. Observable difference in Y between X | Z



Equipment
e.g. OLS Regression model



Ingredients
e.g. Frozen Follow-up Study (FFS)



Attempt
e.g. Estimated difference in Y due to X

IDENTIFICATION → ESTIMATION

Source: Adapted with permission from Peter Tennant.

34

Causal inference as a structured approach

Centre for Longitudinal Studies

CAUSAL ESTIMAND → STATISTICAL ESTIMAND → ESTIMATOR + DATA → ESTIMATE → FALSIFICATION/ ASSUMPTIONS



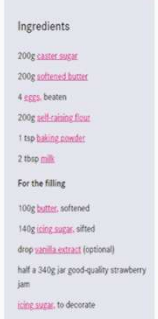
Goal
e.g. True difference in Y due to changing X



Recipe
e.g. Observable difference in Y between X | Z



Equipment
e.g. OLS Regression model



Ingredients
e.g. Frozen Follow-up Study (FFS)



Attempt
e.g. Estimated difference in Y due to X



Quality control
e.g. Negative control

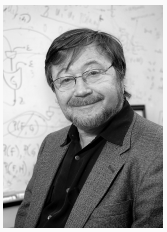
IDENTIFICATION → ESTIMATION

Source: Adapted with permission from Peter Tennant.

35

Causal inference frameworks

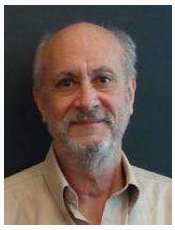
Structural Causal Model (SCM)



Judea Pearl (1936-)

- Directed acyclic graphs (DAGs)
- Structural equations
- Do-calculus
- More popular in computer science

Potential Outcomes Framework



Donald Rubin (1943-)



Jerzy Neyman (1894-1981)

- Potential outcome contrasts
- Explicit focus on estimands
- More popular in epidemiology/ econometrics

- **Largely interchangeable and compatible**
- **Ideas in one can be re-expressed in the other**
- **Often used in combination, e.g. potential outcome notation alongside DAGs**

Centre for Longitudinal Studies

36

Potential outcomes

Suppose we are interested in:



School exclusion in Years 7-11 (X)

Being NEET at Age 21 (Y)

UCL Centre for Longitudinal Studies

37

Potential outcomes

Suppose we are interested in:



School exclusion in Years 7-11 (X)

Being NEET at Age 21 (Y)

In theory, we would need to compare:

Y^0 vs Y^1

If NEET when **X=0 (no exclusion)**

If NEET when **X=1 (exclusion)**

UCL Centre for Longitudinal Studies

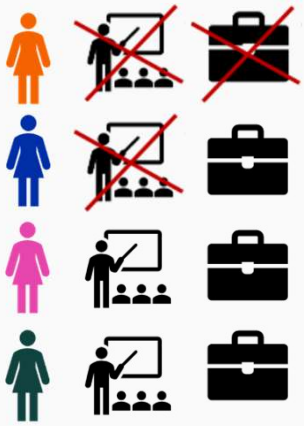
38

Potential outcomes

Each person has two potential outcomes: **Y^0 and Y^1**

But we can only observe **one** of them. We can never know the **counterfactual**.

Known as the "**fundamental problem of causal inference**"



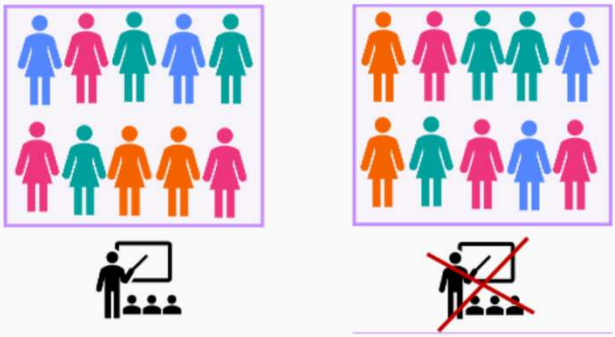
UCL Centre for Longitudinal Studies

39

Individual vs average effects

Instead, we aim to **identify comparable groups.**

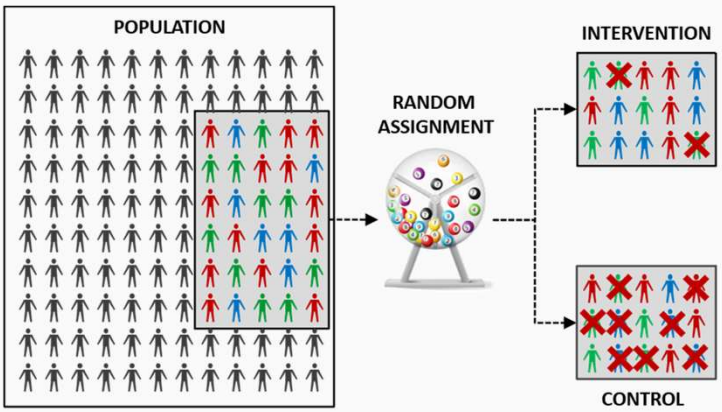
We can then estimate the **average causal effect (a.k.a. average treatment effect, ATE)** by comparing the observed outcomes between these groups, **subject to assumptions**.



UCL Centre for Longitudinal Studies

40

Individual vs average effects



Centre for Longitudinal Studies

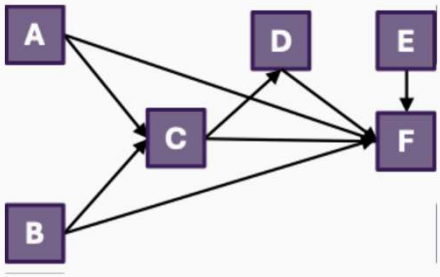
41

Directed acyclic graphs (DAGs)

DAGs are a type of causal diagram.

They are representations of (hypothesised) causal relationships between variables.

Useful for both applied and methodological research.



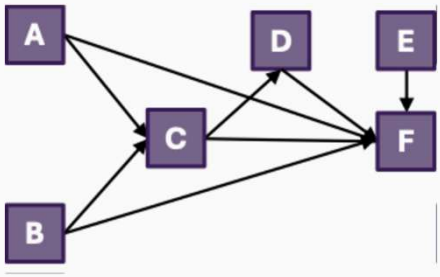
Centre for Longitudinal Studies

42

Directed acyclic graphs (DAGs)

Variables represented as **nodes** (e.g. A-F).

Causal effects represented as directed **arcs** (arrows, edges).



UCL Centre for Longitudinal Studies

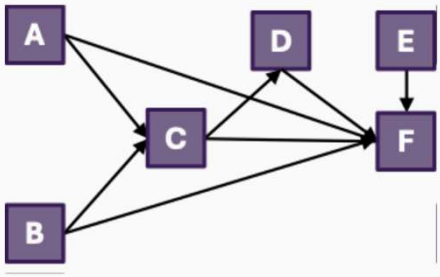
43

Directed acyclic graphs (DAGs)

Directed, i.e. explicitly depicting direction over time.
Acyclic, i.e. no circular paths.

Non-parametric, i.e. do not consider the strength, sign direction, or shape.

Probabilistic, e.g. A changes the probability of C
(but A is not necessary or sufficient for C).

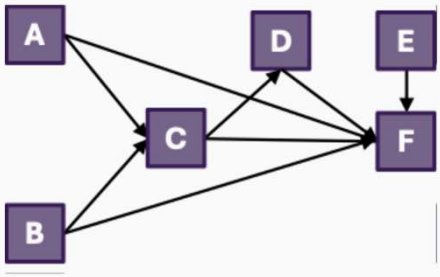


UCL Centre for Longitudinal Studies

44

Paths

A **path** exists between two variables if they are connected by one or more arcs (regardless of direction):
e.g. A → C → F
e.g. D → F ← E

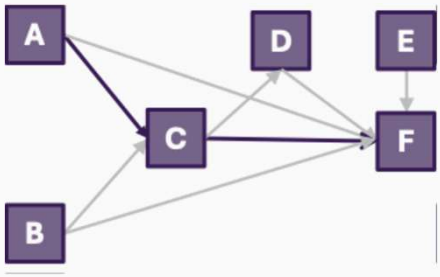


Centre for Longitudinal Studies

45

Paths

A **path** exists between two variables if they are connected by one or more arcs (regardless of direction):
e.g. **A → C → F**
e.g. D → F ← E

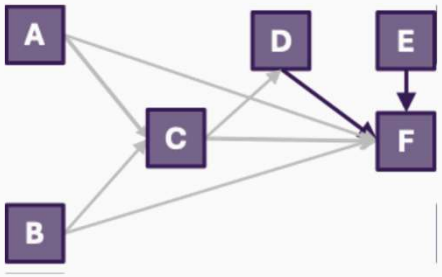


Centre for Longitudinal Studies

46

Paths

A **path** exists between two variables if they are connected by one or more arcs (regardless of direction):
e.g. A → C → F
e.g. **D → F ← E**



UCL Centre for Longitudinal Studies

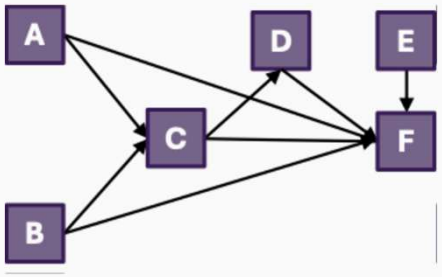
47

Paths

A **path** exists between two variables if they are connected by one or more arcs (regardless of direction):
e.g. A → C → F
e.g. D → F ← E

Open paths transmit associations (dependencies).

Closed paths do not.



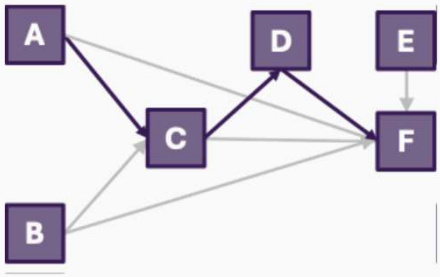
UCL Centre for Longitudinal Studies

48

Causal paths

A **causal** (or **directed**) path exists where all arcs run in the same direction:
e.g. **A → C → D → F**

The path can be closed by **conditioning** (controlling, adjusting, stratifying, matching) for a variable on the path.



UCL Centre for Longitudinal Studies

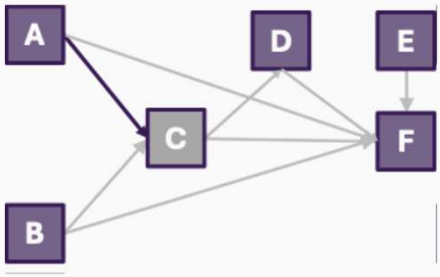
49

Causal paths

A **causal** (or **directed**) path exists where all arcs run in the same direction:
e.g. **A → C → D → F**

The path can be closed by **conditioning** (controlling, adjusting, stratifying, matching) for a variable on the path.

e.g. conditioning on C will close the path



UCL Centre for Longitudinal Studies

50

Main threats to causal inference in observational data

UCL Centre for Longitudinal Studies

51

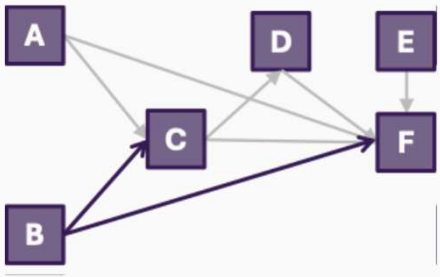
Confounding paths

A **confounding** (or '**backdoor**') path is a type of **non-causal** path.

e.g. in the path **C ← B → F,**

B is a common cause of C and F.

B will introduce a non-causal association between C and F (**confounding bias**).



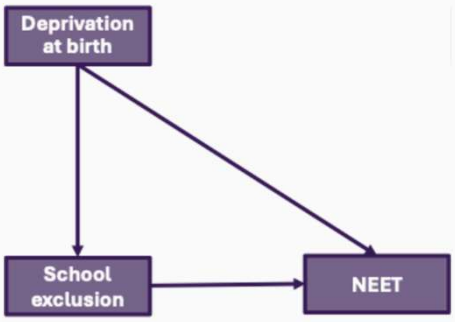
UCL Centre for Longitudinal Studies

52

Confounding bias

The observed association between **school exclusion** and **NEET** will reflect:

- Any true causal paths
- Any non-causal paths (e.g. confounding by **deprivation at birth**)



UCL Centre for Longitudinal Studies

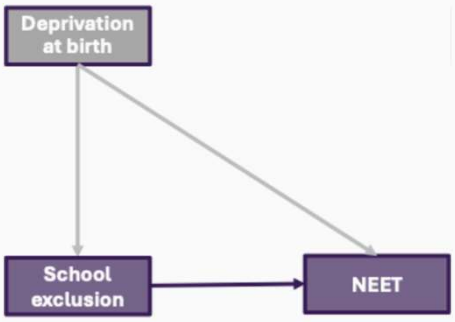
53

Confounding bias

The observed association between **school exclusion** and **NEET** will reflect:

- Any true causal paths
- Any non-causal paths (e.g. confounding by **deprivation at birth**)

Conditioning on **deprivation at birth** will close the confounding path.

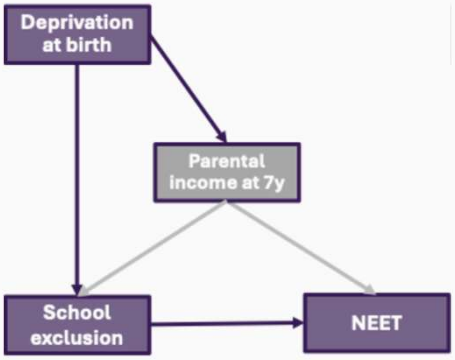


UCL Centre for Longitudinal Studies

54

Confounding bias

If the effect of '**deprivation at birth**' is entirely mediated through (or captured by) '**parental income at age 7**', then conditioning on either will close the confounding path.



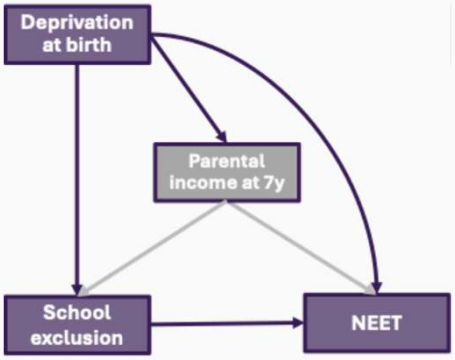
UCL Centre for Longitudinal Studies

55

Confounding bias

If the effect of '**deprivation at birth**' is entirely mediated through (or captured by) '**parental income at age 7**', then conditioning on either will close the confounding path.

If the effect is not fully captured, then **residual confounding** is likely.



UCL Centre for Longitudinal Studies

56

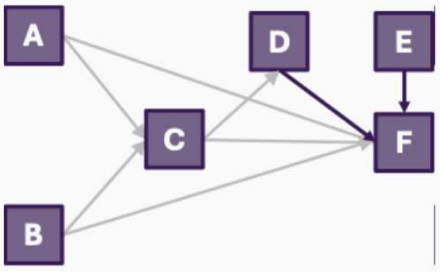
Collider paths

A collider path is another type of **non-causal** path.

e.g. in the path **D → F ← E,**

F is a collider of **D** and **E.**

Collider paths are **closed** and therefore do not automatically introduce any bias.



UCL Centre for Longitudinal Studies

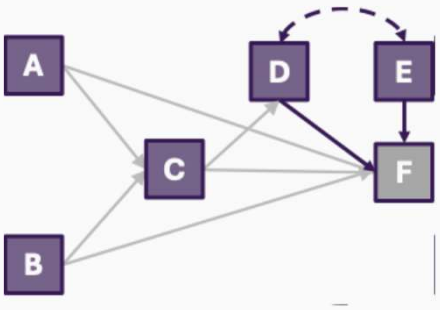
57

Collider paths

But they **can be opened** by conditioning on the collider.

e.g. conditioning on F will open the path **D → F ← E** and introduce a spurious association between **D and E.**

If D and E are the exposure and outcome, then this is will introduce **collider bias**.

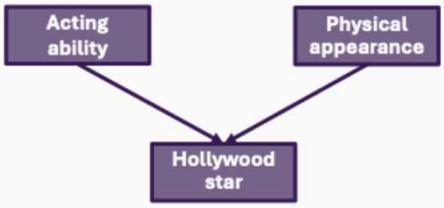


UCL Centre for Longitudinal Studies

58

Collider bias

Suppose there are two reasons why someone may become a Hollywood star: **acting ability** and **physical appearance.**



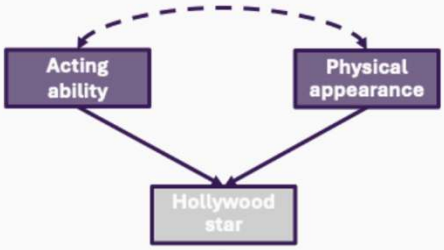
Centre for Longitudinal Studies

59

Collider bias

Suppose there are two reasons why someone may become a Hollywood star: **acting ability** and **physical appearance.**

Conditional on being a Hollywood actor, there will appear to be a negative association between these two reasons.



Centre for Longitudinal Studies

60

Collider adjustment bias

Suppose we are interested in the effect of school exclusion (Years 7-11) on NEET (Age 21).

We are considering whether to adjust for educational attainment at Age 16 (using '**GCSE results**').

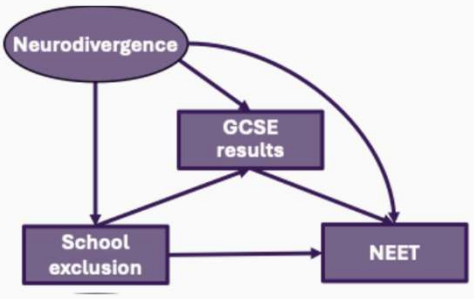


UCL Centre for Longitudinal Studies

61

Collider adjustment bias

There may be many other variables (e.g. neurodivergence) that affect both GCSE results and NEET.



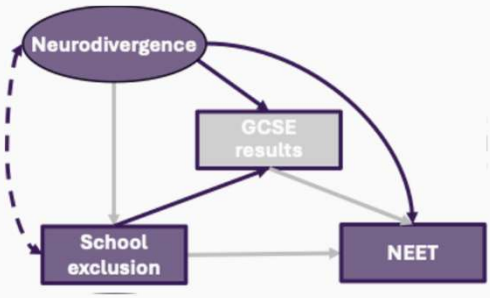
UCL Centre for Longitudinal Studies

62

Collider adjustment bias

There may be many other variables (e.g. neurodivergence) that affect both GCSE results and NEET.

If we naively adjust for GCSE results, without realising it is a **collider**, we risk introducing **collider bias**.

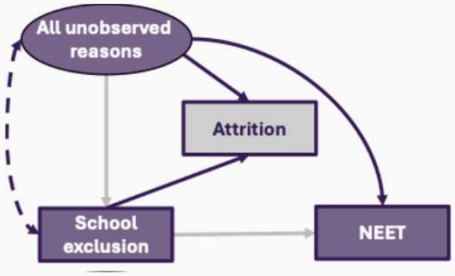


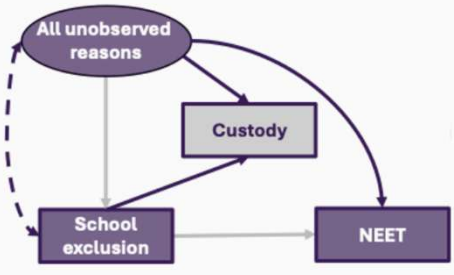
UCL Centre for Longitudinal Studies

63

Collider restriction bias

Collider bias can also be a result of sample selection processes (therefore often referred to as **selection bias**).





UCL Centre for Longitudinal Studies

64

Measurement (information) bias

With additional notation, DAGs can also be modified to depict **measurement error.**

e.g. X* is the measured version of the latent (unobserved) X



UCL Centre for Longitudinal Studies

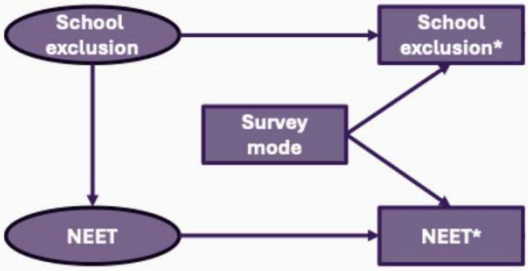
65

Measurement (information) bias

The mode of data collection (e.g. face-to-face interview, web questionnaire) can lead to differential reporting of exclusion and NEET.

Recent waves of CLS data have transitioned to mixed-mode designs.

Survey mode can introduce an additional non-causal path.



UCL Centre for Longitudinal Studies

66

Three approaches to causal inference

Centre for Longitudinal Studies

67

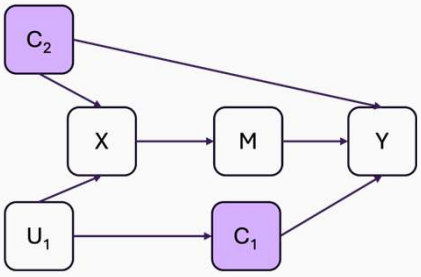
Model–based approach

Idea: Use ***post-hoc*** adjustment to close biasing paths between exposure and outcome.

DAGs show us this can work!

Methods:

- Regression adjustment
- Matching
- Inverse Probability Weighting



Centre for Longitudinal Studies

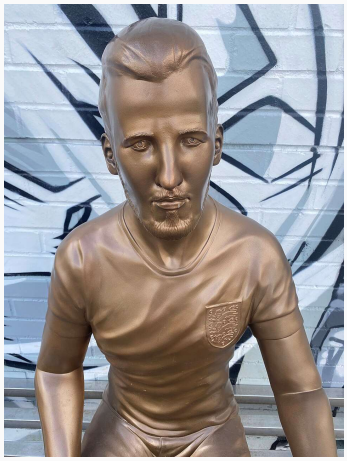
68

Model–based approach

Mechanics: Obtain residual variation in exposure that is **as-if random**

The 'Non-Equivalent Groups Design'

Akin to taking a block of marble and chiselling away to reveal correct form underneath (Huntington-Klein, 2026)



Source: Wikimedia Commons

UCL Centre for Longitudinal Studies

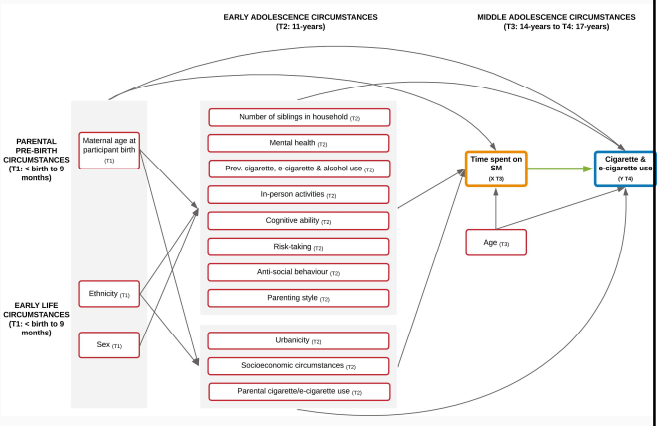
N. Huntington-Klein. *The Effect* (2026) https://theeffectbook.net
Ecco4k, *Harry Kane statue at the Peter May Sports Centre, Walthamstow, London*. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Harry_Kane_statue_2.jpg

69

Model–based approach

Limitations:

1. Requires **knowing** and **having** appropriate measurements of variables
2. Requires modelling these correctly (e.g., non-linearities)
3. Requires detailed understanding of causal structure (no ***Causal Salads;*** McElreath, 2021)



Source: Purba et al. (2025)

UCL Centre for Longitudinal Studies

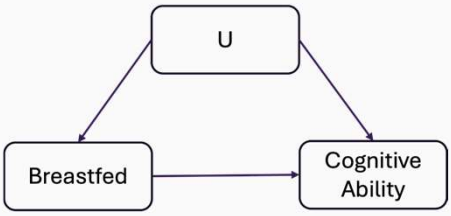
R. McElreath. *Regression, Fire, and Dangerous Things* (2021). https://elevanth.org/blog/2021/06/15/regression-fire-and-dangerous-things-1-3/
A. K. Purba et al., Nicotine and Tobacco Research **27**, 693 (2025). https://doi.org/10.1093/ntr/ntae057

70

Design–based approach

Idea: Identify ***a priori*** variation in exposure that is ***as-if random***

Example #1: Breastfeeding and Child Development (Fitzsimons & Vera-Hernández, 2022)



Centre for Longitudinal Studies

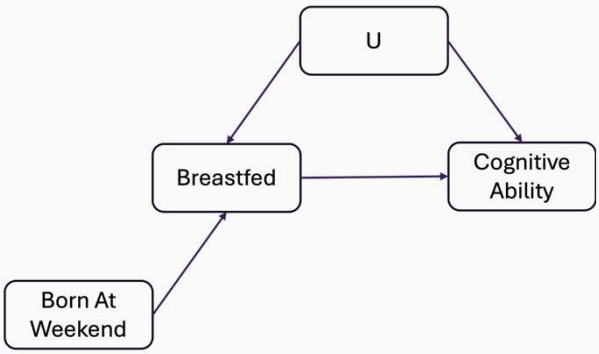
E. Fitzsimons and M. Vera-Hernández, American Economic Journal: Applied Economics **14**, 329 (2022). http://doi.org/10.1257/app.20180385

71

Design–based approach

Idea: Identify ***a priori*** variation in exposure that is ***as-if random***

Example #1: Breastfeeding and Child Development (Fitzsimons & Vera-Hernández, 2022)



Centre for Longitudinal Studies

E. Fitzsimons and M. Vera-Hernández, American Economic Journal: Applied Economics **14**, 329 (2022). http://doi.org/10.1257/app.20180385

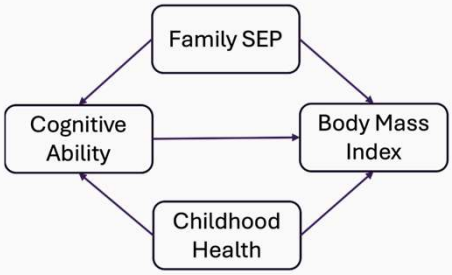
72

Design–based approach

Idea: Identify ***a priori*** variation in exposure that is ***as-if random*** …

… or, at least, unrelated to some usual sources of bias (e.g., *shared family background*)

Example #2: Cognitive Ability and Adult Body Mass Index (Wright et al., 2023)



UCL Centre for Longitudinal Studies

L. Wright, N. M. Davies, and D. Bann, PLoS Med **20**, e1004207 (2023). https://doi.org/10.1371/journal.pmed.1004207

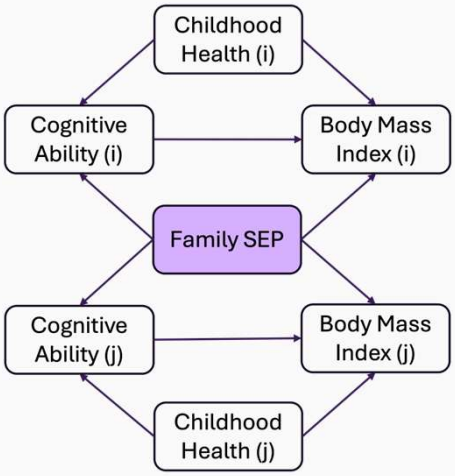
73

Design–based approach

Idea: Identify ***a priori*** variation in exposure that is ***as-if random*** …

… or, at least, unrelated to some usual sources of bias (e.g., *shared family background*)

Example #2: Cognitive Ability and Adult Body Mass Index (Wright et al., 2023)



UCL Centre for Longitudinal Studies

L. Wright, N. M. Davies, and D. Bann, PLoS Med **20**, e1004207 (2023). https://doi.org/10.1371/journal.pmed.1004207

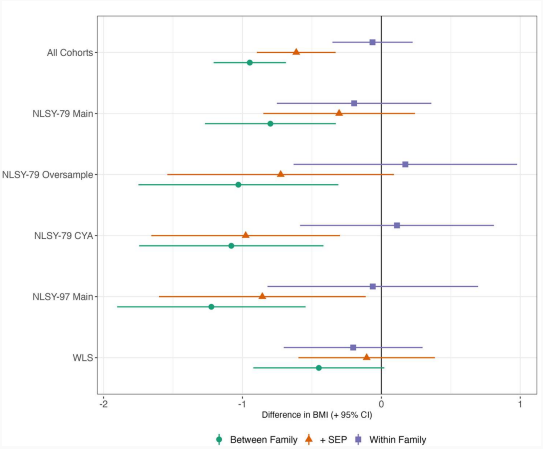
74

Design–based approach

Idea: Identify ***a priori*** variation in exposure that is ***as-if random*** …

… or, at least, unrelated to some usual sources of bias (e.g., *shared family background*)

Example #2: Cognitive Ability and Adult Body Mass Index (Wright et al., 2023)



UCL Centre for Longitudinal Studies

L. Wright, N. M. Davies, and D. Bann, PLoS Med **20**, e1004207 (2023). https://doi.org/10.1371/journal.pmed.1004207

75

Design–based approach

Connected to the idea of 'isolation':

- "A large biased data system may contain within it a smaller, much less-biased natural experiment. Isolation finds or constructs the natural experiment" (Zubizarreta, 2018)



Source: Google Maps

Designs:

- Instrumental Variables
- Regression Discontinuity Design
- Fixed Effects

UCL Centre for Longitudinal Studies

J. R. Zubizarreta, D. S. Small, and P. R. Rosenbaum, CHANCE **31**, 16 (2018). https://doi.org/10.1080/09332480.2018.1549811

76

Design–based approach

Connected to the idea of 'isolation':

- "A large biased data system may contain within it a smaller, much less-biased natural experiment. Isolation finds or constructs the natural experiment" (Zubizarreta, 2018)

Designs:

- Instrumental Variables
- Regression Discontinuity Design
- Fixed Effects



Source: Google Maps

UCL Centre for Longitudinal Studies

J. R. Zubizarreta, D. S. Small, and P. R. Rosenbaum, CHANCE **31**, 16 (2018). https://doi.org/10.1080/09332480.2018.1549811

77

Elaborate Theories

About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate". The reply puzzled me at first, since by Occam's razor the advice usually given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.

William Cochran (1965)

UCL Centre for Longitudinal Studies

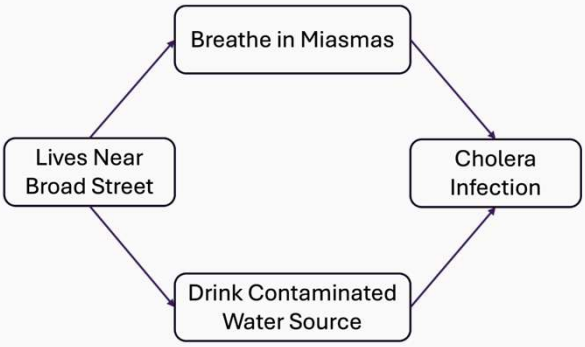
W. G.Cochran, Journal of the Royal Statistical Society. Series A (General) **128**, 234 (1965). https://doi.org/10.2307/2344179

78

Elaborate Theories

Example: John Snow and Cholera

- Snow's Theory: Cholera transmitted via infected water
- Competing Theory: Cholera transmitted via 'miasmas' (bad airs)
- Prediction: Higher caseloads clustered in geographic areas (e.g., around Soho's Broad Street Pump)



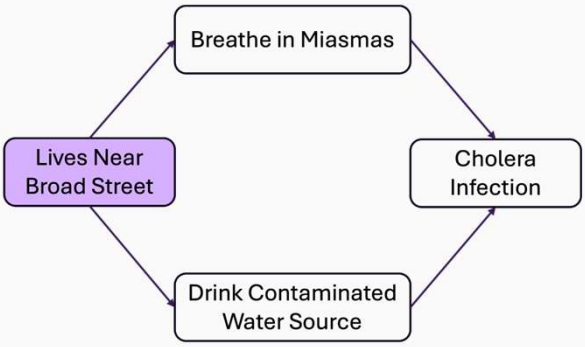
UCL Centre for Longitudinal Studies

79

Elaborate Theories

Elaborate Predictions:

#1: People living close, but drinking alternate water supply will be less likely to acquire cholera



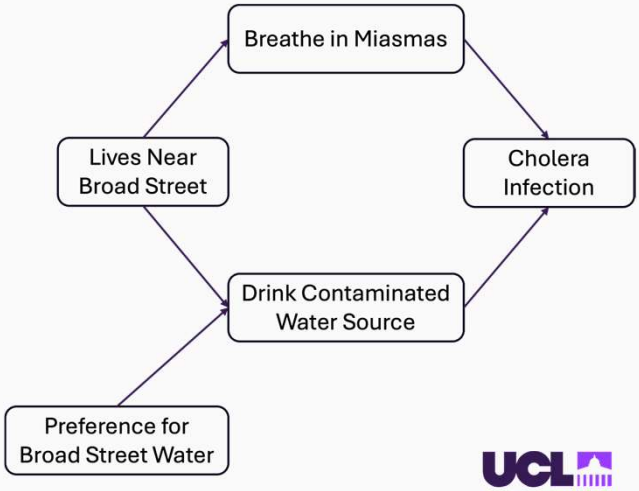
UCL Centre for Longitudinal Studies

80

Elaborate Theories

Elaborate Predictions:

#2: People living far away, but drinking Broad Street water will acquire cholera

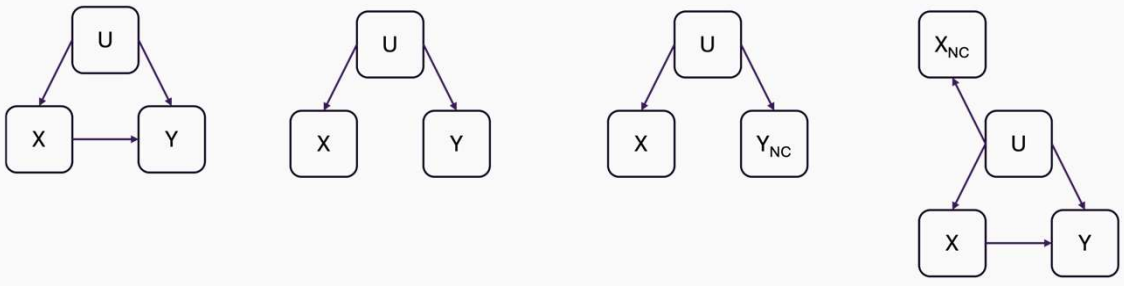


Centre for Longitudinal Studies

81

Negative Control Design

Idea: Identify alternate populations, outcomes or exposures for which a theorised causal effect should not exist



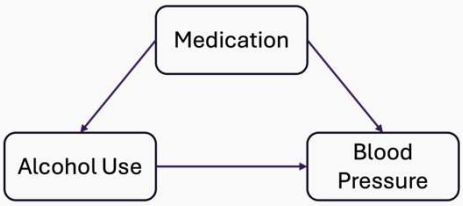
Centre for Longitudinal Studies

82

Negative Control Population Design

Example: ALDH2 and Blood Pressure (Chen et al., 2008)

- Elaborate Theory: ALDH2 should not be associated with blood pressure in populations where alcohol use is low regardless of ALDH2



Centre for Longitudinal Studies

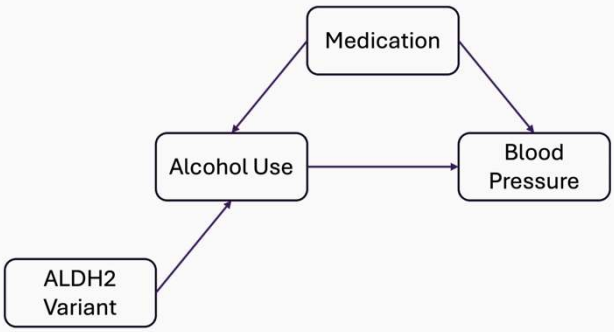
L. Chen, G. Davey Smith, R. M. Harbord, and S. J. Lewis, PLoS Med **5**, e52 (2008). https://doi.org/10.1371/journal.pmed.0050052

83

Negative Control Population Design

Example: ALDH2 and Blood Pressure (Chen et al., 2008)

- Elaborate Theory: ALDH2 should not be associated with blood pressure in populations where alcohol use is low regardless of ALDH2



Centre for Longitudinal Studies

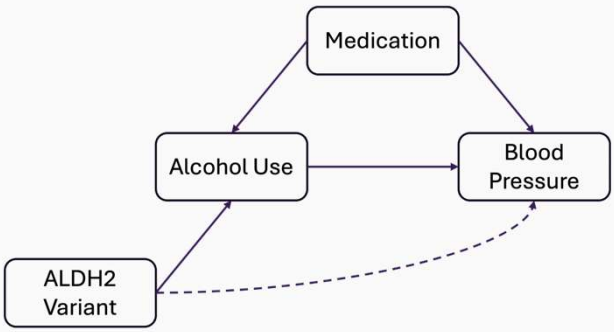
L. Chen, G. Davey Smith, R. M. Harbord, and S. J. Lewis, PLoS Med **5**, e52 (2008). https://doi.org/10.1371/journal.pmed.0050052

84

Negative Control Population Design

Example: ALDH2 and Blood Pressure (Chen et al., 2008)

- Elaborate Theory: ALDH2 should not be associated with blood pressure in populations where alcohol use is low regardless of ALDH2



Centre for Longitudinal Studies

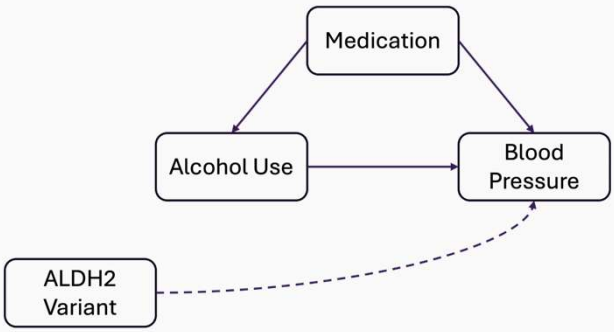
L. Chen, G. Davey Smith, R. M. Harbord, and S. J. Lewis, PLoS Med **5**, e52 (2008). https://doi.org/10.1371/journal.pmed.0050052

85

Negative Control Population Design

Example: ALDH2 and Blood Pressure (Chen et al., 2008)

- Elaborate Theory: ALDH2 should not be associated with blood pressure in populations where alcohol use is low regardless of ALDH2



Centre for Longitudinal Studies

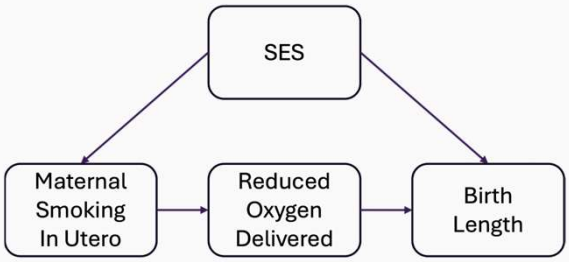
L. Chen, G. Davey Smith, R. M. Harbord, and S. J. Lewis, PLoS Med **5**, e52 (2008). https://doi.org/10.1371/journal.pmed.0050052

86

Negative Control Exposure Design

Example: Maternal Smoking During Pregnancy and Birth Length (Howe et al., 2012)

- Theory: Reduced *in utero* exposure to oxygen – caused by maternal smoking - restricts foetal growth
- Competing Theory: Maternal smoking confounded by SES
- Prediction: Maternal smoking related to shorter birth length



UCL Centre for Longitudinal Studies

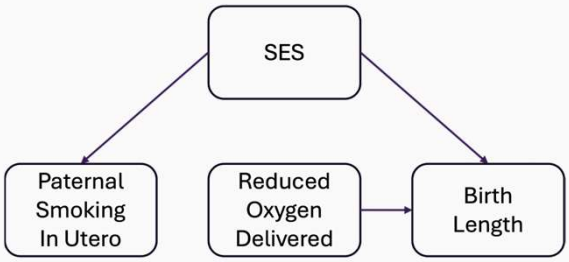
L. D. Howe, A. Matijasevich, K. Tilling, M.-J. Brion, S. D. Leary, G. D. Smith, and D. A. Lawlor, International Journal of Epidemiology **41**, 722 (2012). https://doi.org/10.1093/ije/dys025

87

Negative Control Exposure Design

Elaborate Predictions:

#1: Father's smoking while *in utero* should not be associated with birth length because it is unlikely to have a direct effect.



UCL Centre for Longitudinal Studies

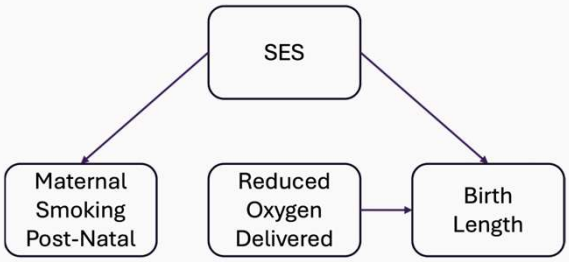
L. D. Howe, A. Matijasevich, K. Tilling, M.-J. Brion, S. D. Leary, G. D. Smith, and D. A. Lawlor, International Journal of Epidemiology **41**, 722 (2012). https://doi.org/10.1093/ije/dys025

88

Negative Control Exposure Design

Elaborate Predictions:

#2: Mother's *post-natal* smoking should not be associated with birth length because it cannot have a causal effect.



UCL Centre for Longitudinal Studies

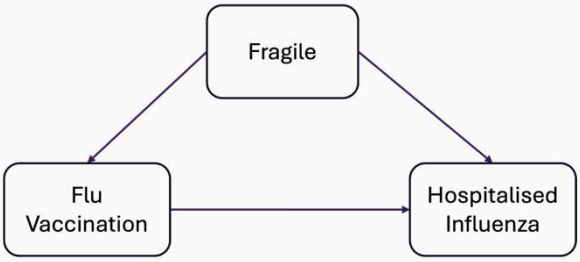
L. D. Howe, A. Matijasevich, K. Tilling, M.-J. Brion, S. D. Leary, G. D. Smith, and D. A. Lawlor, International Journal of Epidemiology **41**, 722 (2012). https://doi.org/10.1093/ije/dys025

89

Negative Control Outcome Design

Example: Flu Vaccination and Influenza Hospitalisation (Jackson et al., 2006)

- Theory: Flu vaccination successfully inoculates against the disease
- Competing Theory: There is confounding e.g., by fragility.
- Prediction: Association between flu vaccination and lower hospitalisation for influenza



UCL Centre for Longitudinal Studies

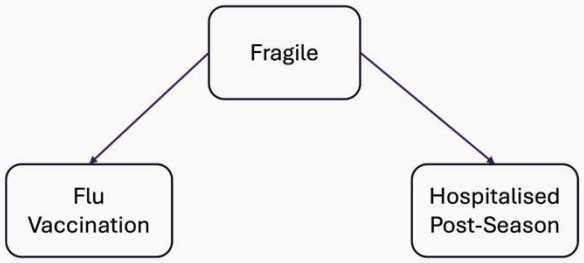
L. A. Jackson, M. L. Jackson, J. C. Nelson, K. M. Neuzil, and N. S. Weiss, International Journal of Epidemiology **35**, 337 (2006). https://doi.org/10.1093/ije/dyi274

90

Negative Control Outcome Design

Elaborate Predictions:

#1: Flu vaccination only has a causal effect when flu is prevalent (i.e., Winter)



Centre for Longitudinal Studies

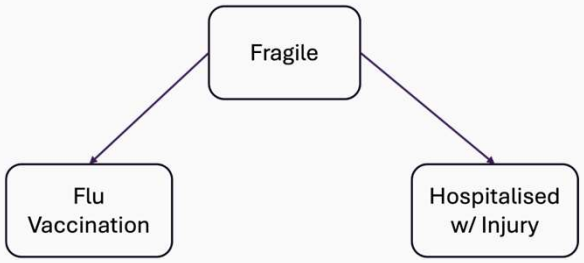
L. A. Jackson, M. L. Jackson, J. C. Nelson, K. M. Neuzil, and N. S. Weiss, International Journal of Epidemiology **35**, 337 (2006). https://doi.org/10.1093/ije/dyi274

91

Negative Control Outcome Design

Elaborate Predictions:

#2: Flu vaccination only has a causal effect on hospitalisation with flu



Centre for Longitudinal Studies

L. A. Jackson, M. L. Jackson, J. C. Nelson, K. M. Neuzil, and N. S. Weiss, International Journal of Epidemiology **35**, 337 (2006). https://doi.org/10.1093/ije/dyi274

92

Properties of CLS cohort data supporting causal inference

Centre for Longitudinal Studies

93

Breadth and depth

- Thousands of variables available in each study, covering period from birth (adolescence in Next Steps) to present day.
- Survey data, bio-samples, diaries, wearables, …
- Mental and physical health, health behaviours, psychosocial development, socioeconomic circumstances, environment, …
- Assumptions of no unmeasured confounding and missingness at random may be more plausible.
- Opportunity for negative controls/falsification tests.
- Data on other family members.

Centre for Longitudinal Studies

94

Data linkages

- Linkages to health, education and place-based data; HM Revenue and Customs and Department of Work and Pensions linkages coming soon.
- Increases breadth and depth of available information.
- Can further help with plausibility of no unmeasured confounding and missingness at random assumptions.
- Opportunities for quasi-experiments, e.g. exploiting arbitrary eligibility thresholds.
- Linkage of contextual information based on birth/interview date.

UCL Centre for Longitudinal Studies

95

Longitudinal assessment

- Same or similar questions are asked across time within each cohort.
- Enables more accurate measurement of variables.
- Can specify the correct temporal ordering.
- Can control for prior exposures and outcomes.
- Can use within-person variation, rather than rely solely on between-person differences.

UCL Centre for Longitudinal Studies

96

Cross-cohort analysis

- Cross-cohort analysis can expand time span or geographic area.
- Confounding structures or causal effects may be expected to differ over time or space (→ elaborate theories).
- Requires (sufficient) harmonisation of data – several projects doing so across CLS cohorts and internationally.

UCL Centre for Longitudinal Studies

97

Quantitative bias analysis

UCL Centre for Longitudinal Studies

98

Quantitative bias analysis (QBA)

- Knowledge of direction of bias can **qualitatively** support causal inference.
- This doesn't allow the analyst to determine **by how much** an estimate is likely to be biased or what an **unbiased estimate** would plausibly look like.
- QBA is a collection of methods designed to incorporate or output **quantitative** information on sensitivity of an estimate to bias.

Fox, M. P., et al. (2021). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer.

UCL Centre for Longitudinal Studies

99

Quantitative bias analysis (QBA)

QBA methods span:

- Application to summary statistics (e.g., regression coefficients) or to record-level data (i.e., observations).
- Provide bounds on bias to explain an association or generate bias-corrected estimates.
- Account for information, confounding and/or selection biases.
- Use deterministic or probabilistic and single- or multi-dimensional inputs.

UCL Centre for Longitudinal Studies

100

Cornfield conditions

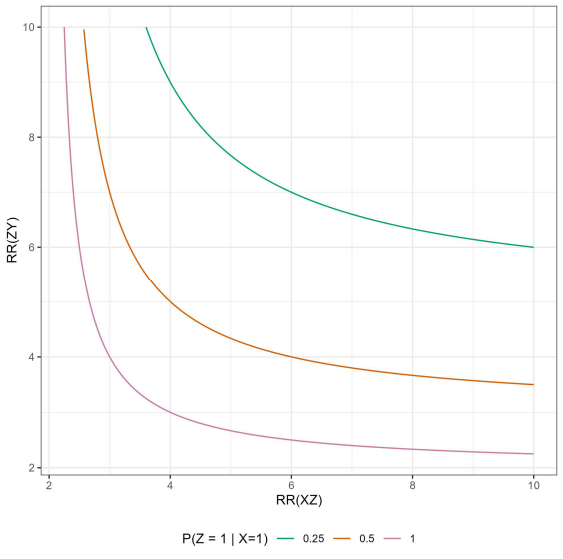
- Lower bounds on the associations a single binary confounder must have with a binary exposure and binary outcome to fully explain an observed association.
- Formalise several intuitions, including:
  - Confounders must have at least as large a causal effect on the outcome as the observed exposure-outcome association.
  - The effect of the confounder on the outcome must be stronger if the effect of the confounder on the exposure is weaker (and vice versa).
  - Unless the confounder has a particularly strong effect on the outcome, there must be strong imbalance in the prevalence of the confounder between levels of the exposure.

Cornfield, J., et al. (1959). *JNCI: Journal of the National Cancer Institute*, *22*(1), 173–203.

101

Cornfield conditions

For RR(XY) = 2



UCL Centre for Longitudinal Studies

102

Cornfield conditions

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

Cornfield, J., et al. (1959). *JNCI: Journal of the National Cancer Institute*, *22*(1), 173–203.

UCL Centre for Longitudinal Studies

103

E-value

- Builds on the Cornfield conditions by making two assumptions:
  1. The exposure-confounder and confounder-outcome effects are equal.
  2. The prevalence of the confounder among the exposed is 100%.
- However, both assumptions likely implausible, particularly #2.
- Can still be useful as a basis for more rigorous discussion of confounders.

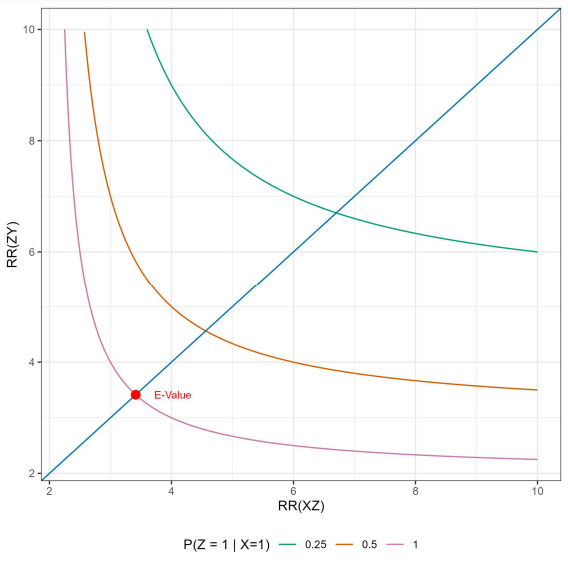
VanderWeele, T. J. & Ding, P. (2017). *Annals of Internal Medicine*, *167*(4), 268.

UCL Centre for Longitudinal Studies

104

E-value

For RR(XY) = 2



Centre for Longitudinal Studies

105

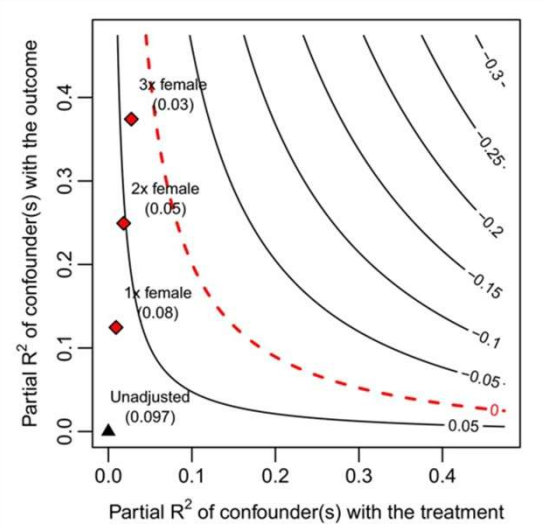
Cinelli & Hazlett

- Reparameterise omitted variable bias in terms of partial R^2 (proportion of residual variance in exposure and outcome a confounder explains after conditioning on covariates).
- Robustness value: minimum partial R^2 required to attenuate the observed association to the null (assuming equal partial R^2 for confounder-exposure and confounder-outcome associations).
- Accompanying R/Stata package also outputs information on the associations between *observed* confounders and exposure and outcome to contextualise consideration of whether unobserved confounders can plausibly explain results.

Cinelli, C. & Hazlett, C. (2020). *Journal of the Royal Statistical Society Series B*, *82*(1), 39–67.

106

Cinelli & Hazlett



Cinelli, C., et al. (2024). *sensemakr: Sensitivity Analysis Tools for Regression Models.*

UCL Centre for Longitudinal Studies

107

Counterfactual simulation

- Simulation-based method to correct observed data (at record level) for assumed error.
- Example: Addressing potential bias due to mode effects.
- Quality of life (CASP-6) measured by web and telephone.
- Concern that respondents may (artificially) report higher CASP-6 by telephone due to interviewer presence.
- Assume mode effect of 1 unit.

UCL Centre for Longitudinal Studies

108

Counterfactual simulation

ID	Mode	CASP-6
1	Tel	4
2	Web	5
3	Tel	5
4	Web	7
5	Tel	3
6	Web	6

Excerpt of a table showing a mock data of six individuals.

UCL Centre for Longitudinal Studies

109

Counterfactual simulation

ID	Mode	CASP-6	Sim
1	Tel	4	
2	Web	5	
3	Tel	5	
4	Web	7	
5	Tel	3	
6	Web	6	

Excerpt of a table showing a mock data of six individuals.

UCL Centre for Longitudinal Studies

110

Counterfactual simulation

ID	Mode	CASP-6	Sim
1	Tel	4	
2	Web	5	5
3	Tel	5	
4	Web	7	7
5	Tel	3	
6	Web	6	6

Excerpt of a table showing a mock data of six individuals.

UCL Centre for Longitudinal Studies

111

Counterfactual simulation

ID	Mode	CASP-6	Sim
1	Tel	4	3
2	Web	5	5
3	Tel	5	4
4	Web	7	7
5	Tel	3	2
6	Web	6	6

Excerpt of a table showing a mock data of six individuals.

UCL Centre for Longitudinal Studies

112

Counterfactual simulation

- Multiple approaches:
  - Best estimate.
  - Intentionally implausible.
  - Tipping point analysis.
- Flexible:
  - Correction can be differential.
  - Can address several issues simultaneously.
  - Deterministic and stochastic versions.
- Not so straightforward for non-continuous variables.

UCL Centre for Longitudinal Studies

113

Final thoughts

UCL Centre for Longitudinal Studies

114

Final thoughts

- Causal inference is one of the central goals of quantitative health and social science, but also one of the most difficult to attain.
- Important to recognise and precisely formulate the causal question.
- Ample opportunities for making causal inferences using CLS cohort data.
- Multiple different methods and broad approaches to utilise when doing so.
- All methods rely on (often untestable) assumptions.

115

Final thoughts

- Number and complexity of sources of bias in observational data (and corresponding solutions) can feel overwhelming.
- Often need to (ideally) address several issues at the same time in a given analysis.
- No single study (even an RCT) is ever likely to be perfect.
- Just addressing one such issue is likely already an improvement.
- We should not just accept bias – but don't let the perfect be the enemy of the good.

UCL Centre for Longitudinal Studies

116