

Don't forget the observables:
Capitalising on the richness of UK longitudinal surveys

George B. Ploubidis

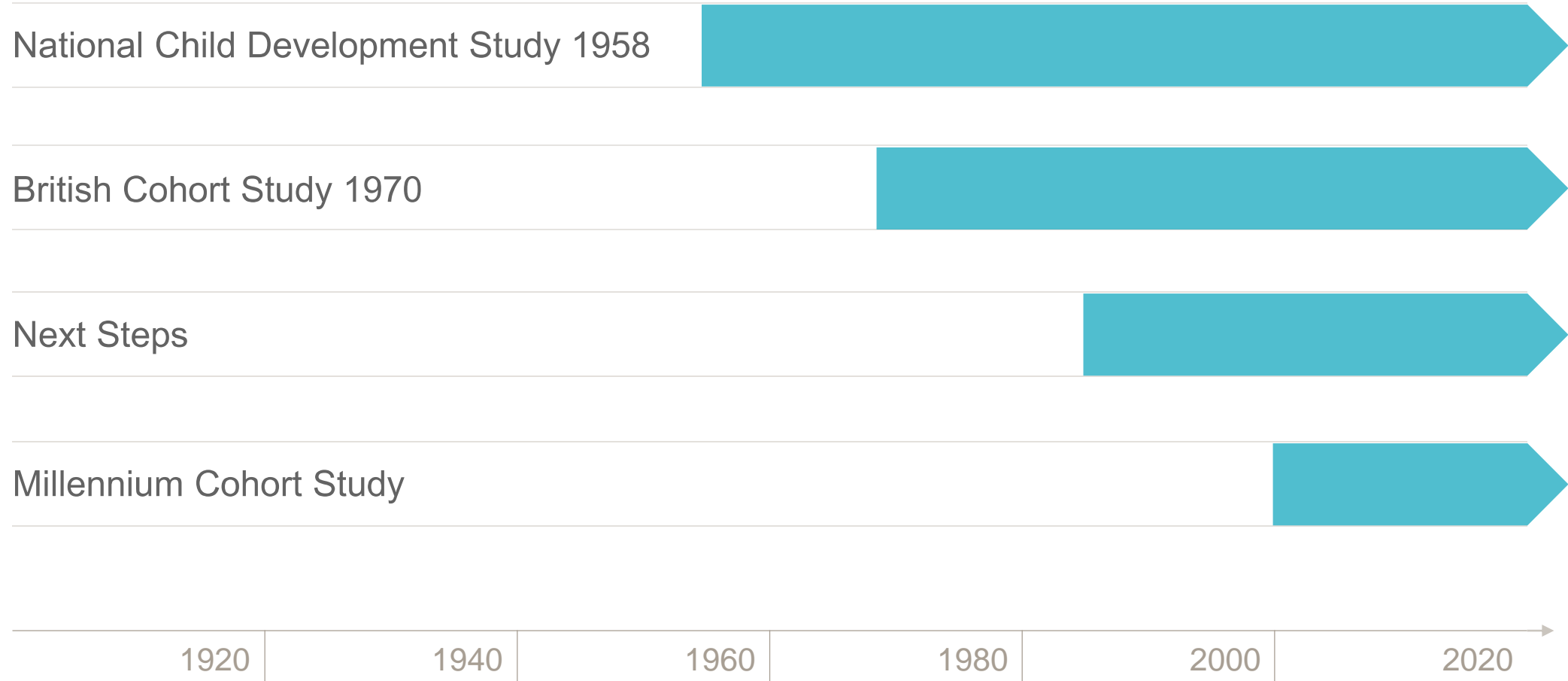
Outline

- The Centre for Longitudinal Studies (CLS)
- The untapped potential of UK longitudinal surveys
- Methodological challenges
- Practical solutions with reasonable assumptions

What we do at CLS

- We run four major national longitudinal studies, which follow people throughout life – “from cradle to grave”
- We are funded by the ESRC as a ‘resource centre’ – to provide data to the wider scientific community
- Deposit of data at UK Data Archive
- Scientific development, and design of the studies
- Training, capacity building, and user support
- Research – Strong multidisciplinary group

We follow people from 'cradle to grave'





Data are free!

- UK Data Service - <https://www.ukdataservice.ac.uk/>
- Takes less than 10 minutes to get access and download the data!

The screenshot shows the UK Data Service website homepage. At the top, there is a navigation bar with the following links: "About us", "Get data", "Use data", "Manage data", "Deposit data", and "News and events". The UK Data Service logo is on the left, and social media icons for LinkedIn, Facebook, Twitter, YouTube, and a purple "it!" icon are on the right. A "Register / Login" button is also present. Below the navigation bar is a dark blue banner with the text "Explore the UK's largest collection of social, economic and population data resources". A search bar with the placeholder "Search data" and a magnifying glass icon is located below the banner. The main content area is divided into two columns. The left column is titled "About the UK Data Service" and features a video player showing a red double-decker bus. The right column is titled "Guides and resources" and lists "Dataset guides", "Topic guides", "Methods and software guides", and "Guides to exploring online", with a "See more" link. Below the "Guides and resources" section is a purple box titled "Video tutorials" with the text "See our growing range of training videos". At the bottom of the page, there is a dark blue banner with the text "See data from all over the world" and a purple button labeled "Browse our data map".

UK Data Service

About us Get data Use data Manage data Deposit data News and events

in f t y it! Register / Login

Explore the UK's largest collection of social, economic and population data resources

Search data

About the UK Data Service

Guides and resources

Dataset guides

Topic guides

Methods and software guides

Guides to exploring online

See more

Video tutorials

See our growing range of training videos

See data from all over the world

Browse our data map

CLS Research Team

- CLS is a truly multidisciplinary group
- **38 full time researchers**
- CLS staff have backgrounds in economics, sociology, psychology, epidemiology, demography, statistics, mixed-methods research, survey methods and techniques of policy evaluation
- Experience in using the CLS datasets and other UK/International longitudinal surveys

CLS Scientific Themes – What the CLS Research Team Does

- Improving social mobility and generational change
- Tackling the obesity challenge
- Improving mental health across the whole of life
- Cognitive development and cognitive decline
- Healthy ageing and productive working older lives
- Families and intra-household dynamics
- Applied Statistical Methods
- Survey Methods
- Social Genomics

CLS Scientific Themes – What the CLS Research Team Does

- Improving social mobility and generational change
- Tackling the obesity challenge
- Improving mental health across the whole of life
- Cognitive development and cognitive decline
- Healthy ageing and productive working older lives
- Families and intra-household dynamics
- Applied Statistical Methods
- Survey Methods
- Social Genomics

CLS Applied Statistical Methods

- Applied methodological – **translational** - work which aims to reduce bias from the three major challenges in observational longitudinal data:
 - ✓ Missing data
 - ✓ Measurement error
 - ✓ Causal inference
- **Interdisciplinary approach**: Applying in the CLS data methods/ideas from Statistics/Biostatistics, Epidemiology, Econometrics, Psychometrics and Computer Science
- **Support and enable users** of the CLS data (including the CLS Research Team) to tackle these important challenges

Major methodological challenges in observational longitudinal data

Major methodological challenges in observational longitudinal data

- Missing data
- Causal inference
- Measurement error

Today

- Missing data
- Causal inference

- ✓ Both challenges directly related to (un)observed information
- ✓ Avoiding/minimising bias a function of the **observed information** and assumptions about unobservables
- ✓ Not all observational data are the same!

UK Longitudinal Surveys

- There is a **plethora of information** available in the UK's longitudinal surveys
- Untapped potential of **available information (the observables)**
- Implications for missing data handling and causal inference
- Relatively straightforward methods (and software!) available

Causal inference

Causal inference in observational data: A nearly alchemic task

- ✓ We cannot observe two values of a variable in the same unit/person at the same time
- ✓ Since only one value can be observed, the other value only exists in a **counterfactual**
- ✓ We can only estimate **Average Effects**, assuming no unmeasured confounders/omitted variable bias
- ✓ **Randomisation** removes bias, but...

Causal inference in observational data: A nearly alchemic task

- ✓ **Many (majority?) interesting and policy relevant topics** can only be studied with non – randomised/experimental data
- ✓ **Observational data:** Systematic way to collect data by observing people in natural situations or settings
- ✓ In epidemiology, social sciences, psychology and statistics, an observational study draws inferences from a **sample to a population** where the independent variable is not randomised because of ethical concerns or logistical constraints
- ✓ Other examples?

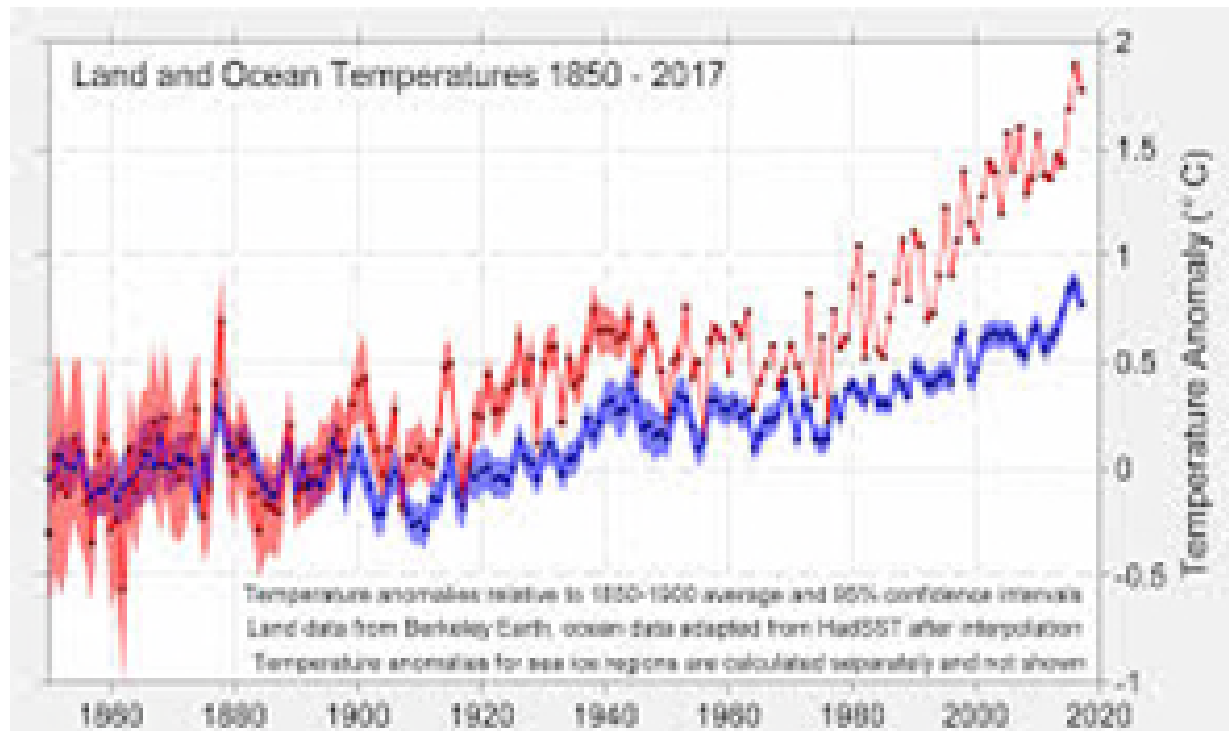
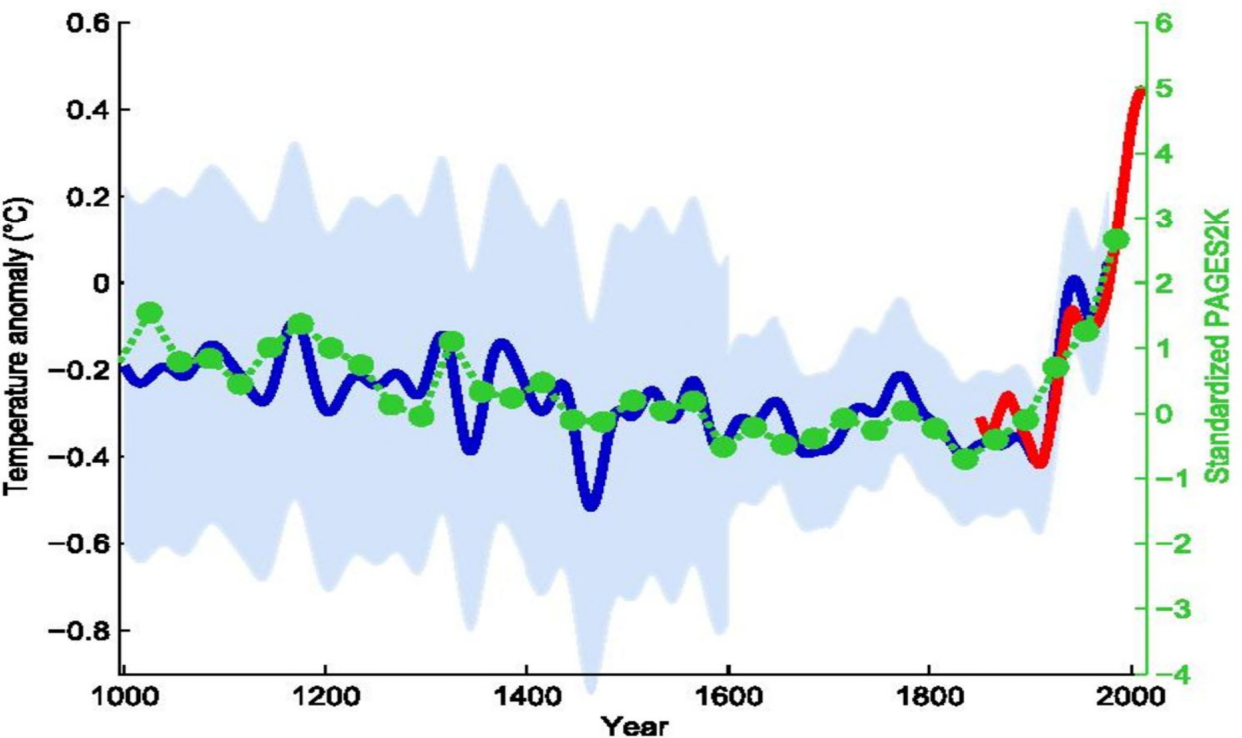


The Intergovernmental Panel on Climate Change

The Intergovernmental Panel on Climate Change (IPCC) is the United Nations body for assessing the science related to climate change.

PREVIOUS WEBSITE

SROCC



Causal inference in observational data: A nearly alchemic task

Available methods for causal inference in observational data:

- ✓ Instrumental Variables / Mendelian Randomisation
- ✓ Two sample Mendelian Randomisation
- ✓ Regression Discontinuity
- ✓ Fixed Effects / Dynamic fixed effects/Correlated Random Effects
- ✓ Multivariable Adjustment (MVA)

All methods rely on **untestable assumptions about unobservables**

Today

Available methods for causal inference in observational data:

- ✓ Instrumental Variables / Mendelian Randomisation
- ✓ Two sample Mendelian Randomisation
- ✓ Regression Discontinuity
- ✓ Fixed Effects / Dynamic fixed effects/Correlated Random Effects
- ✓ Multivariable Adjustment (MVA)

Most widely used

Only MVA estimates **Average Treatment Effects (ATE)**

All other methods estimate **Local Average Treatment Effects (LATE)**

Causal bounds

- 1) Not all association studies/observational data are alike: **varying extent to which confounders are able to be controlled for**
- 2) **Causal bounds** can illustrate what degree of unmeasured confounding would overturn an observed exposure-outcome association
 - ❖ Advantage: we do not have to fall back on estimating local effects from an IV approach (or making some of the other strong assumptions required for the IV)
 - ❖ Disadvantage: Does not explicitly deal with unmeasured confounding/omitted variable bias
- 3) Sometimes we can use '**negative controls**' to check how effective available data are to successfully control for confounding

Sensitivity Analysis Without Assumptions

Peng Ding^a and Tyler J. VanderWeele^b

RESEARCH AND REPORTING METHODS **Annals of Internal Medicine**

Sensitivity Analysis in Observational Research: Introducing the E-Value

Tyler J. VanderWeele, PhD, and Peng Ding, PhD

Sensitivity analysis is useful in assessing how robust an association is to potential unmeasured or uncontrolled confounding. This article introduces a new measure called the “E-value,” which is related to the evidence for causality in observational studies that are potentially subject to confounding. The E-value is defined as the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates. A large E-value implies that considerable unmeasured confounding would be needed to explain away an effect estimate. A small E-value implies little unmeasured confounding would be needed to explain away an effect estimate.

The authors propose that in all observational studies intended to produce evidence for causality, the E-value be reported or some other sensitivity analysis be used. They suggest calculating the E-value for both the observed association estimate (after adjustments for measured confounders) and the limit of the confidence interval closest to the null. If this were to become standard practice, the ability of the scientific community to assess evidence from observational studies would improve considerably, and ultimately, science would be strengthened.

Ann Intern Med. 2017;167:268-274. doi:10.7326/M16-2607

For author affiliations, see end of text.

This article was published at Annals.org on 11 July 2017.

Annals.org

E-Value: intuition

- In a model where we are interested in estimating the risk ratio RR associated with a particular exposure, E, on binary outcome, D
- Bias in estimates due to unmeasured confounding depends on the association between the unmeasured confounder(s) and
 - the exposure, (RR_{EU})
 - the outcome, (RR_{UD})
- The stronger these associations (conditional on other measured co-variates), the bigger the bias

Example given in Ding et al: breastfeeding & infant death by respiratory infection, **estimated reduction in relative risk 3.9**

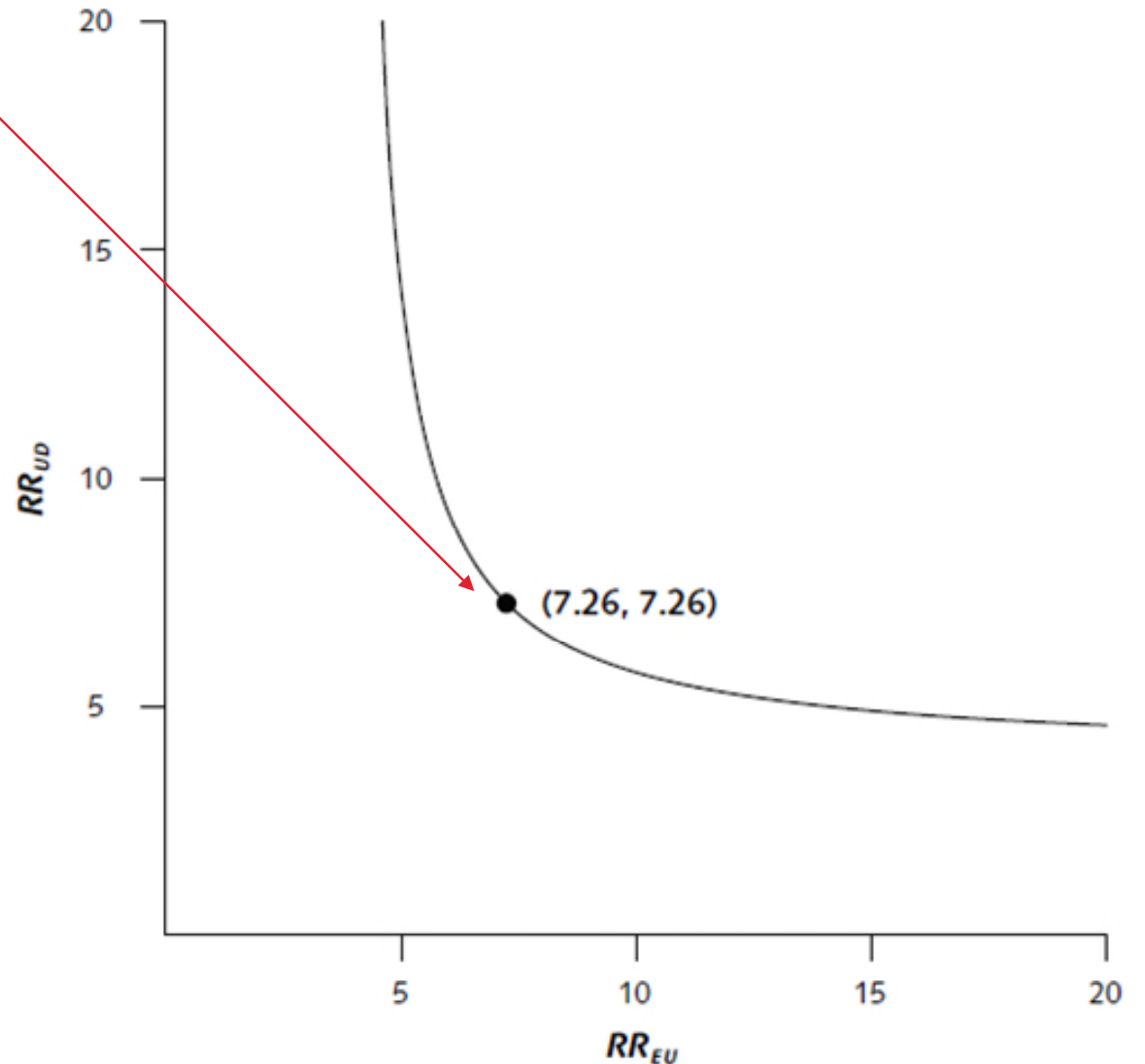


E-Value

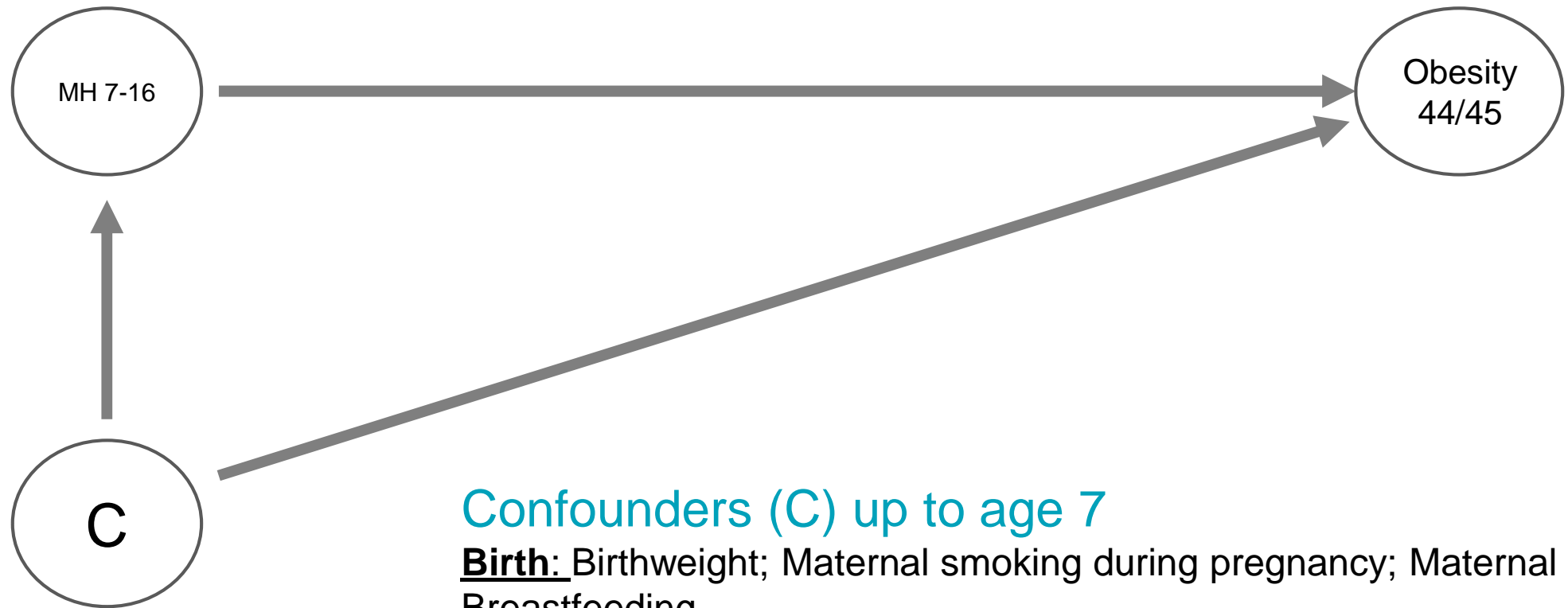
$$E\text{-value} = RR + \sqrt{RR \times (RR - 1)}$$

The E-value is the minimum strength of association, on the risk ratio scale, that an unmeasured confounder(s) would need to have with **both** the exposure and outcome, **conditional on the measured covariates**, to fully explain away a specific exposure–outcome association

In this example, the estimated RR is 3.9, and the E-Value is 7.26



Association study of childhood mental health and adult obesity in the 1958 cohort



Confounders (C) up to age 7

Birth: Birthweight; Maternal smoking during pregnancy; Maternal age; Breastfeeding

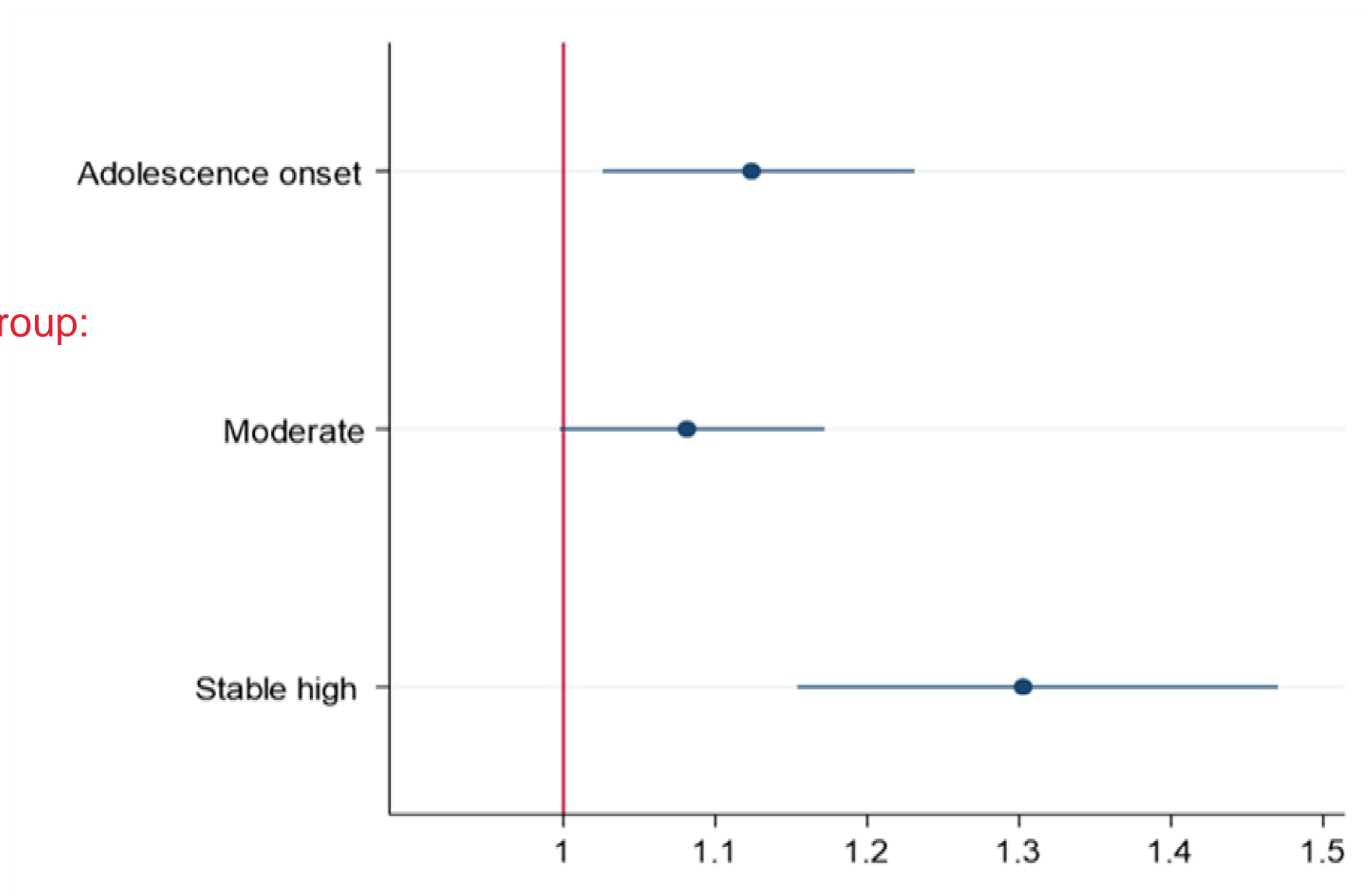
Parents: Mother working up to 5; Parents read to child; Parental interest in school; Divorce; Separation from child

SEP: Paternal social class at birth; Financial difficulties; Age mother stayed at school; Housing tenure; Access to amenities; Housing difficulties

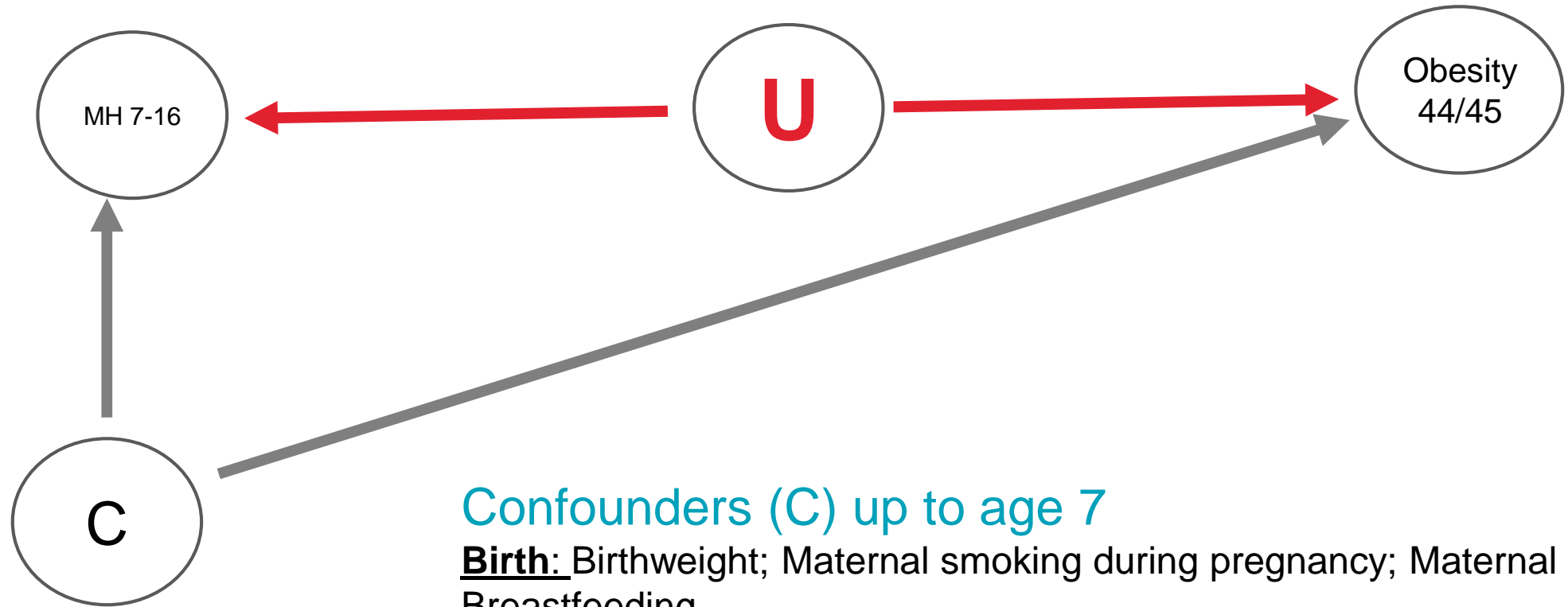
Cohort member: Cognitive ability; Enuresis; Summary of objectively assessed health conditions; BMI

Child mental health, and risk of adult abdominal obesity

Reference group:
Stable Low



Association study of childhood mental health and adult obesity in 1958 cohort



Confounders (C) up to age 7

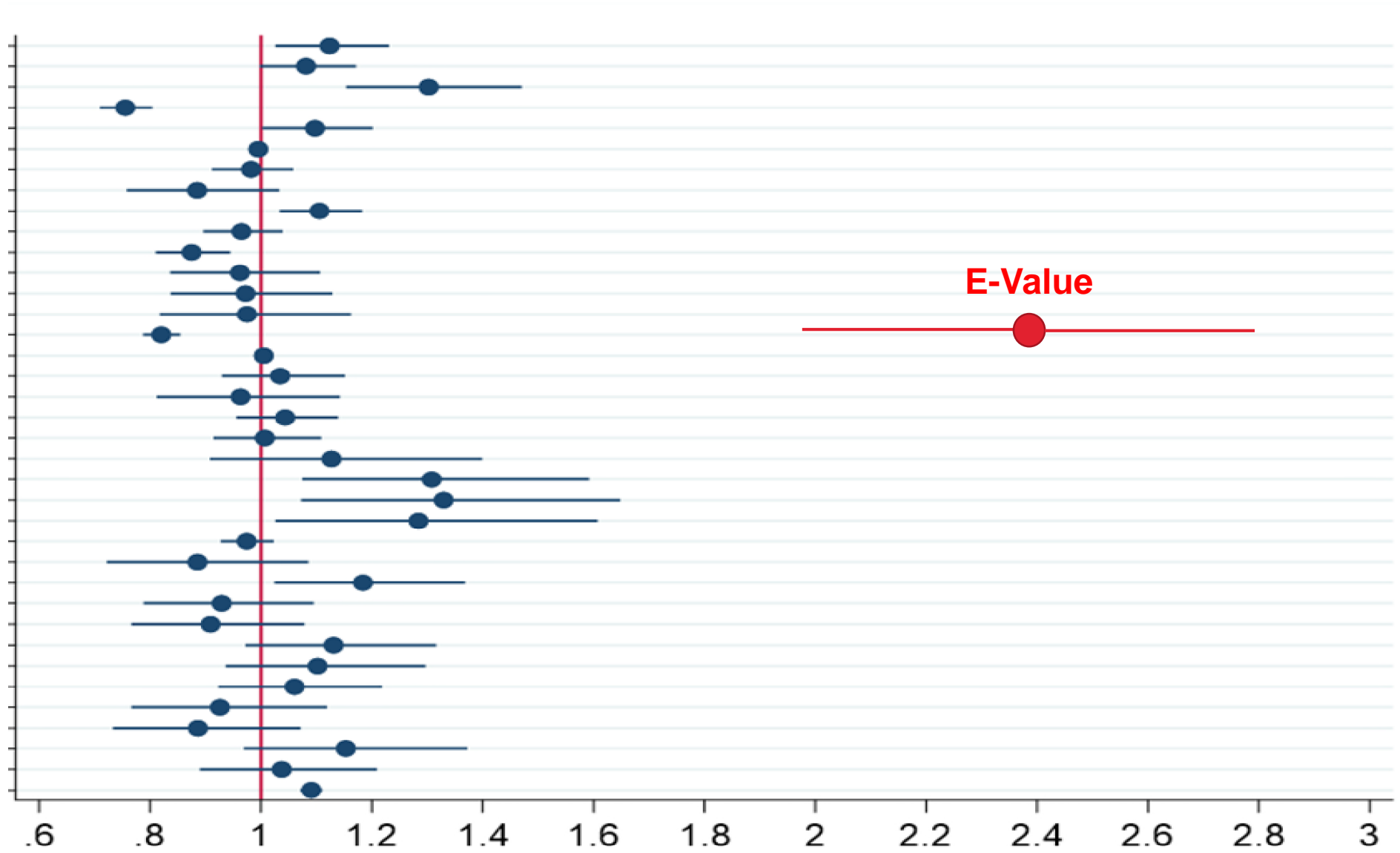
Birth: Birthweight; Maternal smoking during pregnancy; Maternal age; Breastfeeding

Parents: Mother working up to 5; Parents read to child; Parental interest in school; Divorce; Separation from child

SEP: Paternal social class at birth; Financial difficulties; Age mother stayed at school; Housing tenure; Access to amenities; Housing difficulties

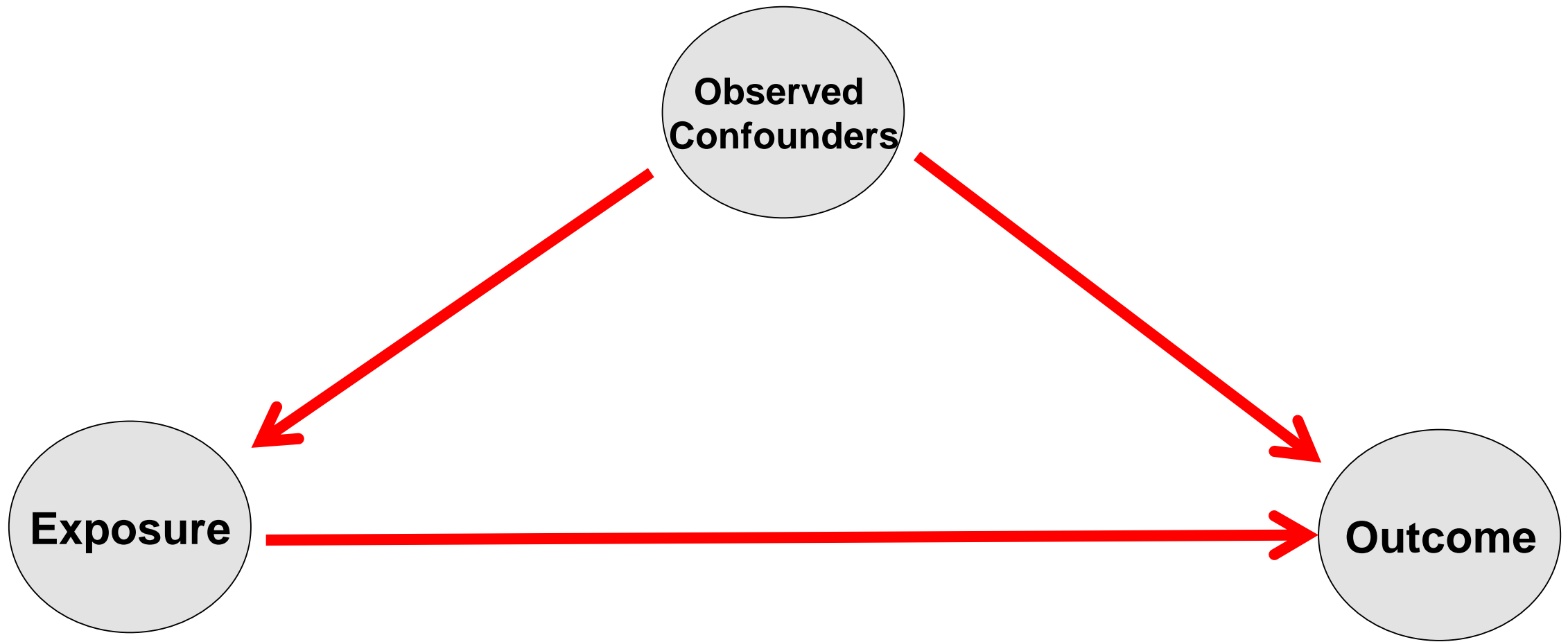
Cohort member: Cognitive ability; Enuresis; Summary of objectively assessed health conditions; BMI

E – Value vs observed confounders



Sensitivity analysis with negative controls

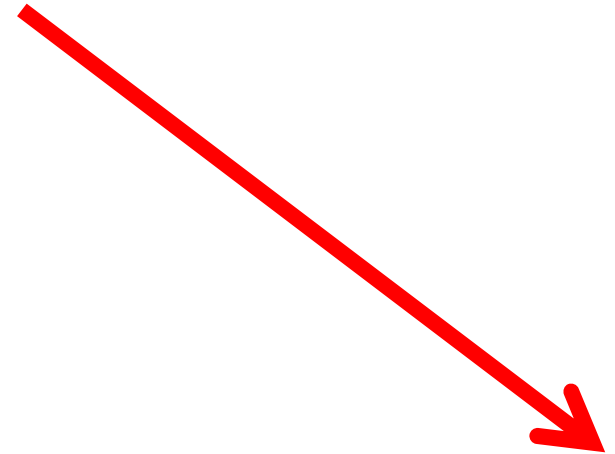
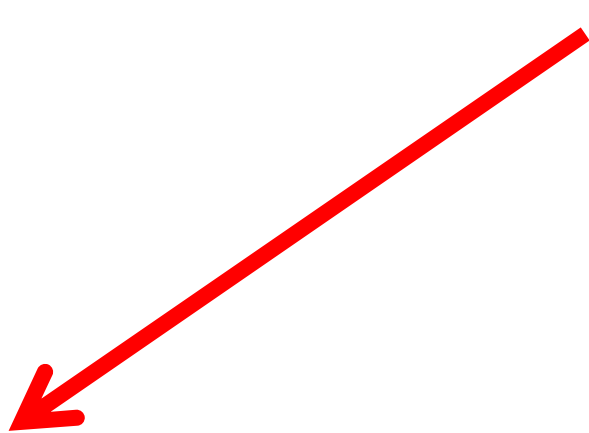
- A simple idea from Biology
- A negative control is a variable (outcome or exposure) for which there is **no plausible Mechanism Of Action (MOA)** other than confounding or measurement error that links it with the actual exposure or outcome in the MOI
- Given the observables there shouldn't be any association with the negative control
- Opportunity to benefit from the richness of the UK longitudinal surveys and probe the assumptions of no unmeasured confounding/omitted variable bias and measurement without error



**Observed
Confounders**

Exposure

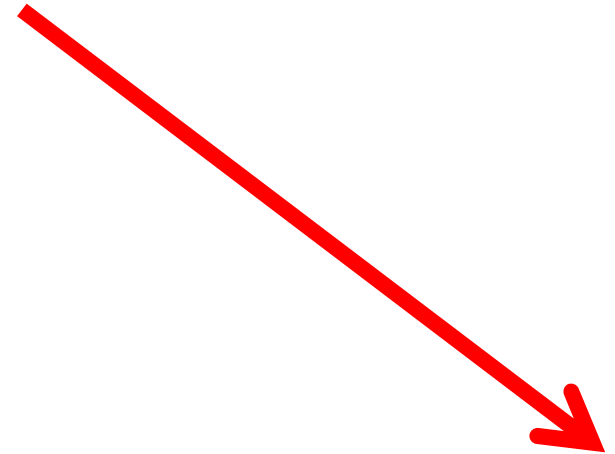
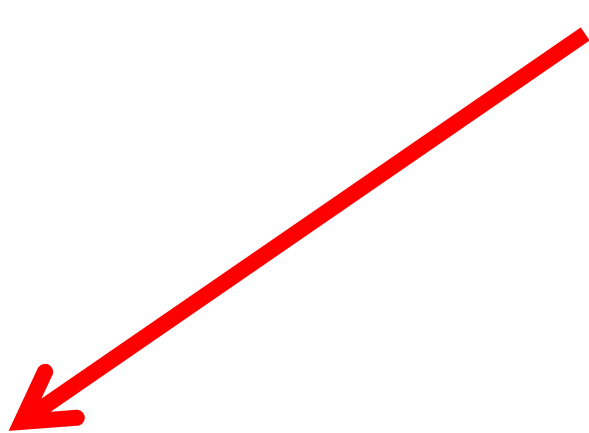
**“Negative”
Outcome**



**Observed
Confounders**

**“Negative”
Exposure**

Outcome



Breastfeeding example, Lawlor et al

“Confounding by socioeconomic position (SEP) is a major concern in associations of breastfeeding with later maternal or offspring outcomes. Therefore we wanted a negative control outcome that is likely to be influenced by SEP but would not plausibly be influenced by being breastfed. We also (ideally) wanted the outcome to be scaled similarly to the real outcome (BMI).”



International Journal of Epidemiology, 2016, 1866–1886

doi: 10.1093/ije/dyw314

Advance Access Publication Date: 20 January 2017

Original article



Approaches to causal inference

Triangulation in aetiological epidemiology

Debbie A Lawlor,^{1,2,*} Kate Tilling^{1,2} and George Davey Smith^{1,2}

Mice & pigeons

Richer families



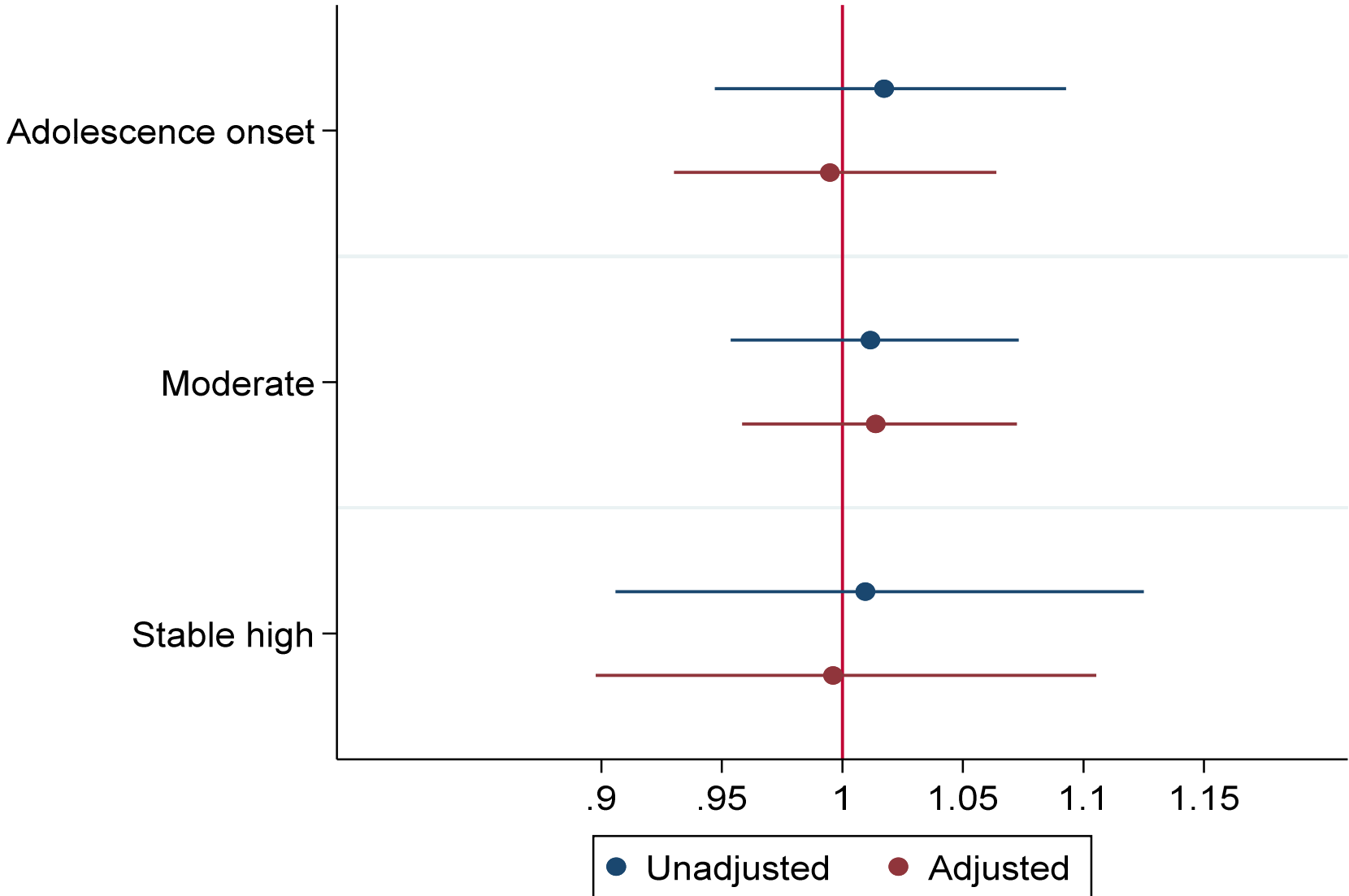
Poorer families



“We found that having been breastfed was inversely associated with obesity at age 7, but it was also inversely associated (with the same magnitude) with parental report of home invasion by pigeons and positively associated (stronger magnitude) with report of invasion by mice”

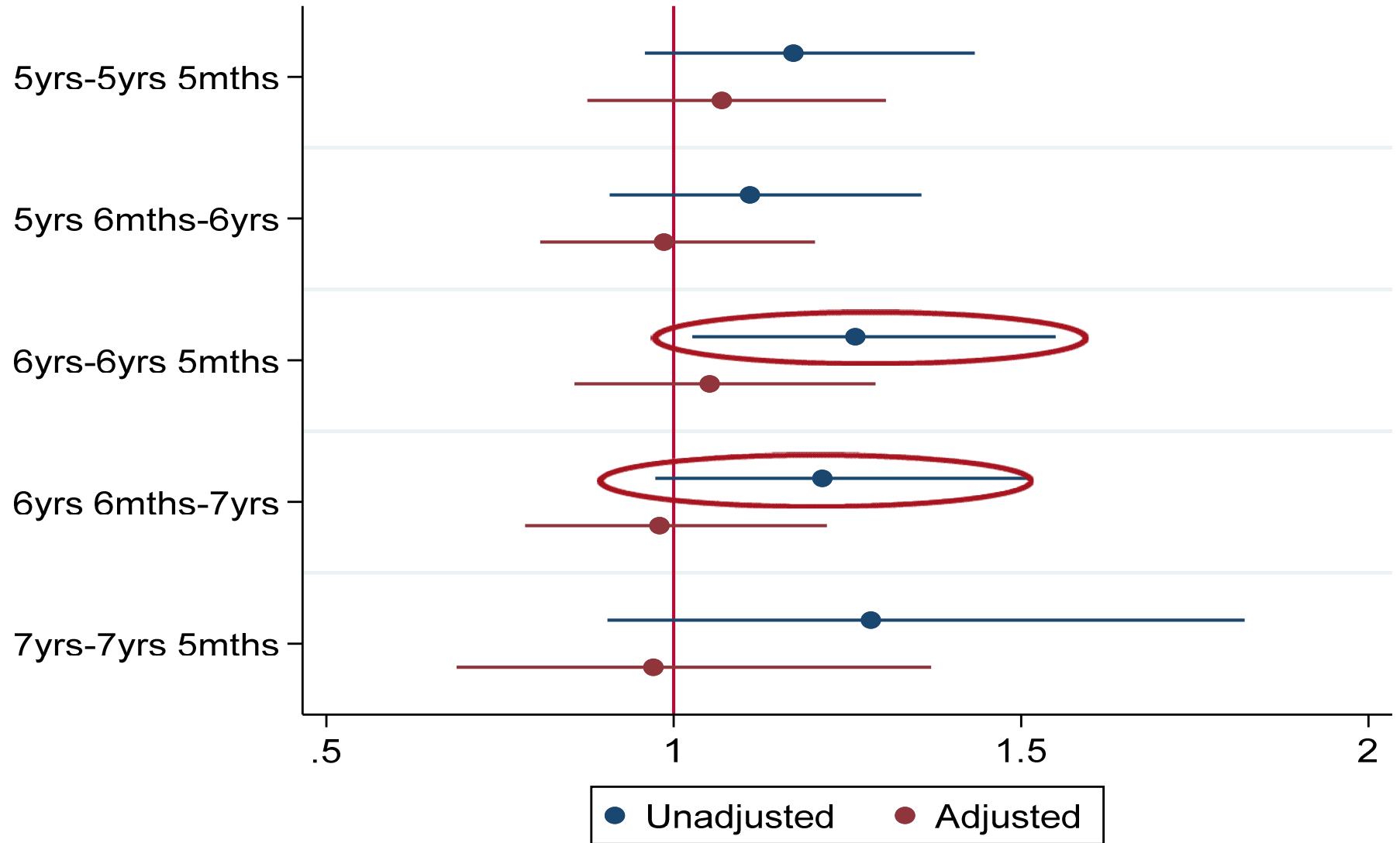
Child mental health example: association with hair colour (negative control outcome)

Reference group:
Stable Low



Negative control exposure: Age started phonetics training

Reference group:
< 5 years 5mths



Conclusion

- Negative controls provide an opportunity for straightforward sensitivity analysis
- If an association is found, researchers can investigate the source(s) of bias
 - ✓ Add more observed confounders in MVA or find another instrument in IVM/MR
 - ✓ Attempt to correct for measurement error
 - ✓ Speculate about the possible source(s) of bias to inform future research
- Useful for future research in other datasets
- Future data collections: suspected unmeasured confounders added in questionnaire/data linkage plans

Missing data

Missing Data

- Selection bias, in the form of incomplete or missing data, is unavoidable in longitudinal surveys
- Smaller samples, incomplete histories, lower statistical power
- **Threat to representativeness**
- Unbiased estimates cannot be obtained without properly addressing the implications of incompleteness
- Statistical methods available to **exploit the richness of longitudinal data** to address bias

Rubin's framework

- A simple Directed Acyclic Graph (DAG)
- Y is an outcome
- X is an exposure (assumed complete/no missing)
- R_Y is binary indicator with $R = 1$ denoting whether a respondent has a missing value on Y

Missing Completely At Random - MCAR



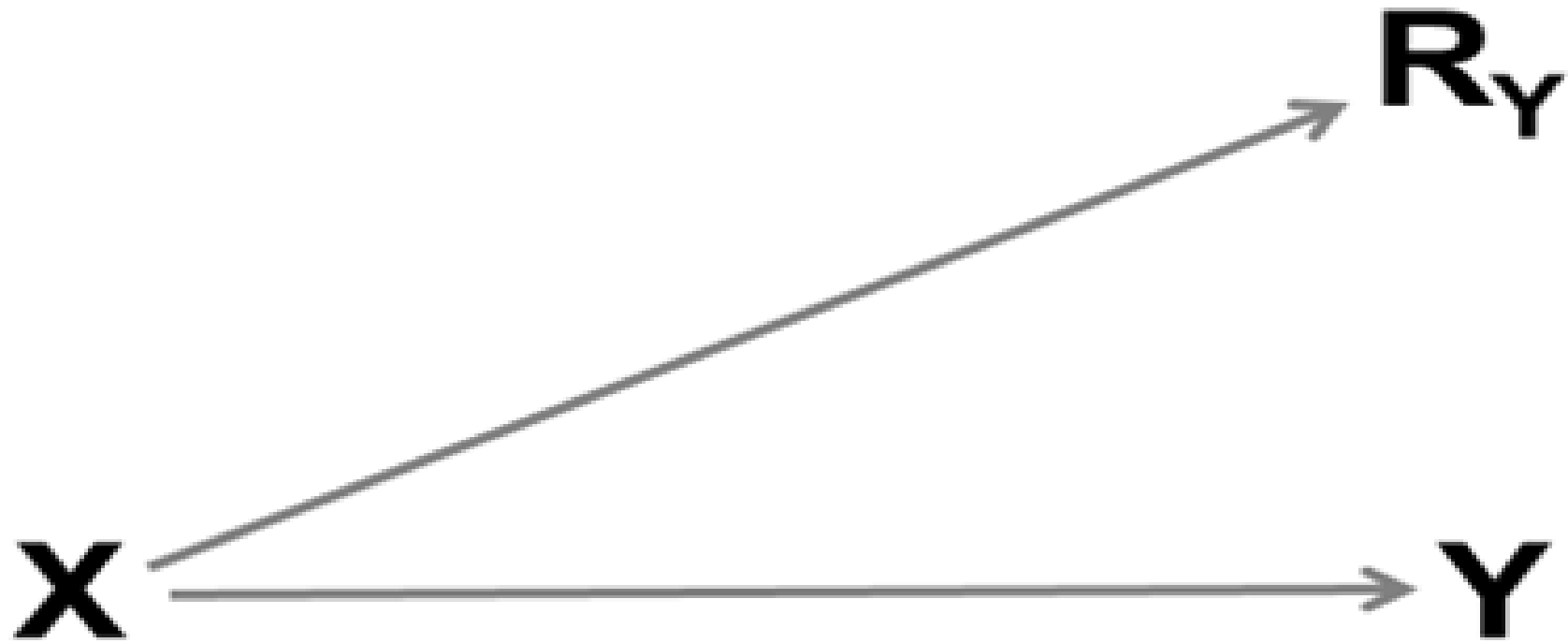
Rubin's framework in the context of longitudinal surveys

- Missing Completely At Random (MCAR): There are no systematic differences between the missing values and the observed values
- Missing At Random (MAR): Systematic differences between the missing values and the observed values can be explained by observed data
- Missing Not At Random (MNAR): Even after accounting for all observed information, differences remain between the missing values and the observed values

Rubin's framework in the context of longitudinal surveys

- Missing Completely At Random (MCAR): There are no systematic differences between the missing values and the observed values – **Never holds in longitudinal surveys**
- Missing At Random (MAR): Systematic differences between the missing values and the observed values can be explained by observed data
- Missing Not At Random (MNAR): Even after accounting for all observed information, differences remain between the missing values and the observed values

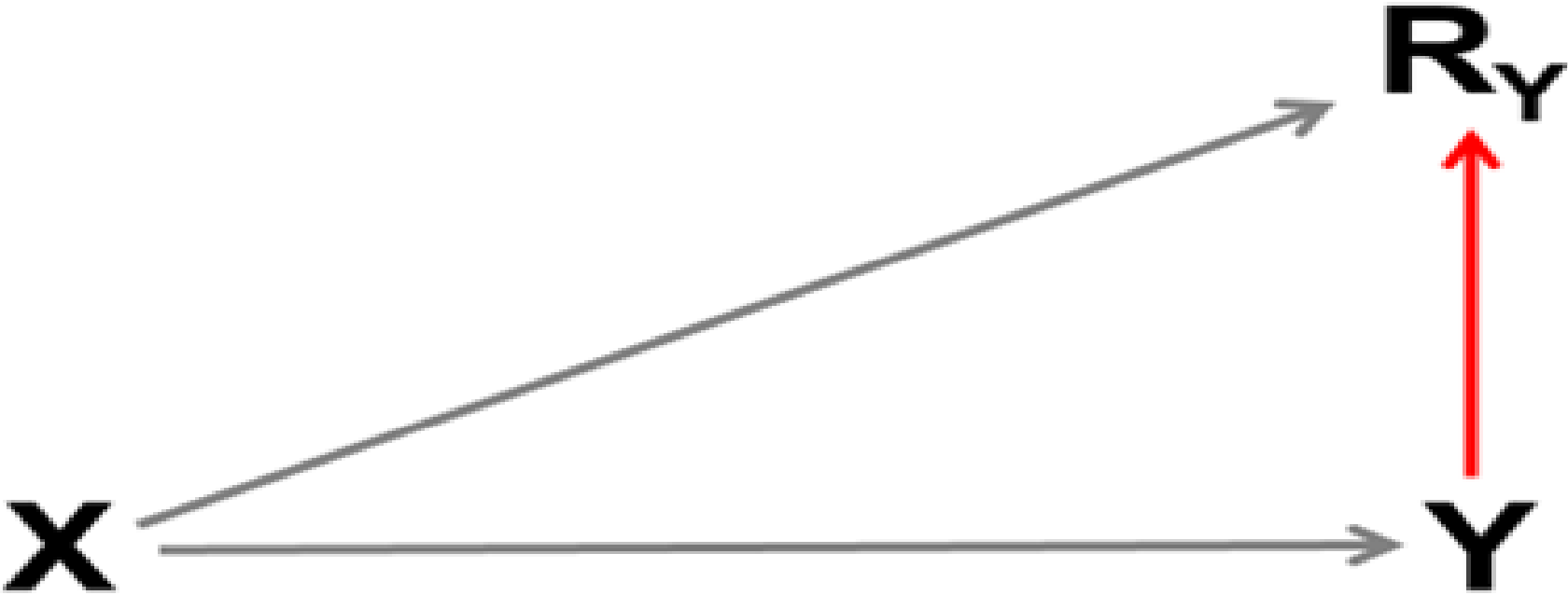
Missing At Random DAG



Rubin's framework in the context of longitudinal surveys

- Missing Completely At Random (MCAR): There are no systematic differences between the missing values and the observed values – **Never holds in longitudinal surveys**
- Missing At Random (MAR): Systematic differences between the missing values and the observed values can be explained by observed data – **Which variables?**
- Missing Not At Random (MNAR): Even after accounting for all observed information, differences remain between the missing values and the observed values

Missing Not At Random - DAG



Rubin's framework in the context of longitudinal surveys

- Missing Completely At Random (MCAR): There are no systematic differences between the missing values and the observed values – **Never holds in longitudinal surveys**
- Missing At Random (MAR): Systematic differences between the missing values and the observed values can be explained by observed data – Which variables?
- Missing Not At Random (MNAR): Even after accounting for all observed information, differences remain between the missing values and the observed values – **Strong distributional assumptions**

Rubin's framework and representativeness

- **MCAR:** No selection, sample is “representative”/balanced
- **MAR:** Observed variables account for selection. Given these, sample is representative/balanced
 - ✓ Can **observables restore/maintain** representativeness?
 - ✓ Does **maximising the plausibility of MAR** help with representativeness?
- **MNAR:** Observed variables do not account for selection (selection is due to unobservables too)

Target population and sample representativeness






- Representative of what? Generalisable where?
- **Any study** (RCT or observational, small or large) that publishes standard errors has a target population
- **Assumptions of generalisability**: are the results transportable to other populations?
- Which populations? Are the assumptions reasonable?
- Missing data analysis is an attempt to **restore sample representativeness to its target population**

Missing data in longitudinal surveys

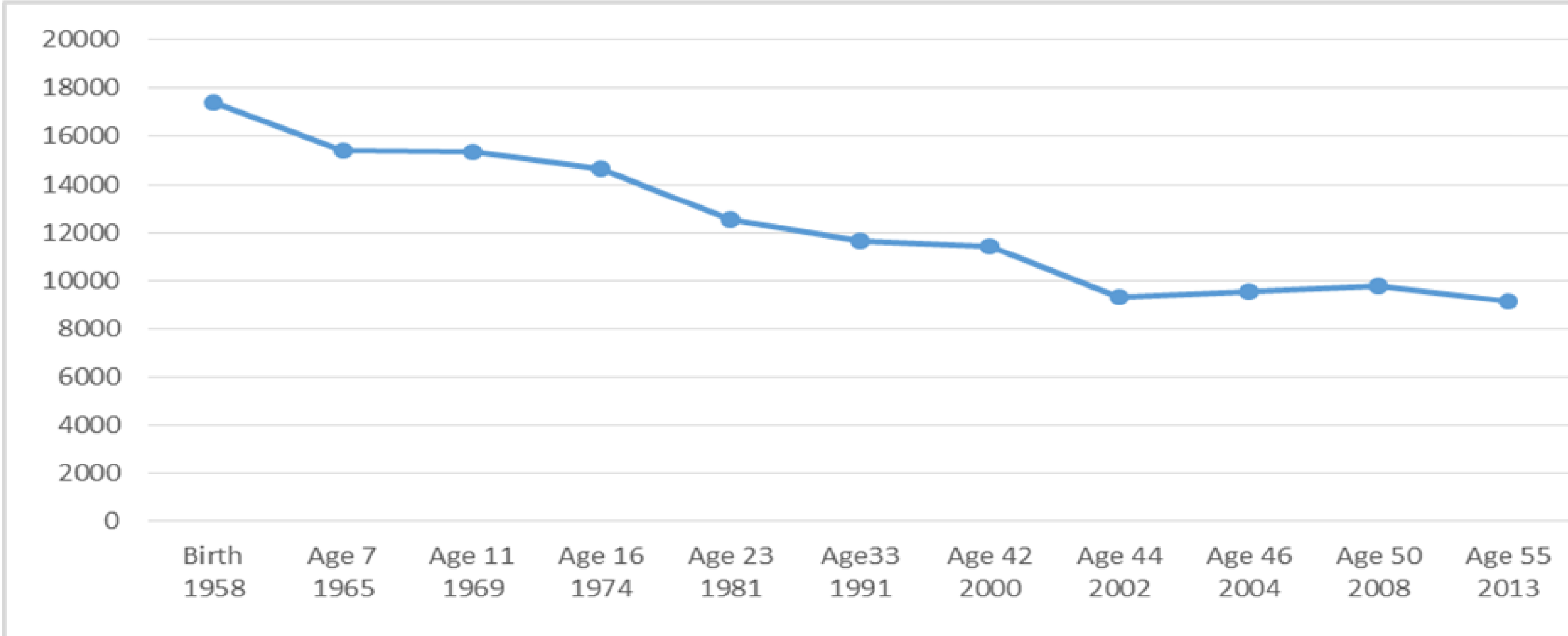
- MAR and MNAR largely untestable
- **Non monotone missing data patterns are more likely to be MNAR** and have implications for the use/derivation of response weights
- We assume that after introducing observables with a principled method (MI, FIML, Fully Bayesian, IPW, Linear Increments) our data are either MAR, or not far from being MAR, so bias is negligible
- **Reasonable assumption**
 - ✓ Richness of longitudinal data
 - ✓ MAR methods have been shown to perform well even when data are MNAR
- Arguably MAR methods **more suitable** than MNAR methods in **rich longitudinal studies**

The National Child Development Study (NCDS -1958 cohort)

CENTRE FOR
LONGITUDINAL
STUDIES

	1958 Birth	1965 7	1969 11	1974 16	1981 23	1991 33	2000 42	2003 45	2004 46	2008 50	2013 55
 main respondent	mother	parent	parent	cohort member / parent	cohort member	cohort member	cohort member	cohort member	cohort member	cohort member	cohort member
 secondary respondent	medical	school medical	school medical	school medical		partner mother children			medical		
 survey instruments		cognitive tests	cognitive tests	cognitive tests						cognitive tests	
 linked data					exams					consents	
 response	17,415	15,425	15,337	14,654	12,537	11,469	11,419	9,377	9,534	9,790	9,137

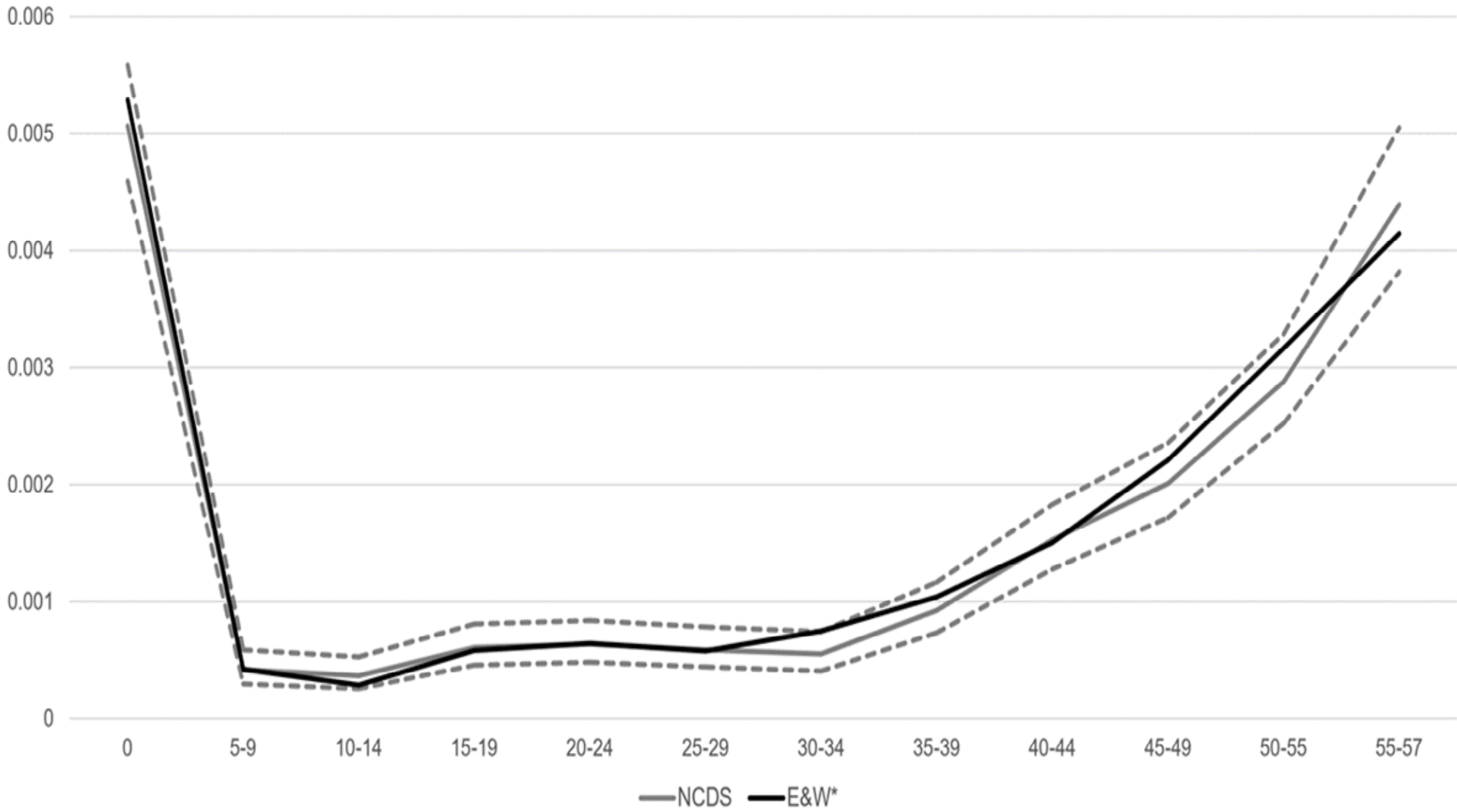
Response in NCDS



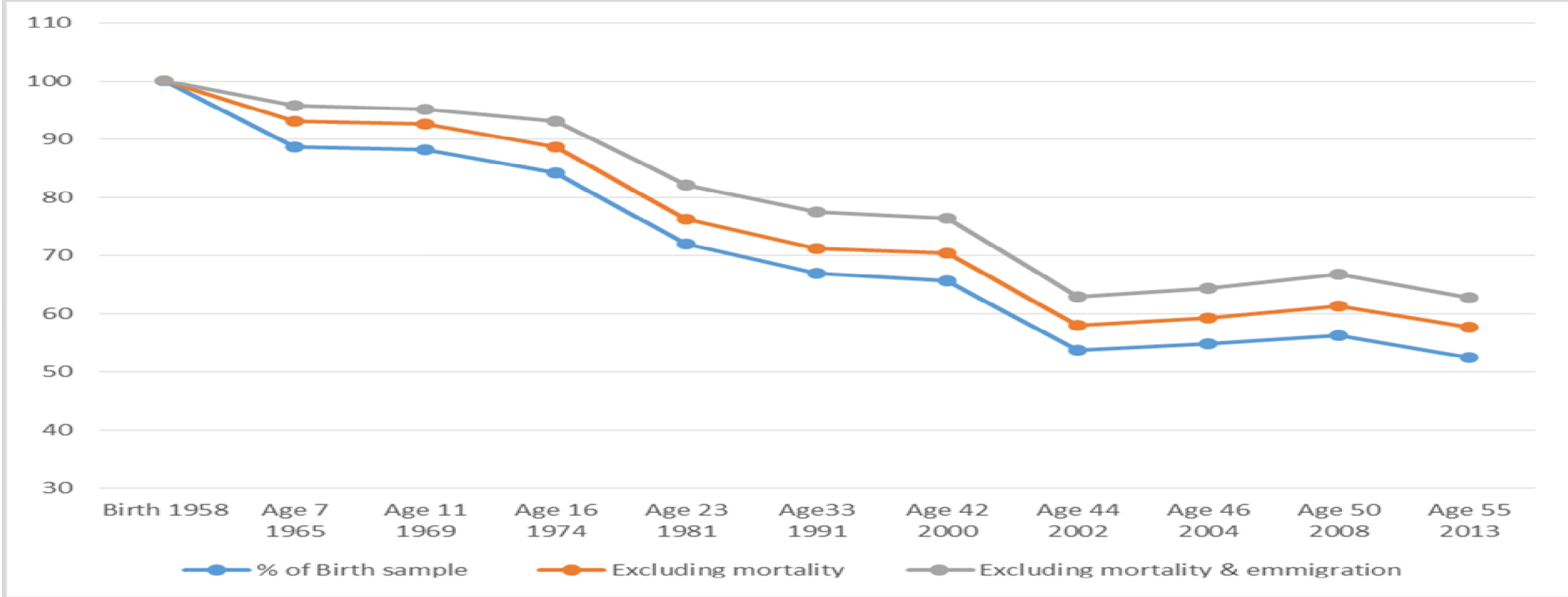
Non response in NCDS

Types of non-response	Wave 0	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8	Wave 9
Age	Birth	7	11	16	23	33	42	46	50	55
Non-contact	223	1,042	410	786	1,867	1,529	1,832	612	835	664
Not issued	920	542	271	0	0	0	1,415	4,248	3,553	4,698
Refusal	0	80	797	1,151	1,160	1,776	1,148	1,448	1,214	582
Other unproductive	0	173	202	295	838	1,399	263	109	332	491
Not issued - emigrant	0	475	701	799	1,196	1,335	1,268	1,272	1,293	1,287
Not issued - dead	0	821	840	873	960	1,050	1,200	1,324	1,460	1,503
Ineligible	0	0	0	0	0	0	13	11	81	0
Total	1,143	3,133	3,221	3,904	6,021	7,089	7,139	9,024	8,768	9,225





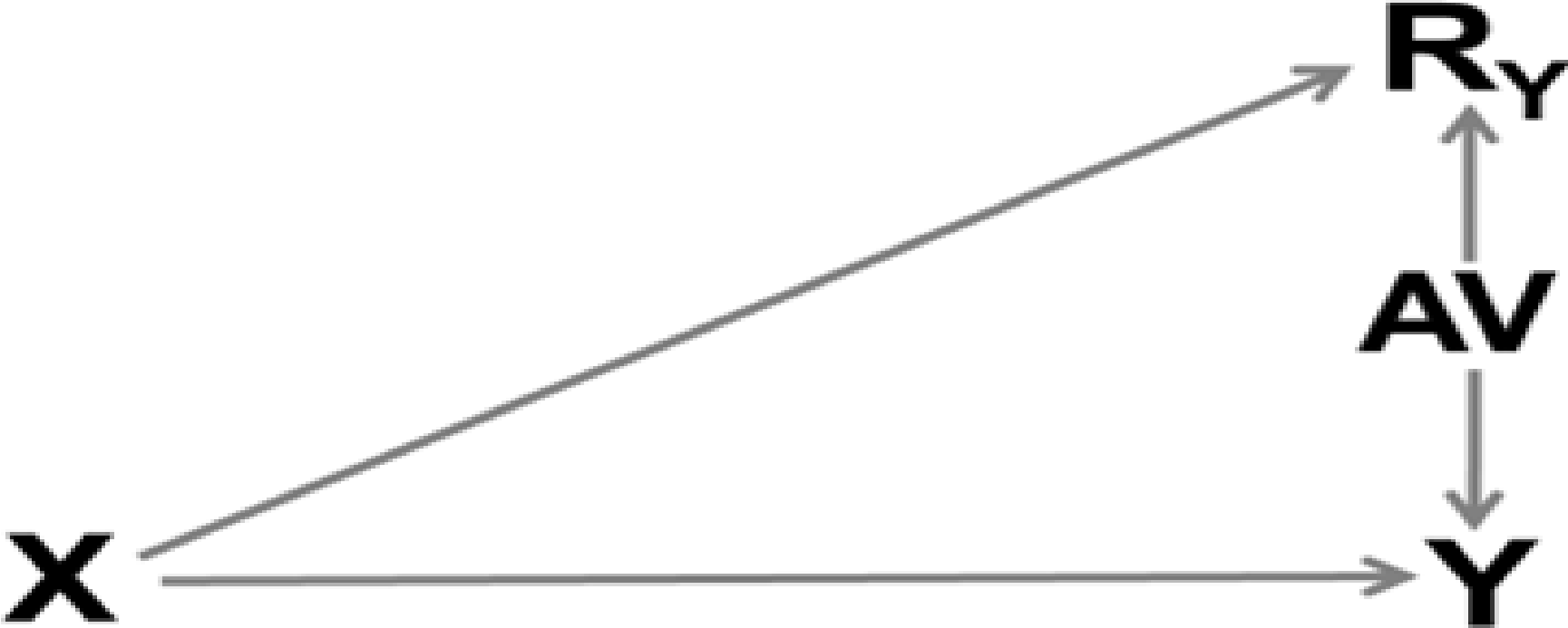
Sample size in the 1958 cohort as % of the original sample



CLS Missing Data Strategy

- A simple idea
- **Data driven approach** to maximise the plausibility of the MAR assumption by exploiting the richness of longitudinal data
- In longitudinal surveys the information that maximises the plausibility of MAR is finite – the information **that matters in practice** can be at least approximated
- We can identify the variables that are associated with non response/attrition
- **Auxiliary variables** – to be used **in conjunction** with variables in the substantive model/Model of Interest (MoI)
- Substantive interest in **understanding the drivers of non response**
- Age, period and cohort effects

How to turn MNAR into MAR (or at least attempt to)



Variables in
NCDS up to 50

“Eligible” vars

Routed
Binary <1%,
Missing > 50%
Summary variables
Similar concepts

17,412

587

Age 0 Age 7 Age 11 Age 16 Age 23 Age 33 Age 42 Age 44 Age 46 Age 50

Stage 1 Input

25	52	53	58	38	77	105	52	67	89
----	----	----	----	----	----	-----	----	----	----

Stage 1 Output

12	22	16	15	15	18	16	14	7	11
----	----	----	----	----	----	----	----	---	----

Non-Response

Age 7 Age 11 Age 16 Age 23 Age 33 Age 42 Biomed Age 46 Age 50 Age 55

Stage 2 Input

6	16	19	35	39	31	68	56	55	70
---	----	----	----	----	----	----	----	----	----

Stage 2 Output

2	4	5	13	11	16	15	27	19	23
---	---	---	----	----	----	----	----	----	----

Strong predictors of participating in NCDS

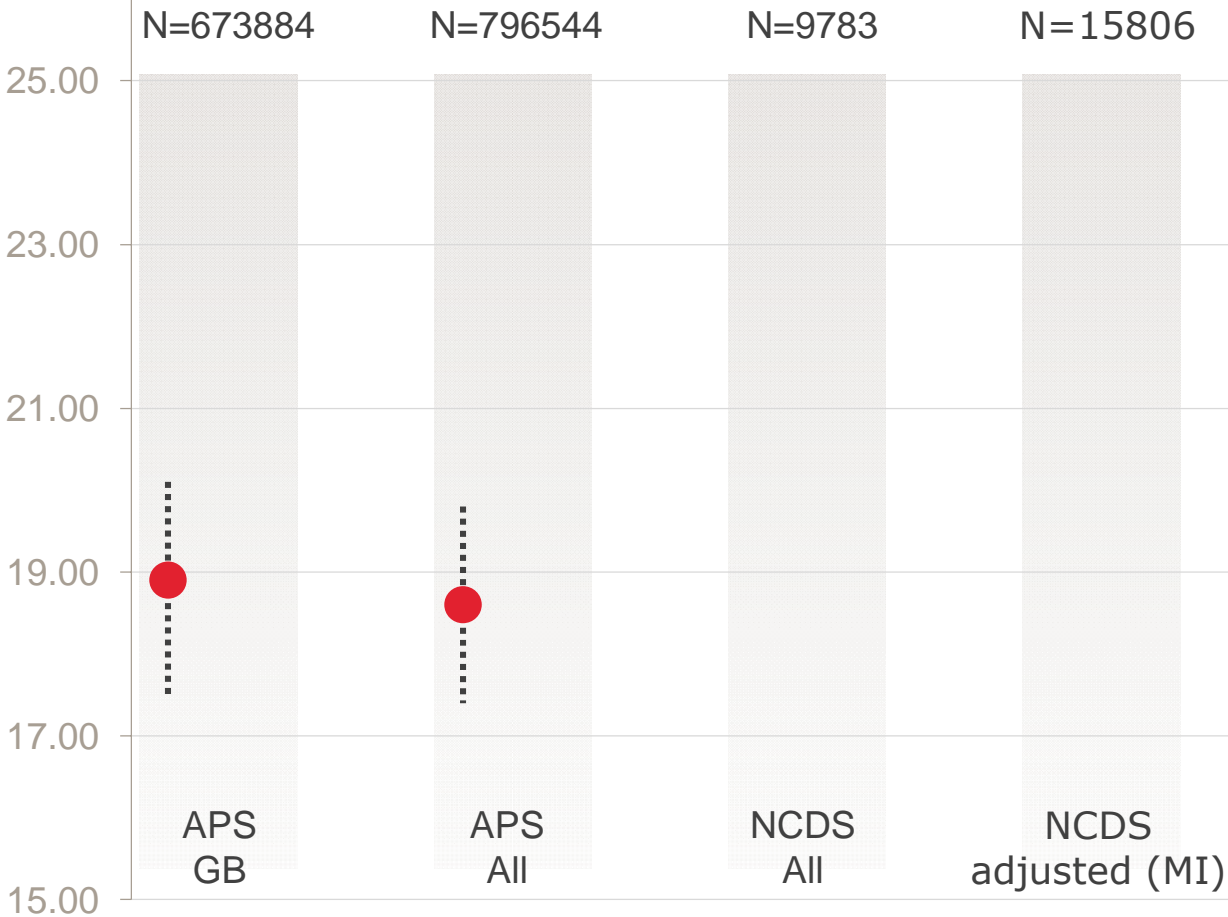
- Voting
- Union membership
- Early life mental health
- Social class – Advantaged background
- Early life cognitive ability

“Experiment” – Education at 50

- “Known” population distributions from the Annual Population Survey – Office for National Statistics
- Multiple Imputation (MI) with chained equations in Stata v15: 20 imputations
- Can we replicate the “known” population distribution of educational qualifications using NCDS after handling missing data with MI?
- Known population distribution (ONS Annual Population Survey data) represents those born in 1958 in GB

Education at age 50

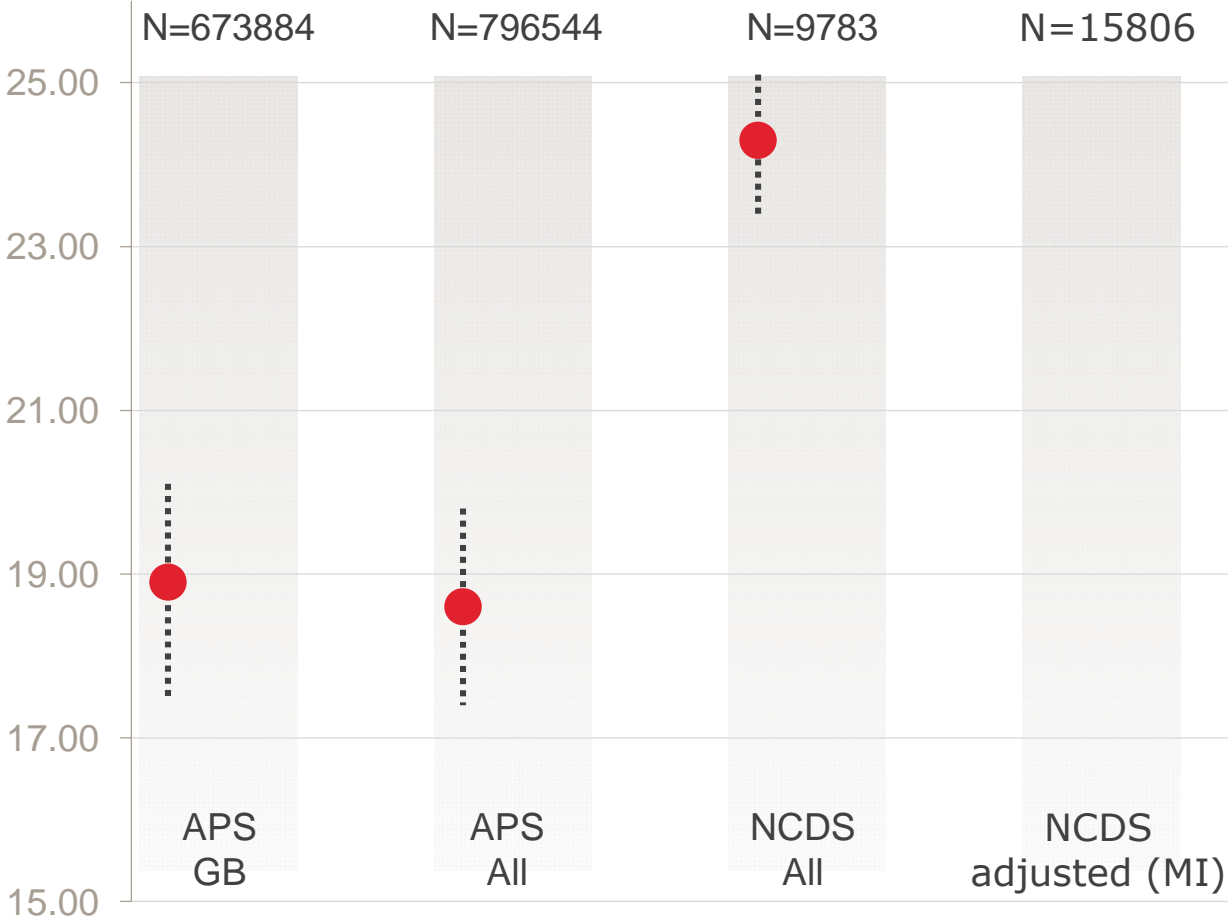
Percentage in Category 1: Degree or equivalent



● Degree or equivalent

Education at age 50

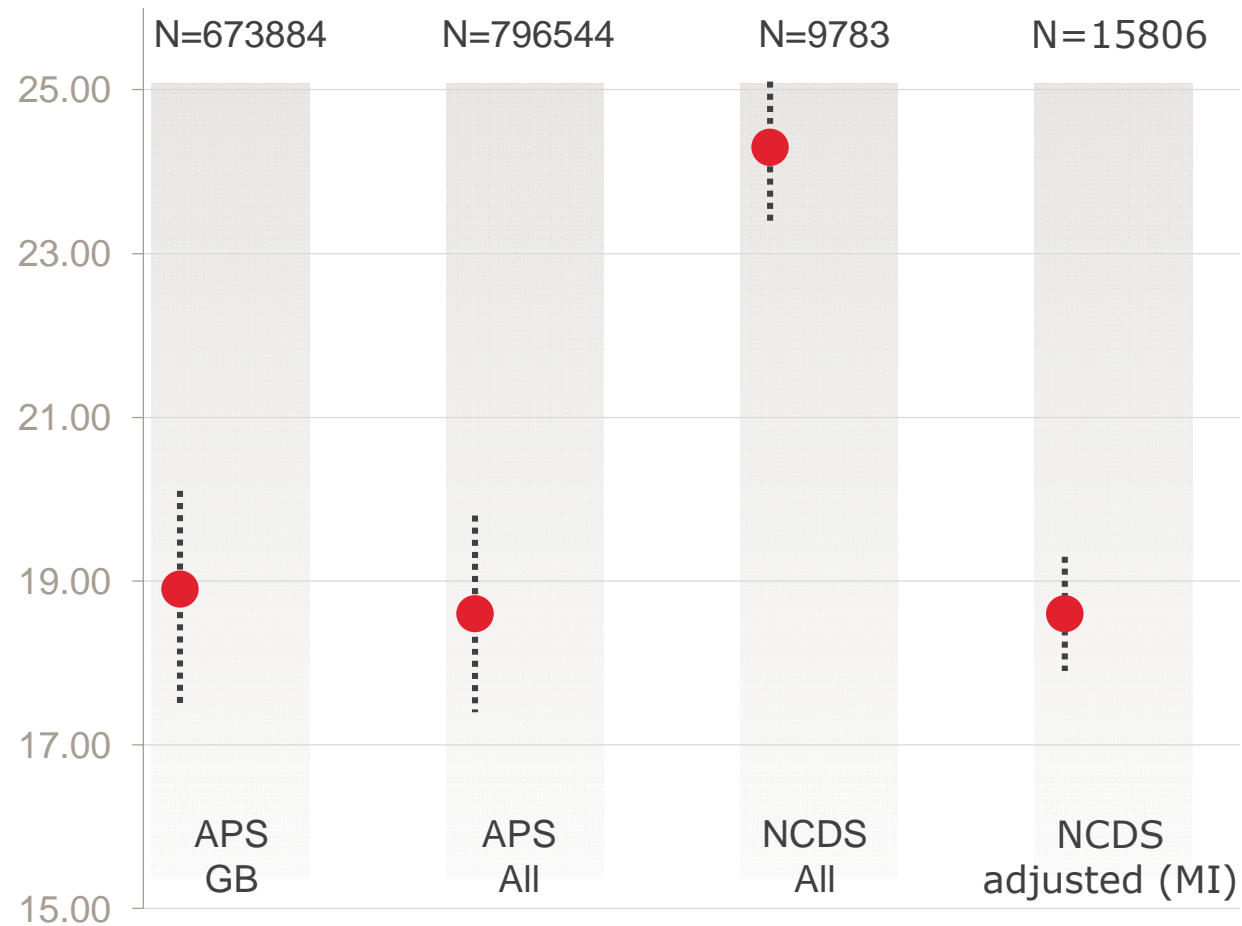
Percentage in Category 1: Degree or equivalent



● Degree or equivalent

Education at age 50

Percentage in Category 1: Degree or equivalent



● Degree or equivalent

Conclusion

- In NCDS maximising the plausibility of the MAR assumption with observed data has the potential to **restore/maintain sample representativeness**
- **Reasonable approximation** of contemporary population totals
- Reassuring for substantive research as bias is reduced
- **Not** a test for MAR vs MNAR
- Bias due to missing data still possible

Outputs

- User guide for missing data analysis & list of auxiliary variables for users to **adapt** to their analysis – [available end of October 2019!](#)
- Peer reviewed papers
- **Dynamic process**, the results will be updated when new waves or other forms of data become available (administrative data for example)
- Handling Missing Data in British Birth Cohorts Training: [December 2nd 2019](#)
- User guide with causal inference examples: [Available in February 2020!](#)
- Measurement error in mental health and cognitive ability measures in the British birth cohorts reports: [Available in November 2019](#)

Thank you for your attention!

G.Ploubidis@UCL.ac.uk

 [@GeorgePloubidis](https://twitter.com/GeorgePloubidis)