

# Missing data theory

George B. Ploubidis

# 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



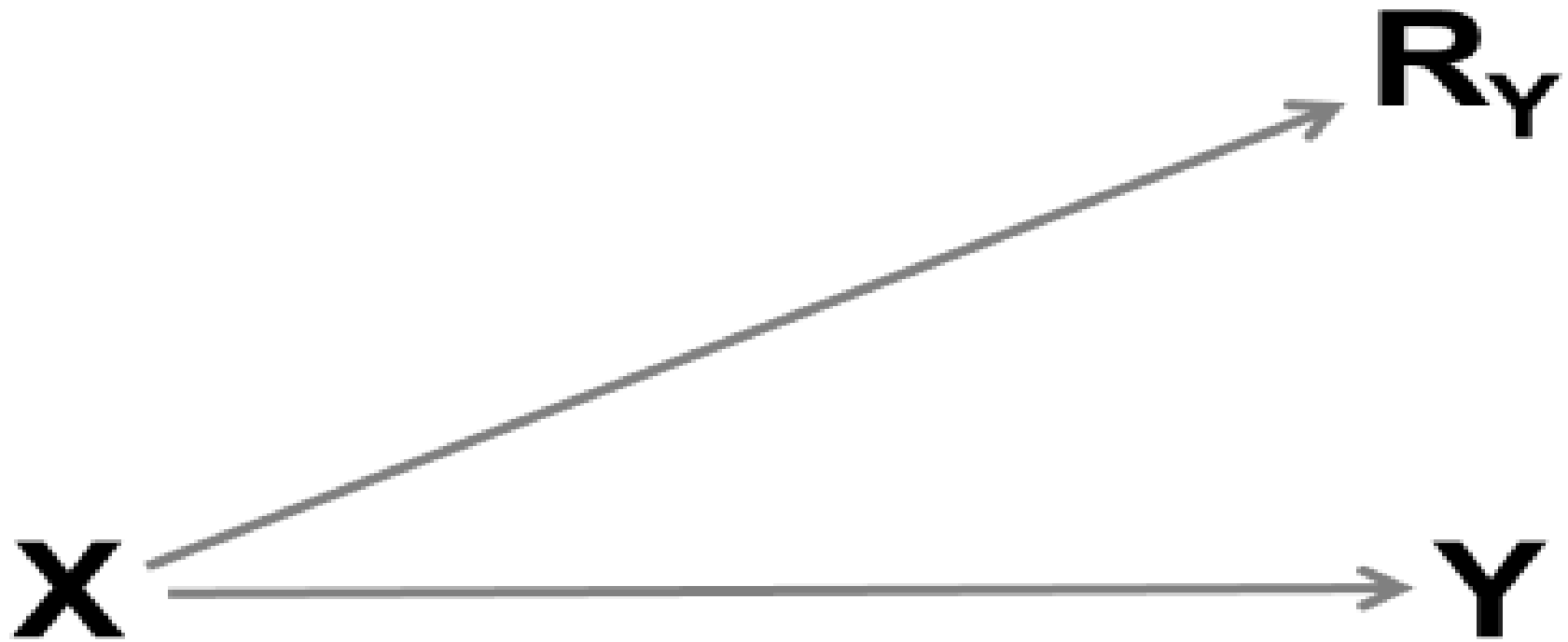
# Rubin's framework in the context of longitudinal surveys

- 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

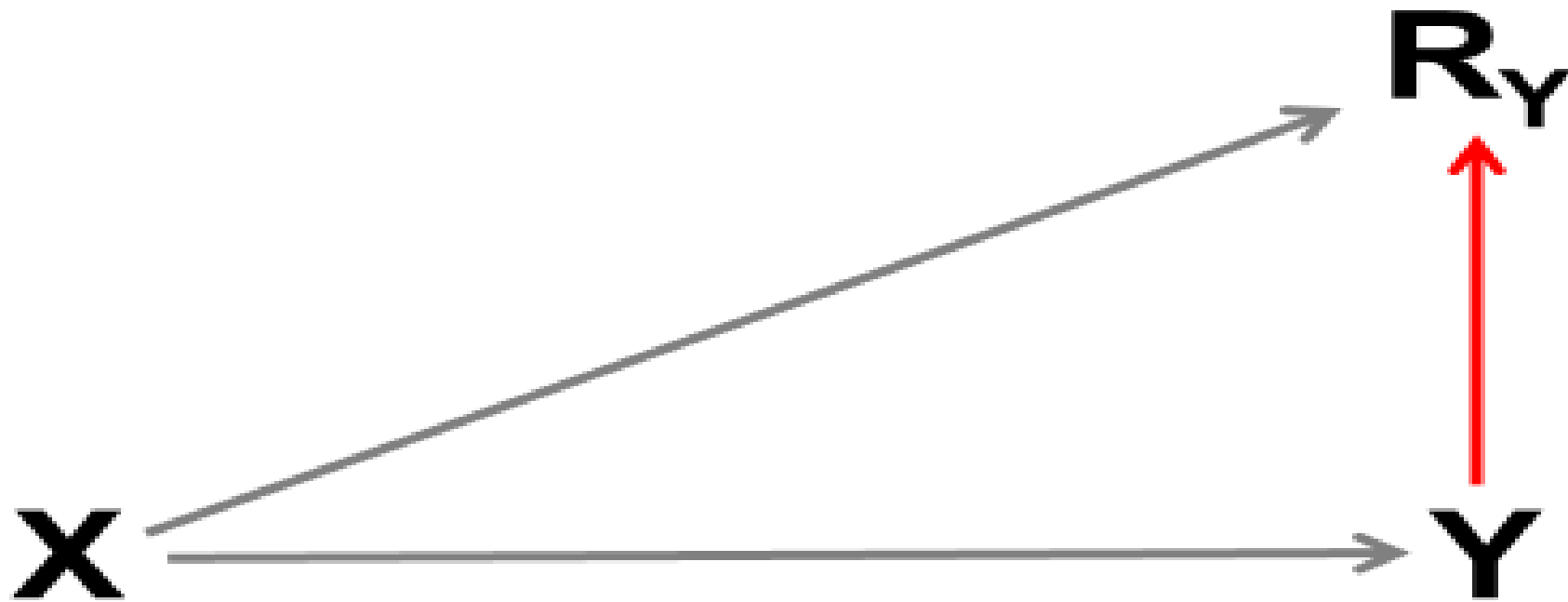


# Rubin's framework in the context of longitudinal surveys

- 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)

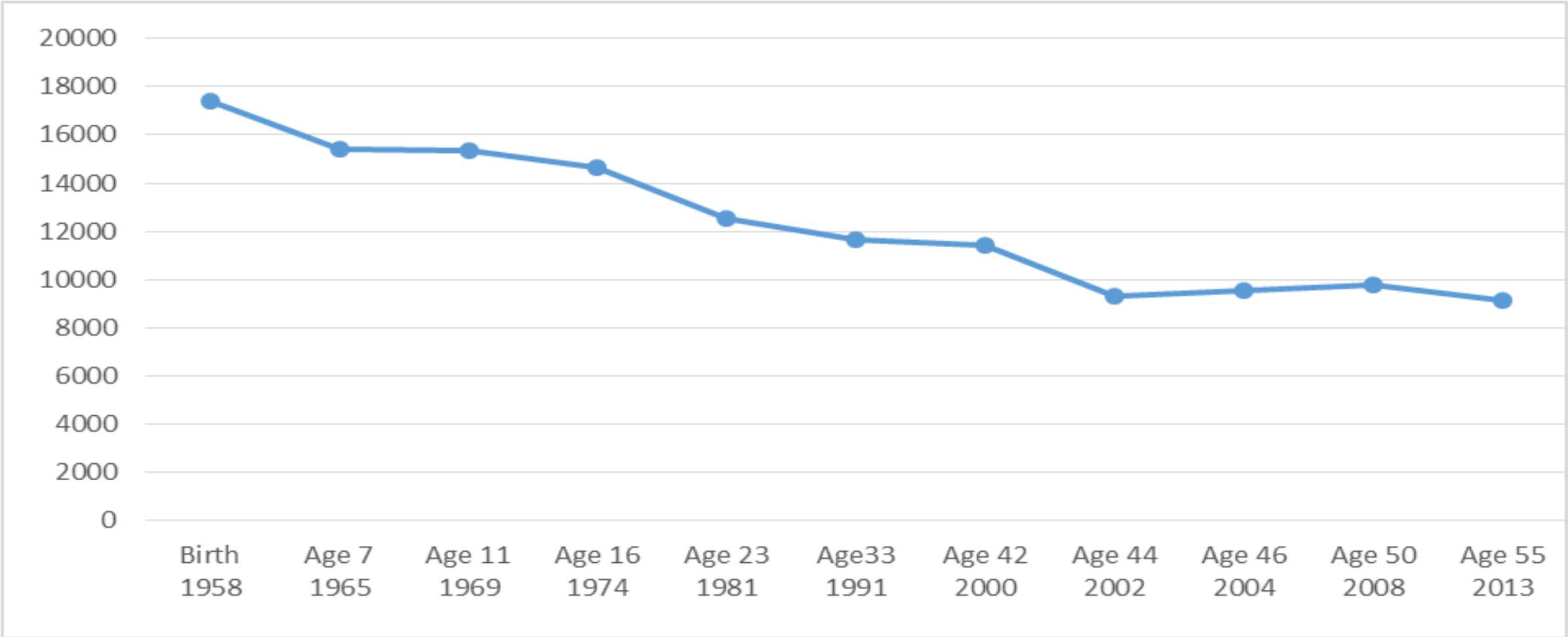
# Target population and sample representativeness

- 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**

# The National Child Development Study (NCDS -1958 cohort)

	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

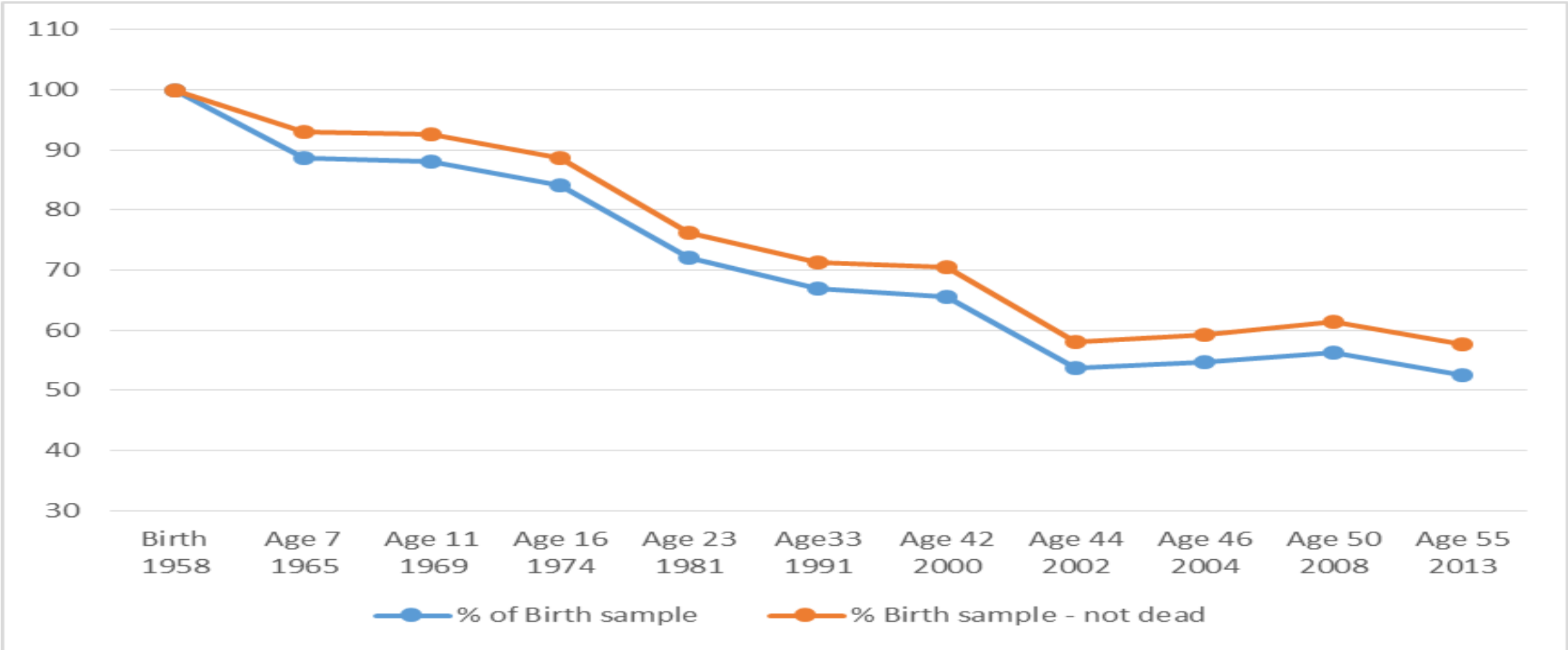
# Response in NCDS



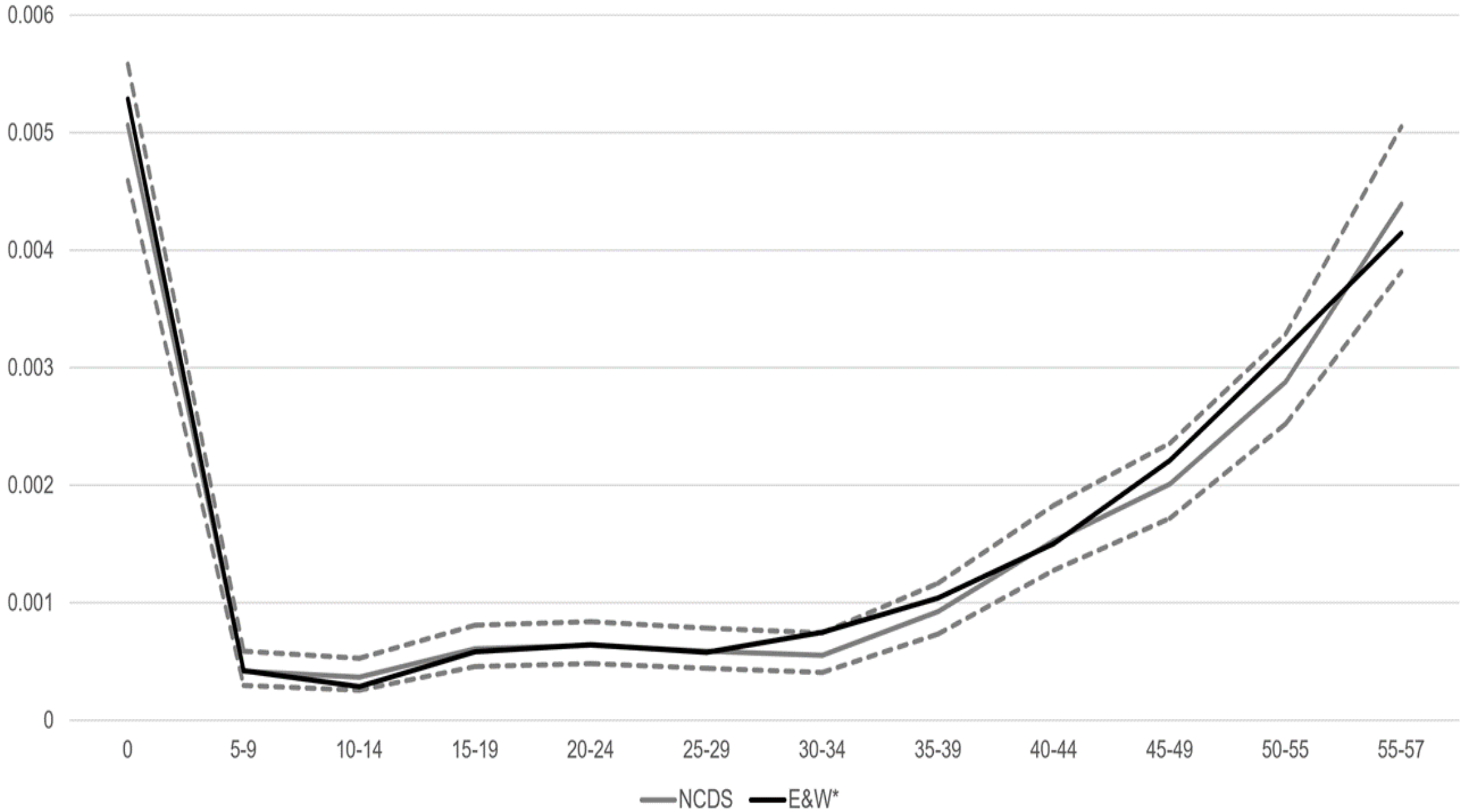
# Non response in NCDS

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

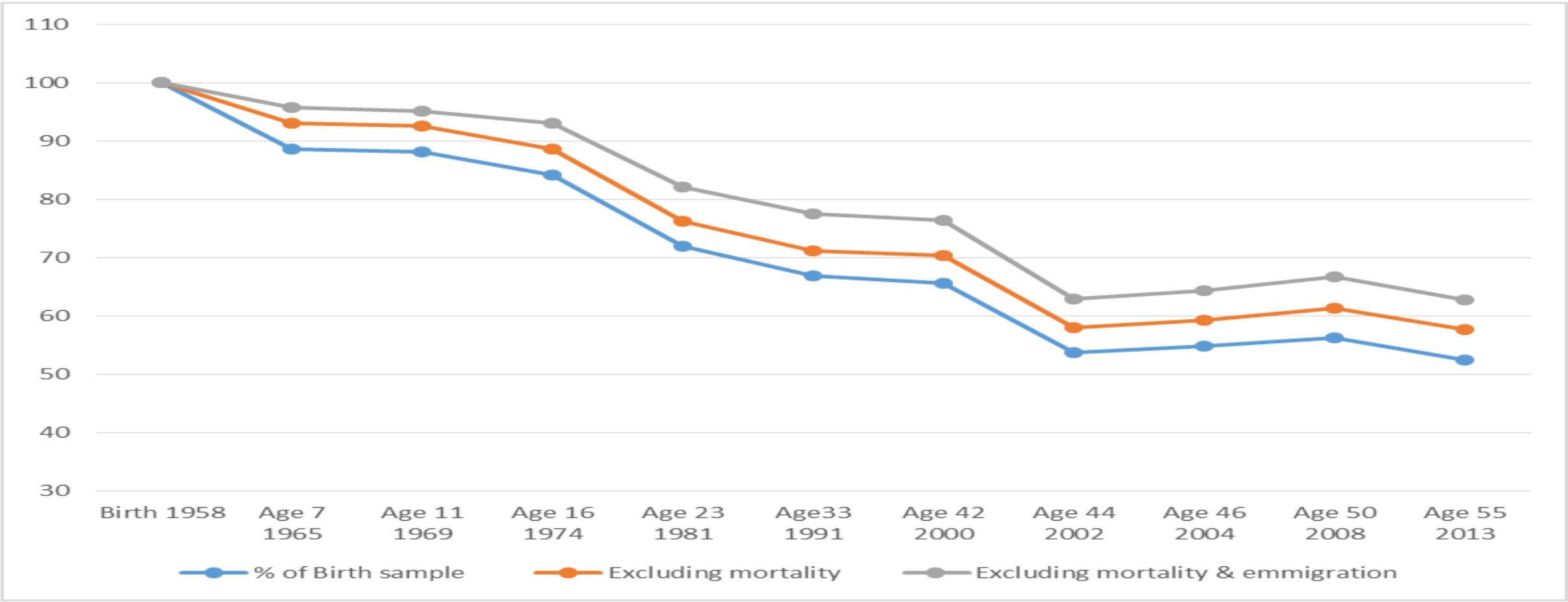
# Sample size in the 1958 cohort as % of the original sample







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



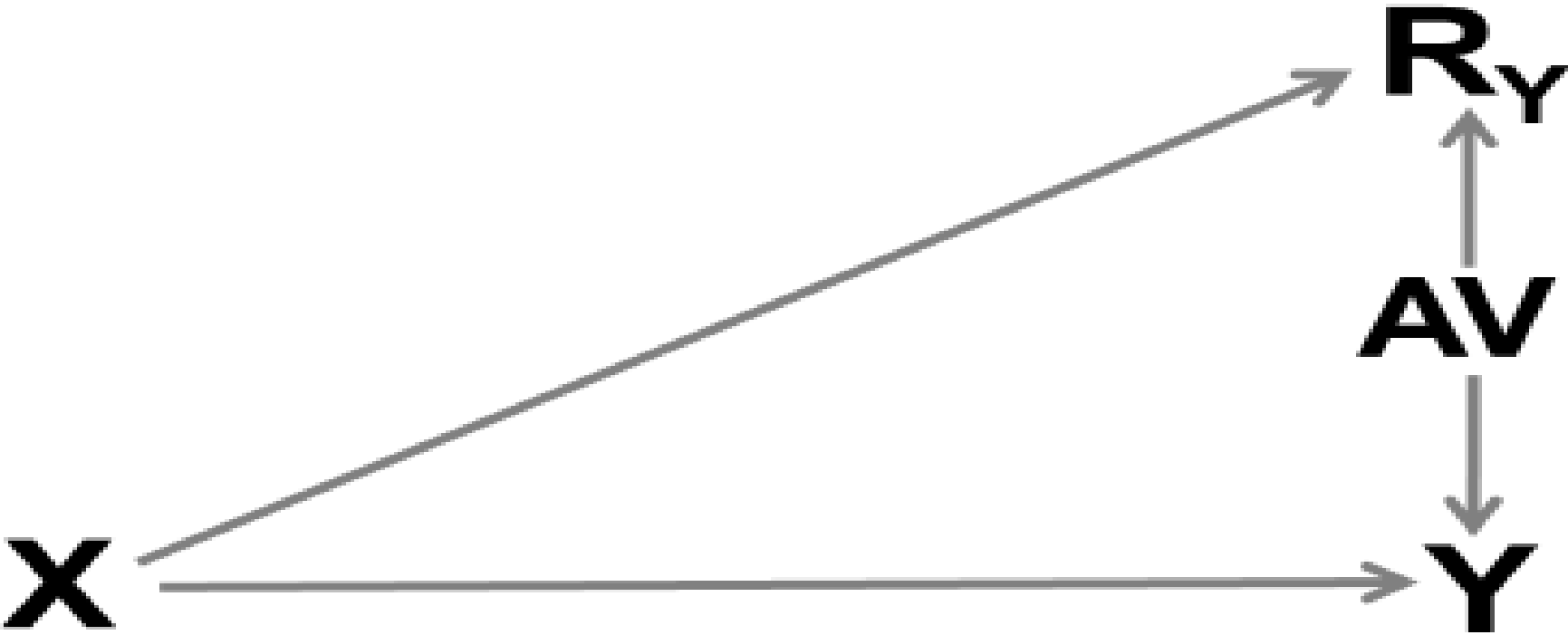
# Missing data in longitudinal surveys

- 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**

# 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**
- **Generational differences** in the drivers of non-response

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



# Outputs

- **User guides** for missing data analysis & list of auxiliary variables for users to **adapt** to their analysis
- Working papers/Peer reviewed papers
- **Dynamic process**, the results will be updated when new waves or other forms of data become available (administrative data for example)
- Training



[Home](#) [Data access and training](#)

## Handling missing data

We know different types of people tend to drop out of our studies at different times, depending on their individual circumstances and characteristics. To support researchers in producing robust analysis, we have developed comprehensive advice on how to deal with missing data. The approaches we recommend to researchers capitalise on the rich data cohort members provided over the years before their non-response. These include well known methods such as multiple imputation, inverse probability weighting, and full information maximum likelihood.

### WORKING PAPERS

## A data driven approach to understanding and handling non-response the Next Steps cohort – CLS working paper 2020/5



This paper presents a systematic data-driven approach to identify predictors of non-response at wave 8 (age 25-26 years) in Next Steps and demonstrates that including such variables in analyses with principled methods can reduce bias due to missing data.

Author: Richard J. Silverwood, Lisa Calderwood, Joseph W Sakshaug and George B. Ploubidis

Date published: 27 April 2020

ORIGINAL ARTICLE | [VOLUME 136, P44-54, AUGUST 01, 2021](#)

## Missing at random assumption made more plausible: evidence from the 1958 British birth cohort

[Tarek Mostafa](#) • [Martina Narayanan](#) • [Benedetta Pongiglione](#) • [Brian Dodgeon](#) • [Alissa Goodman](#) • [Richard J. Silverwood](#)   • [George B. Ploubidis](#) • [Show less](#)

Published: February 26, 2021 • DOI: <https://doi.org/10.1016/j.jclinepi.2021.02.019> •



Check for updates

Thank you for your attention!

[G.Ploubidis@UCL.ac.uk](mailto:G.Ploubidis@UCL.ac.uk)

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