

Compensating for Missing Data in Longitudinal Surveys Workshop

George B. Ploubidis

Outline

- Rubin's classification
- Data driven approach
- Results from the 1958 cohort – Work in progress
- Relevance to users of CLS data
- Outputs

CLS Applied Statistical Methods

- Applied methodological 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

Missing data

- Selection bias, in the form of incomplete or missing data, is unavoidable in longitudinal surveys
- Smaller samples, incomplete histories, lower statistical power
- 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



Missing Completely At Random - MCAR

- There are no systematic differences between the missing values and the observed values
- There isn't any association between observed or unobserved variables and non response
- Partially testable, since we can find out whether variables available in our data are associated with missingness
- However, if we fail to find such associations, we cannot be certain that unmeasured variables are not associated with the probability of non-response

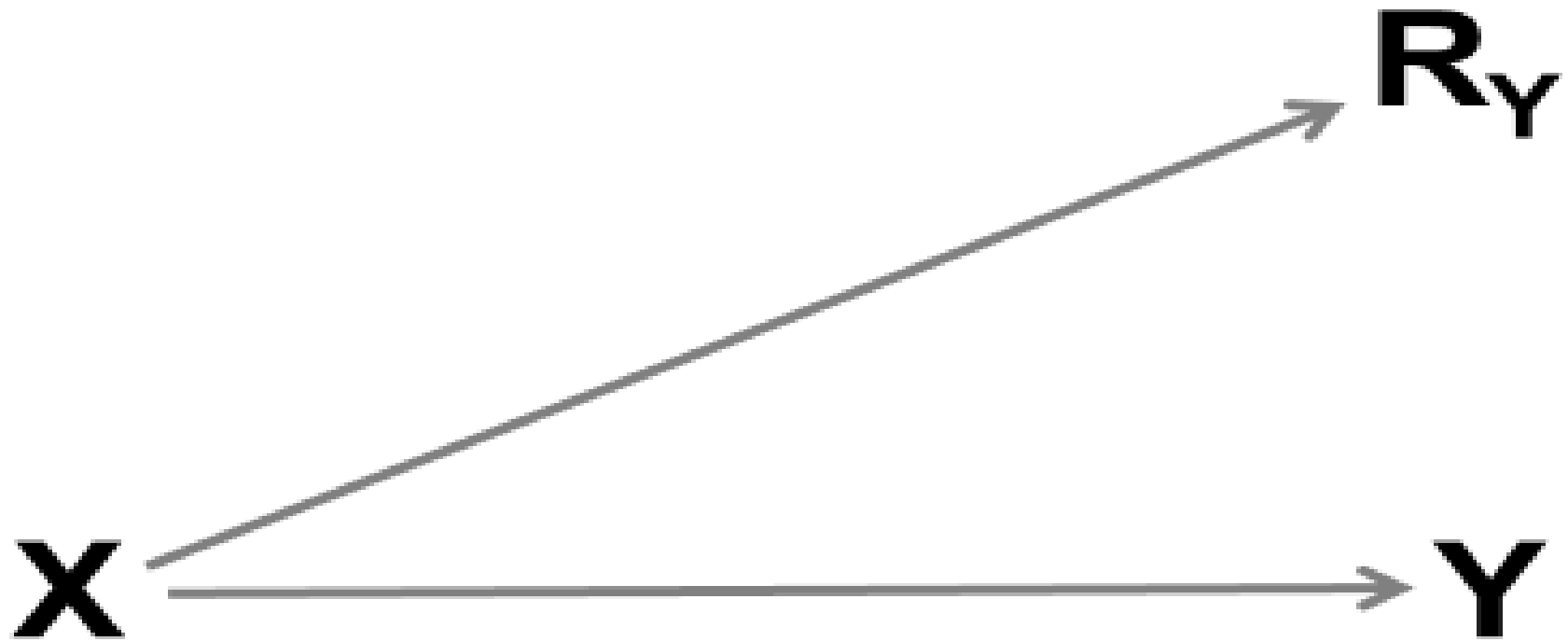
Complete Case Analysis (CCA)

- OK when not much missing data (<10%) or FICO <10%
- Valid under MCAR – Assumes that complete records do not differ from incomplete
- But! CCA can be unbiased (but obviously less efficient) in specific scenarios even if the complete records are systematically different - not true that CCA is always biased if data are not MCAR
- Outcome Y missing/**complete exposures X**, probability of missing on Y independent of observed values of Y, given the exposures/covariates X in the substantive model
- Exposure X missing/**outcome Y complete**, probability of missing on X independent of Y or observed values of X, given the exposures/covariates X in the substantive model

Complete Case Analysis (CCA)

- In longitudinal studies, usually there are incomplete records on both exposure and outcome (on confounders and mediators too)
- In most scenarios missing data and FICO will be >10%
- Auxiliary variables/predictors of response not in the substantive model of interest may be available
- In the majority of research scenarios in longitudinal surveys CCA will be biased

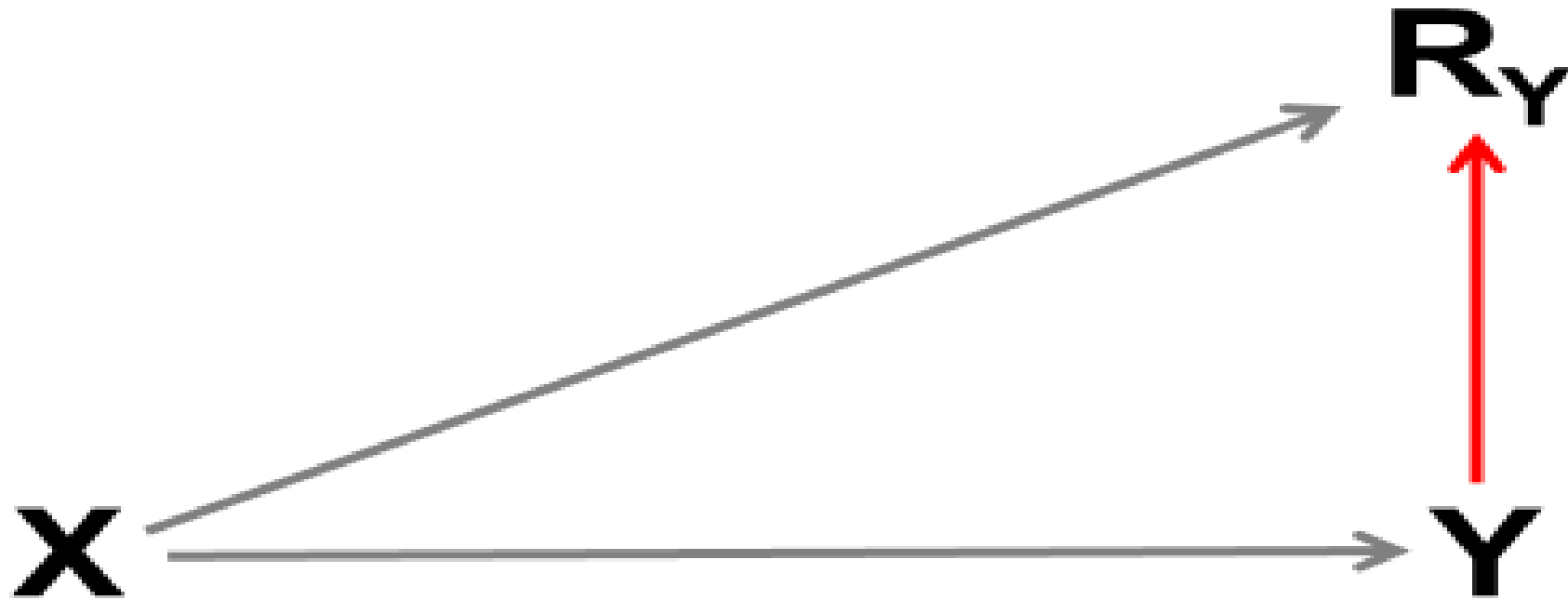
Missing At Random DAG



Missing At Random - MAR

- Systematic differences between the missing values and the observed values can be explained by observed data
- Given the observed data, missingness does not depend on unobserved variables
- $\Pr(R_Y) = \Pr(R_Y | X)$
- MAR methods: Multiple Imputation (various forms of), Full Information Maximum Likelihood, Inverse Probability Weighting, Fully Bayesian methods, Linear Increments, Doubly Robust Methods (IPW +MI)
- All methods assume that all/most important drivers of missingness are available
- MAR largely untestable, but see Karthika et al, 2013 & Seaman et al, 2013 for exceptions
- **Which variables?**

Missing Not At Random - DAG



Missing Not At Random - MNAR






- Even after accounting for all observed information, differences remain between the missing values and the observed values
- Unobserved variables are responsible for missingness
- $\Pr(R_Y) = \Pr(R_Y | Y, X)$
- Untestable!
- Selection models and/or pattern mixture models
- Both approaches make unverifiable distributional assumptions!
- Choice depends on the complexity of the substantive question
- Choice between MAR and MNAR models not straightforward

Rubin's framework and representativeness/balanced samples




- **MCAR:** No selection, sample is “representative”/balanced
- **MAR:** Observed variables account for selection. Given these, sample is representative/balanced
- **MNAR:** Observed variables do not account for selection (selection is due to unobservables too)
- MAR and MNAR are largely untestable*, but if a “gold standard” for the target population exists, we could test whether after accounting for selection with auxiliary variables the distribution of target variables is similar to that observed in the population
- Even when distributions are similar the target variables can still be MNAR, but the bias (for this specific variable) is probably negligible

The National Child Development Study (NCDS- 1958 cohort)

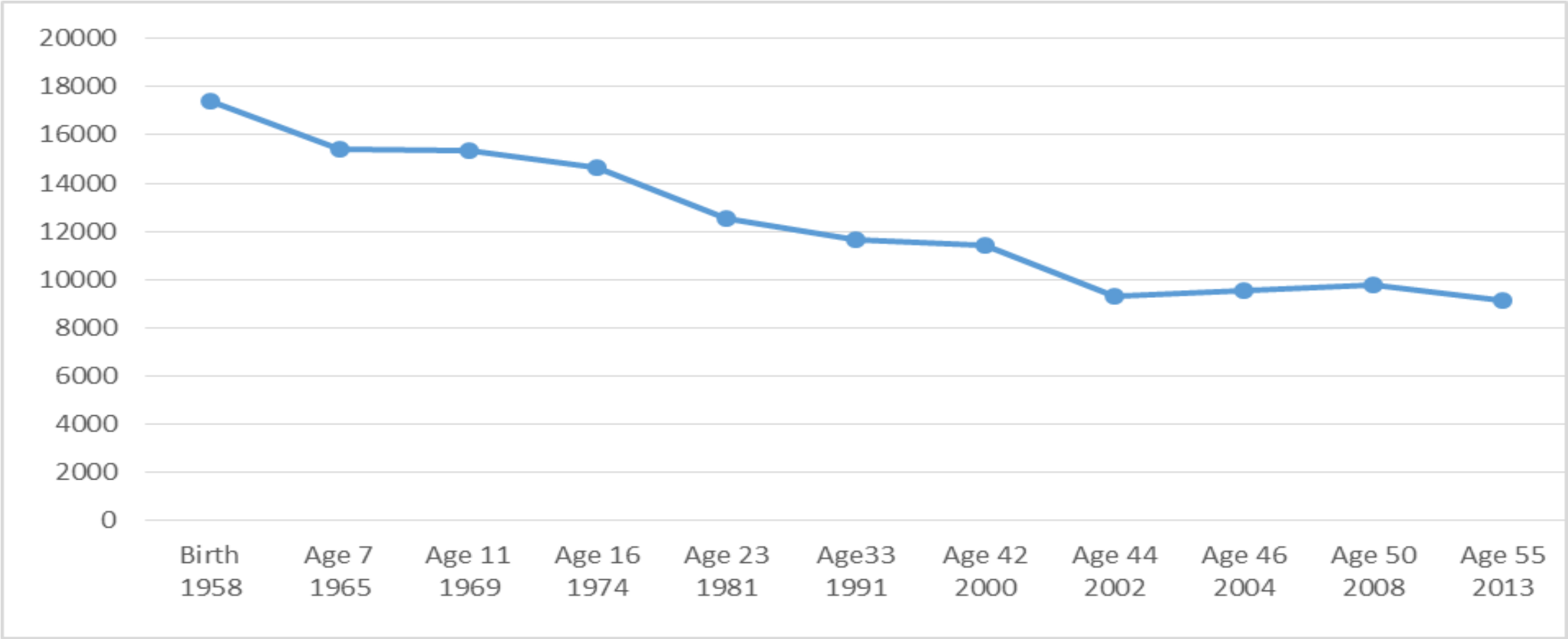
CENTRE FOR LONGITUDINAL STUDIES

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

Types of information covered

 Birth	 School years	 Adult
Household composition Parental social class Obstetric history Smoking in pregnancy Pregnancy (problems, antenatal care) Labour (length, pain relief, problems) Birthweight, length	Household composition Parental social class Parental employment Financial circumstances Housing Health Cognitive tests Emotions and behaviour School Views and expectations Attainment	Household composition Employment Social class Income Housing Health Well-being and mental health Health-related behaviour Training and qualifications Basic skills Cognitive tests Views and expectations

Response in NCDS



Monotone vs. Non-monotone response

Response patterns	Freq.	Percent
Monotone	5688	30.64
Non-monotone	8329	44.89
All waves	4,541	24.47
Total	18558	100

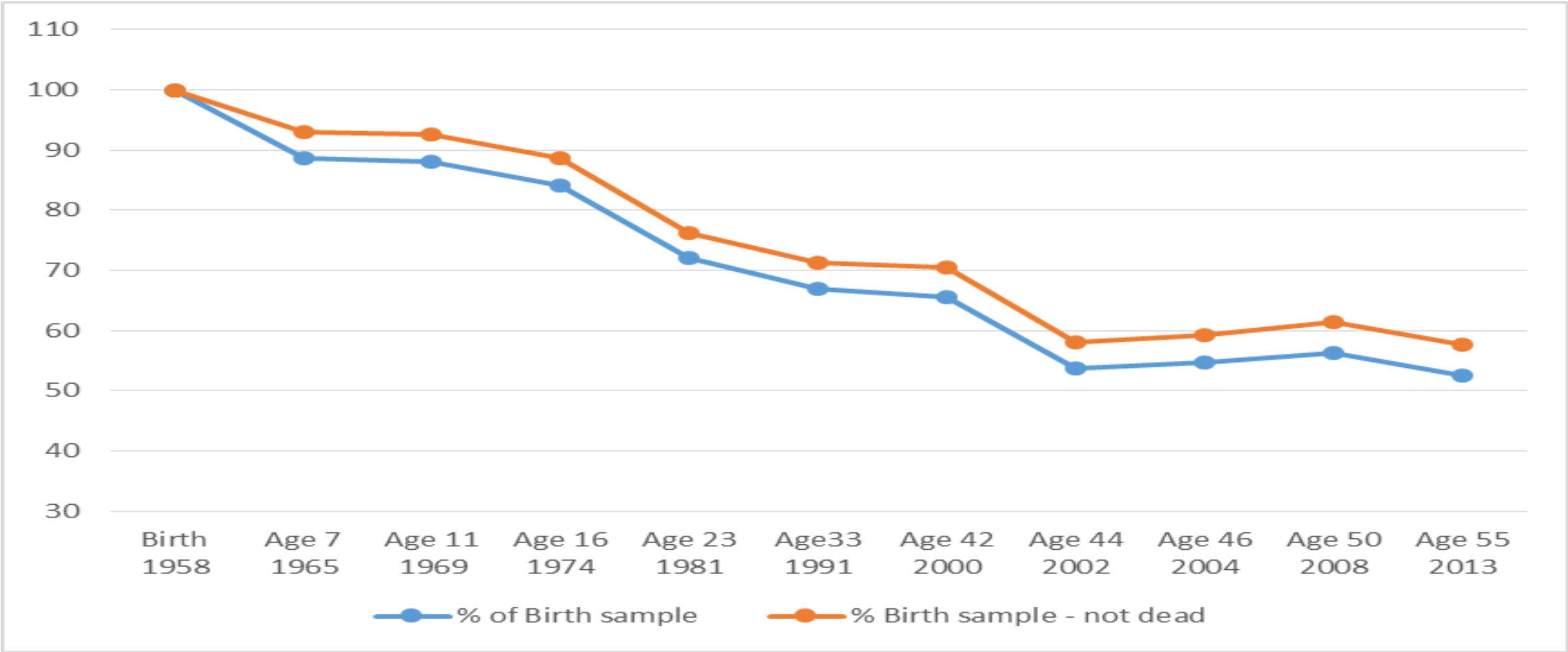
Non response in NCDS

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

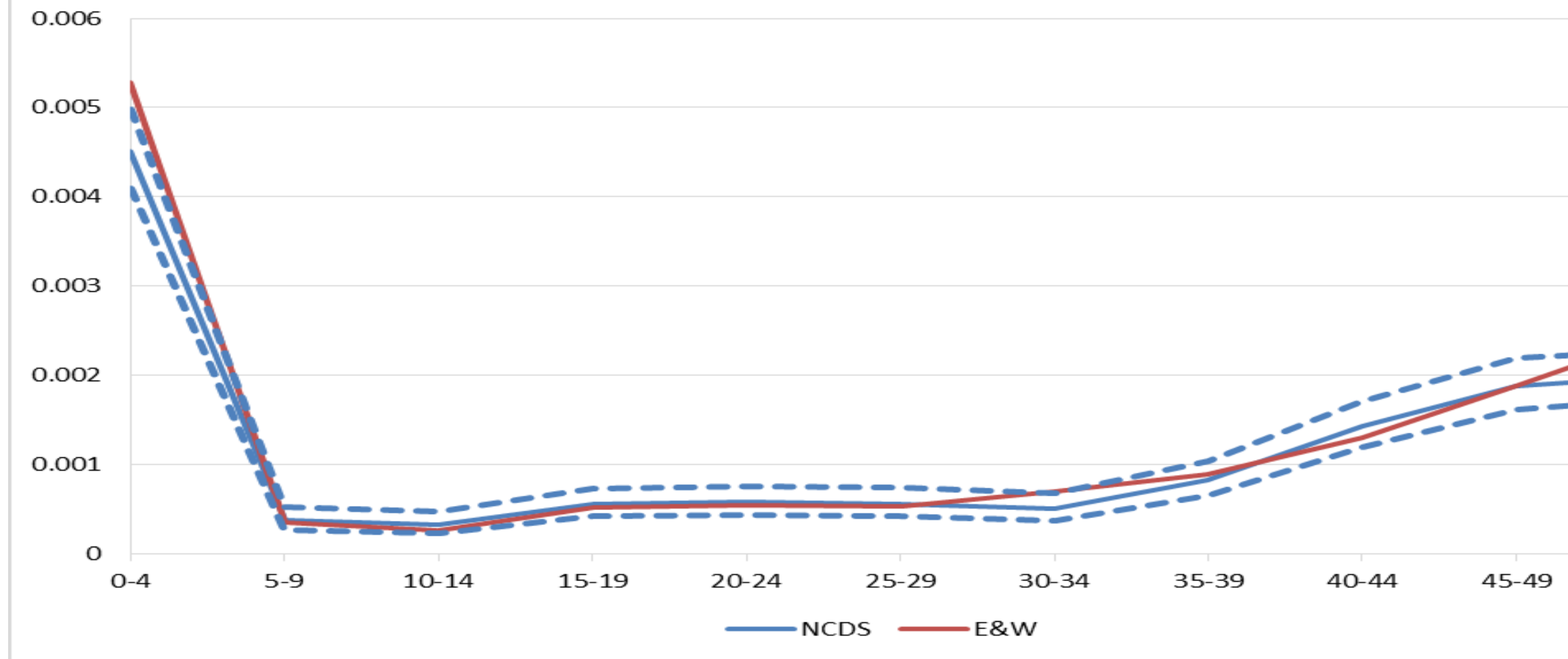
Non response in NCDS

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

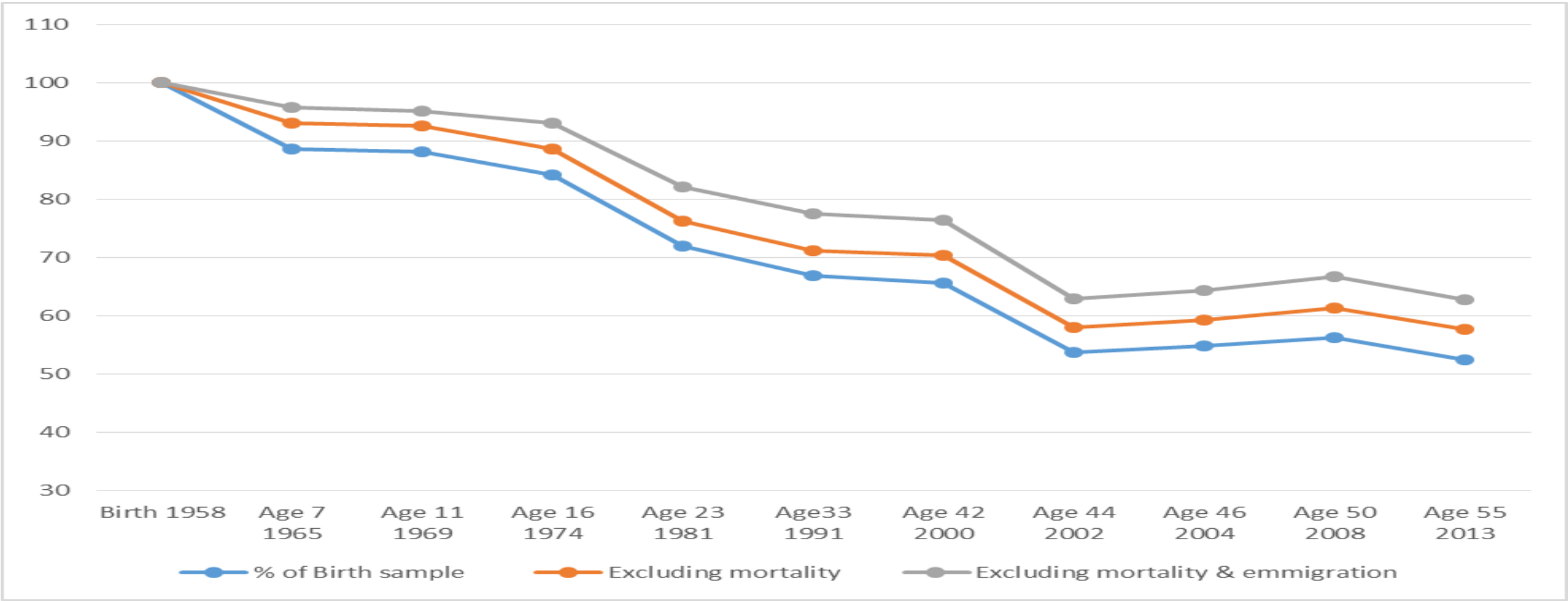
Sample size in the 1958 cohort as % of the original sample



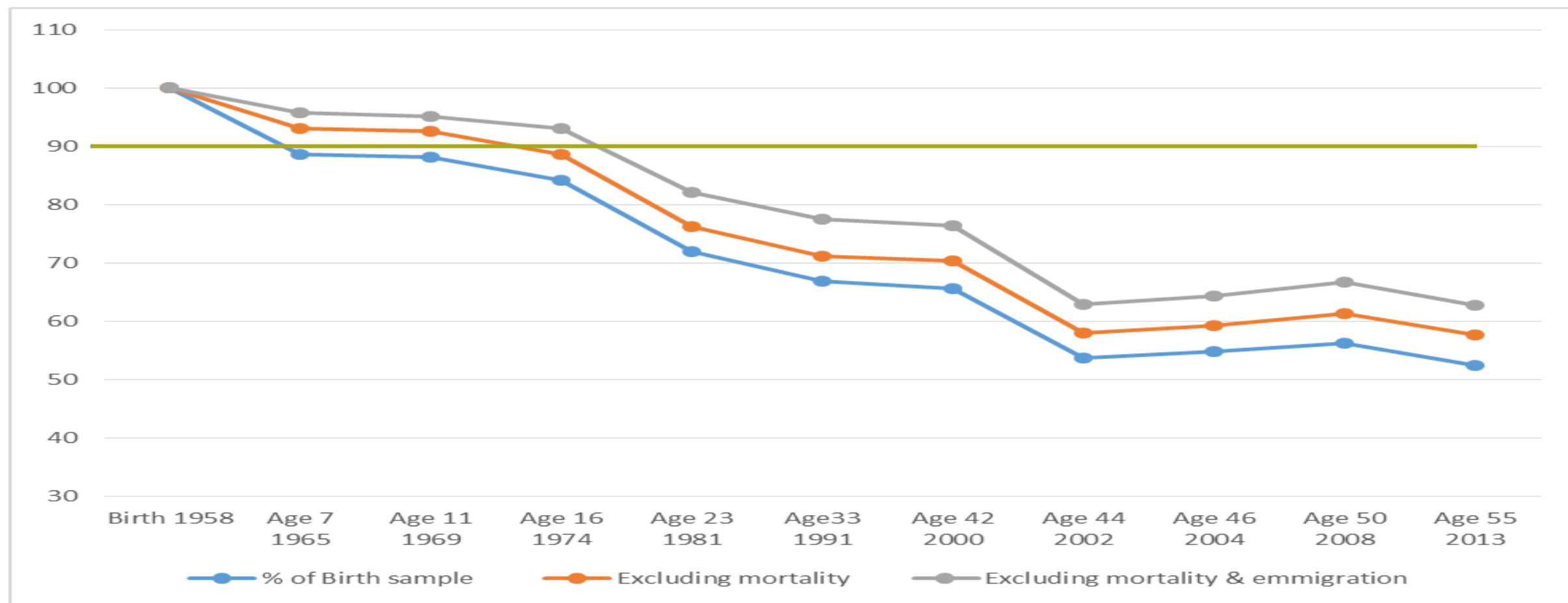
NCDS and England and Wales Mortality Rates



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



The 10% rule (of thumb)



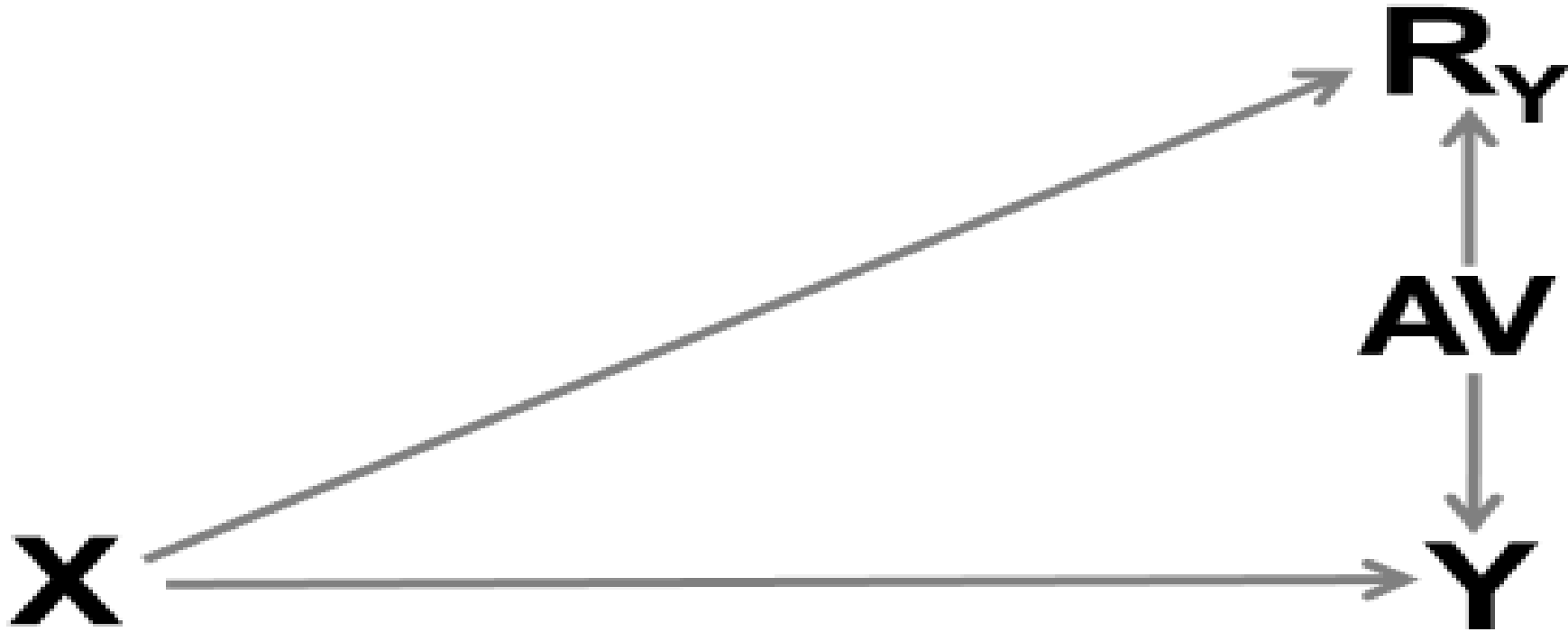
What happens in the 1958 cohort?

- We know that the missing data generating mechanism is not MCAR
- For the majority of research scenarios CCA will either be biased and/or inefficient
- CCA probably OK if outcome and exposure up to age 11
- Missing data generating mechanism is either MAR or MNAR
- Both largely untestable* – rely on unverifiable assumptions
- MAR vs MNAR related to omitted variable bias/unmeasured confounding bias
- In the majority of research scenarios in the 1958 cohort a principled approach to the analysis of incomplete records is needed

CLS Missing Data Strategy

- Applied methodological work
- A simple idea - Maximise the plausibility of the MAR assumption
- Exploit the richness of longitudinal data to address sources of bias
- In the 1958 cohort (and any study) the information that maximises the plausibility of MAR is finite
- The information maximising MAR **that matters in practice** can be at least approximated
- We can identify the variables that are associated with non response
- Auxiliary variables – not in the substantive model

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



A data driven approach to maximise the plausibility of MAR

- Data driven approach to identify predictors of non response in all waves of the CLS studies
- Substantive interest: Understanding non response
- Is early life more important, or it's all about what happened in the previous wave?
- Can we maximise the plausibility of MAR with sets of early life variables, or later waves are needed too?
- Are the drivers of non response similar between cohorts?
- The goal is to understand non response and in the process identify auxiliary variables that can be used in realistically complex models that assume MAR
- AV's to be used in addition to the variables in the substantive model and predictors of item non response and/or strong predictors of the outcome

MAR vs MNAR

- Some missing data patterns/variables may be MNAR even after the introduction of auxiliary variables
- Non monotone patterns are more likely to be MNAR (Robins & Gill, 1997)
- We assume that after the introduction of AV's our data is either MAR, or not far from being MAR, so bias is negligible
- Reasonable assumption - Richness of longitudinal data
- Can't be sure, but MAR methods have been shown to perform well even when data are MNAR (Collins et al, 2001)
- Our results will inform sensitivity analyses for departures from MAR
- Arguably MAR methods **more suitable** than MNAR methods in rich longitudinal studies

A data driven approach

- Identify predictors of response for each wave of NCDS: Variables from each wave can only predict response on subsequent waves
- About 17000 variables! => Selection is done in three stages

Pre – selection

- We exclude routed variables, binary variables <1%, item non response > 50%

Analysis:

- Stage 1: univariate regressions within wave
- Stage 2: multivariable regressions within wave
- Stage 3: multivariable regression across waves
- Variable selection repeated with machine learning algorithms

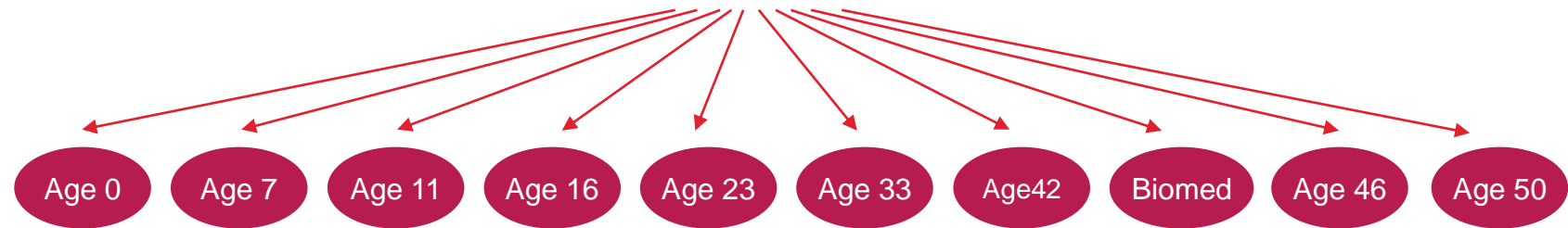
Variables in
NCDS up to age
50

“Eligible” vars

17,412

Routed
Binary <1%,
Missing > 50%

1,048



Stage 1	19	50	48	58	73	100	276	81	155	194
---------	----	----	----	----	----	-----	-----	----	-----	-----

Stage 2	11	36	34	47	50	68	114	39	53	120
---------	----	----	----	----	----	----	-----	----	----	-----

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

Stage 3 Input	4	13	23	36	52	54	81	86	88	116
---------------	---	----	----	----	----	----	----	----	----	-----

Stage 3 Output	4	9	11	17	24	18	40	34	35	39
----------------	---	---	----	----	----	----	----	----	----	----

Stage 1: Univariate regressions

- Response: All waves
- Predictors: waves up to age 50
- Univariate logistic regressions to predict response in subsequent waves

N of predictors	W0	W1	W2	W3	W4	W5	W6	Biomed	W7	W8
Stage 1	19	50	48	58	73	100	276	81	155	194

- Regressions for: (wave0= 19*10); (wave5=100*5); ...
- Stata loop
- Retained predictors with a “significant” impact on response at $p < 0.01$

Stage 2

N of predictors	W0	W1	W2	W3	W4	W5	W6	Biomed	W7	W8
Stage 1	19	50	48	58	73	100	276	81	155	194
Stage 2 Input	11	36	34	47	50	68	114	39	53	120

- Stage 2 consists of multivariable regressions using all predictors within wave (i.e. 11 for W0, 36 for W1, 120 for W8) that were retained after Stage 1
- **Stages 1 & 2 combined** in Machine Learning replication of “traditional” approach
- Variables compete against each other within wave
- Log binomial models – Risk Ratio (outcome not always rare – non collapsibility of the OR)
- Least Absolute Shrinkage and Selection Operator with glmnet in R using cyclical coordinated descent & Forward Stepwise with LARS in Stata
- This results in **different subsets of predictors from each wave**

Stage 3

- At this stage we let predictors from different waves compete against each other (i.e. subset of predictors from W0 and W1 predict response in W2; subset of predictors from W0, 1, 2, 3, 4 predict response in W5 etc)

Variables entering Stage 3:

N of predictors	Resp1	Resp2	Resp3	Resp4	Resp5	Resp6	Biomed	Resp7	Resp8	Resp9
Stage 3 Input	4	13	23	36	52	54	81	86	88	116

- Challenge:** predictors from different waves => different levels of missingness

Stage 3

- Progressive imputation and alternation of MI and response modelling
- We impute missing predictors then estimate response models and so on for each wave
- MI with chained equations
- Response modelling with multivariable regressions (Log binomial) and Machine learning (logit)
- LASSO with glmnet in R, Forward Stepwise with LARS in Stata (work in progress)

Results from Stage 3

- The number of predictors of response at each wave is reduced

N of predictors	Resp1	Resp2	Resp3	Resp4	Resp5	Resp6	Biomed	Resp7	Resp8	Resp9
Stage 3: input	4	13	23	36	52	54	81	86	88	116
Stage 3: output	4	9	11	17	24	18	40	34	35	39

- Predictors include all types of variables: Social, economic, health, and survey related (cooperation with sub-studies)

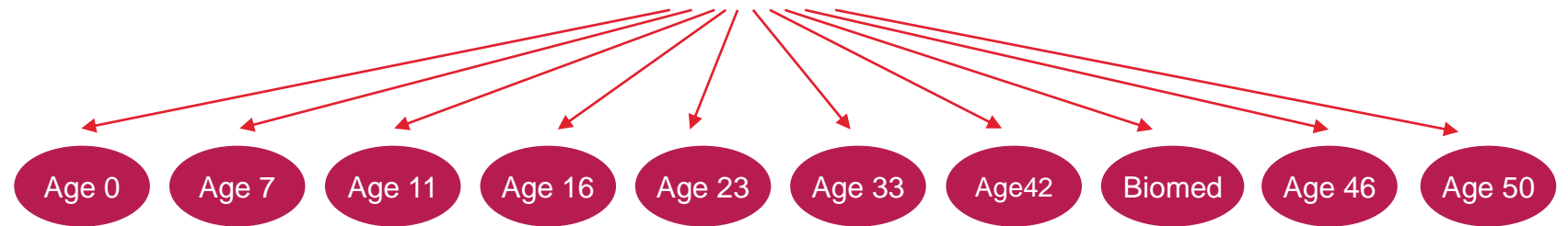
Variables in
NCDS up to age
50

“Eligible” vars

17,412

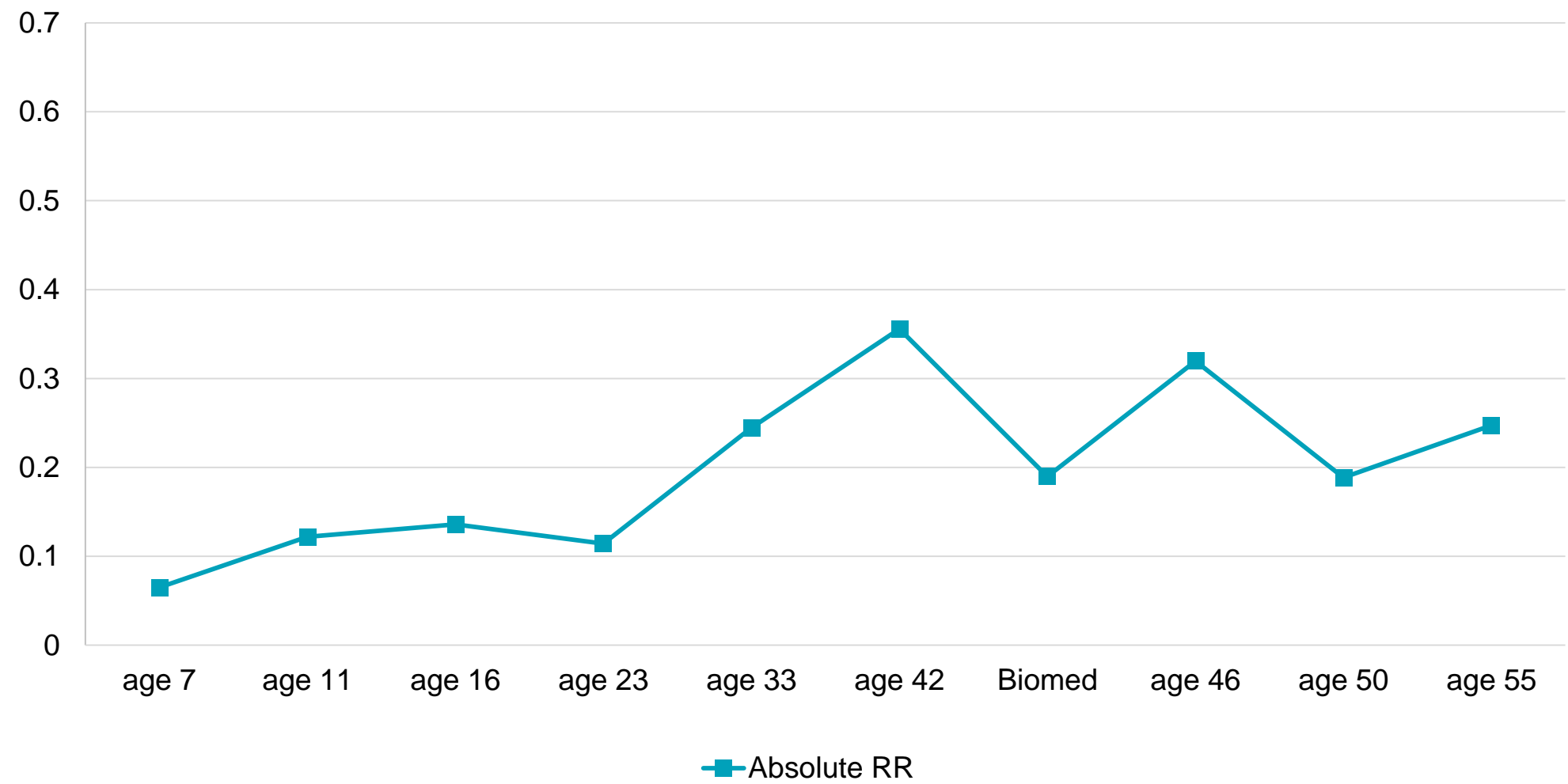
Routed
Binary <1%,
Missing > 50%

1,048



Stage 1	19	50	48	58	73	100	276	81	155	194
Stage 2	11	36	34	47	50	68	114	39	53	120
Response	Age 7	Age 11	Age 16	Age 23	Age 33	Age 42	Biomed	Age 46	Age 50	Age 55
Stage 3 Input	4	13	23	36	52	54	81	86	88	116
Stage 3 Output	4	9	11	17	24	18	40	34	35	39

Stroger predictor highest absolute Risk Ratio difference from 1



Highest predictors of missingness at age 42

Variable	Effect	RR	95% CI
Member of a Trade Union/Staff Association at age 33	Those who were members of a staff association at age 33 are less likely to be present at age 42, compared to those members of trade unions.	0.64	(0.6; 0.69)
Ever been convicted of traffic offence at age 33	Those convicted of traffic offence at age 33 are less likely to be present at age 42, compared to those who were not.	0.79	(0.75; 0.83)
Type of accommodation at age 33	Those living in a terraced house at age 33 are less likely to be present at age 42, compared to those living in detached house or similar.	0.88	(0.86; 0.91)
In a steady relationship at age 33	Those not having a steady relationship at age 33 are less likely to be present at age 42, compared to those who had a steady relationship at that age.	0.89	(0.84; 0.94)

Highest predictors of missingness at age 55

Variable	Effect	RR	95% CI
Region at age 46 (based on post-1974 regions)	Those living in Wales at age 46 are less likely to be present at age 55, compared to those living in the North.	0.75	(0.67; 0.84)
Standard (Statistical) Region at Interview at age 46	Those living in Yorkshire and Humberberside at age 46 are less likely to be present at age 55, compared to those living in the North.	0.84	(0.78; 0.89)
Participation at age 23	Those who did not participate at NCDS 5 (age 33) are less likely to be present at age 55, compared to those who did participate.	0.84	(0.8; 0.88)
Region at birth	Those who lived in the South at birth (1958) are less likely to be present at age 55, compared to those living in the North.	0.87	(0.75; 0.99)

Are they any good?

- So, we identified predictors of response, but are they any good?
- How effective are the identified “auxiliary” variables in reducing bias?
- Two “experiments” can shed some light into this
 - i) Can we replicate the composition of the sample at birth despite attrition?
 - ii) Can we replicate the “known” population distribution despite attrition?
- Results from (i) available, working with ONS (Census, Annual Population Survey, Integrated Household Survey) on (ii), results available soon

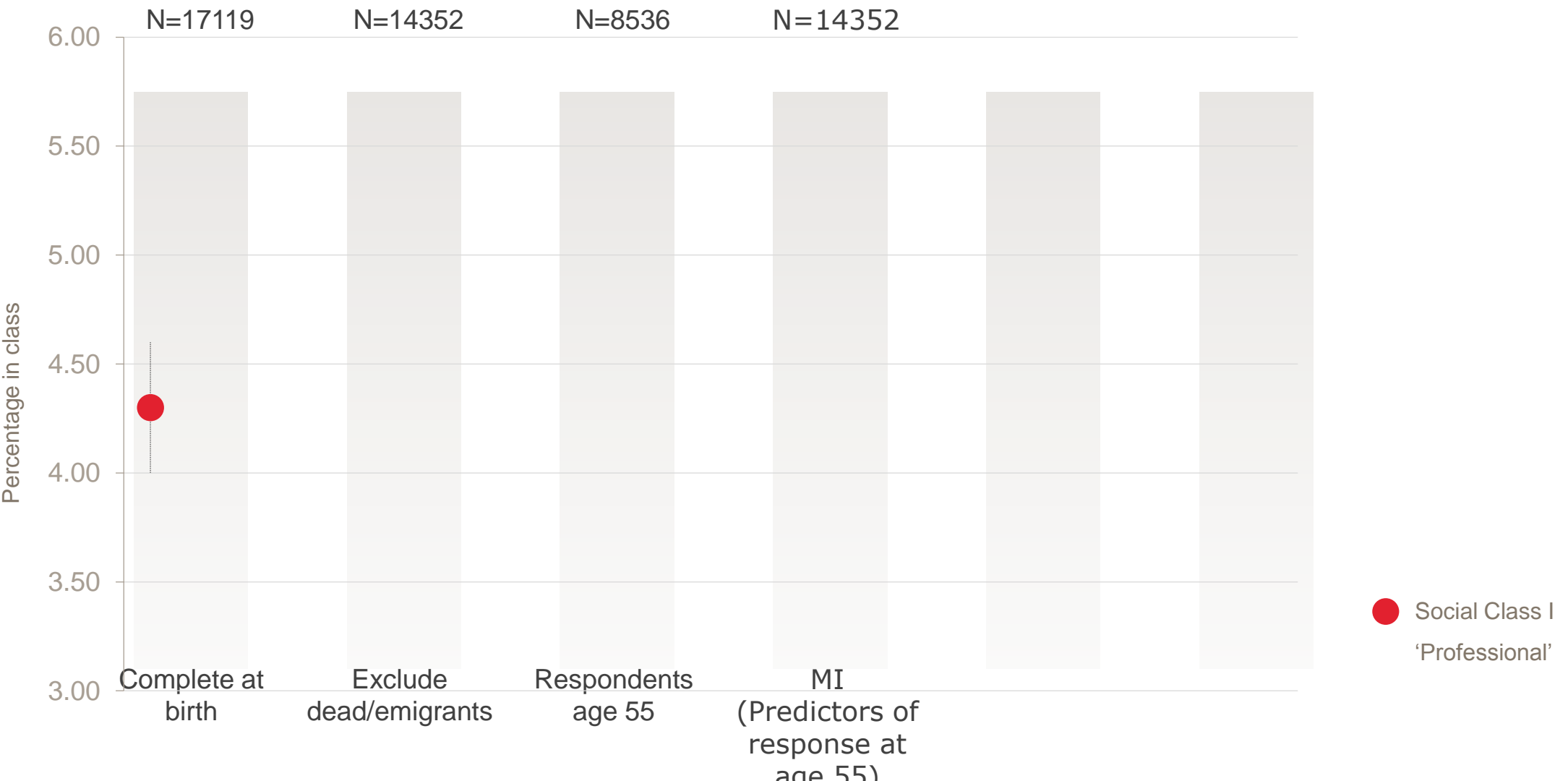
Social Class of mother's husband 1958

Variable	Complete at birth			
	Freq.	Percent	Confidence Intervals	
Social Class of mother's husband 1958				
I Professional	731	4.3	4.0	4.6
II Intermediate	2,113	12.3	11.9	12.8
III NM Skilled non-manual	1,565	9.1	8.7	9.6
III M Skilled manual	8,253	48.2	47.5	49.0
IV Semi-skilled manual	1,958	11.4	11.0	11.9
V Unskilled manual & other	2,499	14.6	14.1	15.1

- Can we replicate the composition of Social Class at birth with participants at 55 (N = 8536)?
- MI with chained equations, 20 imputations using auxiliary variables

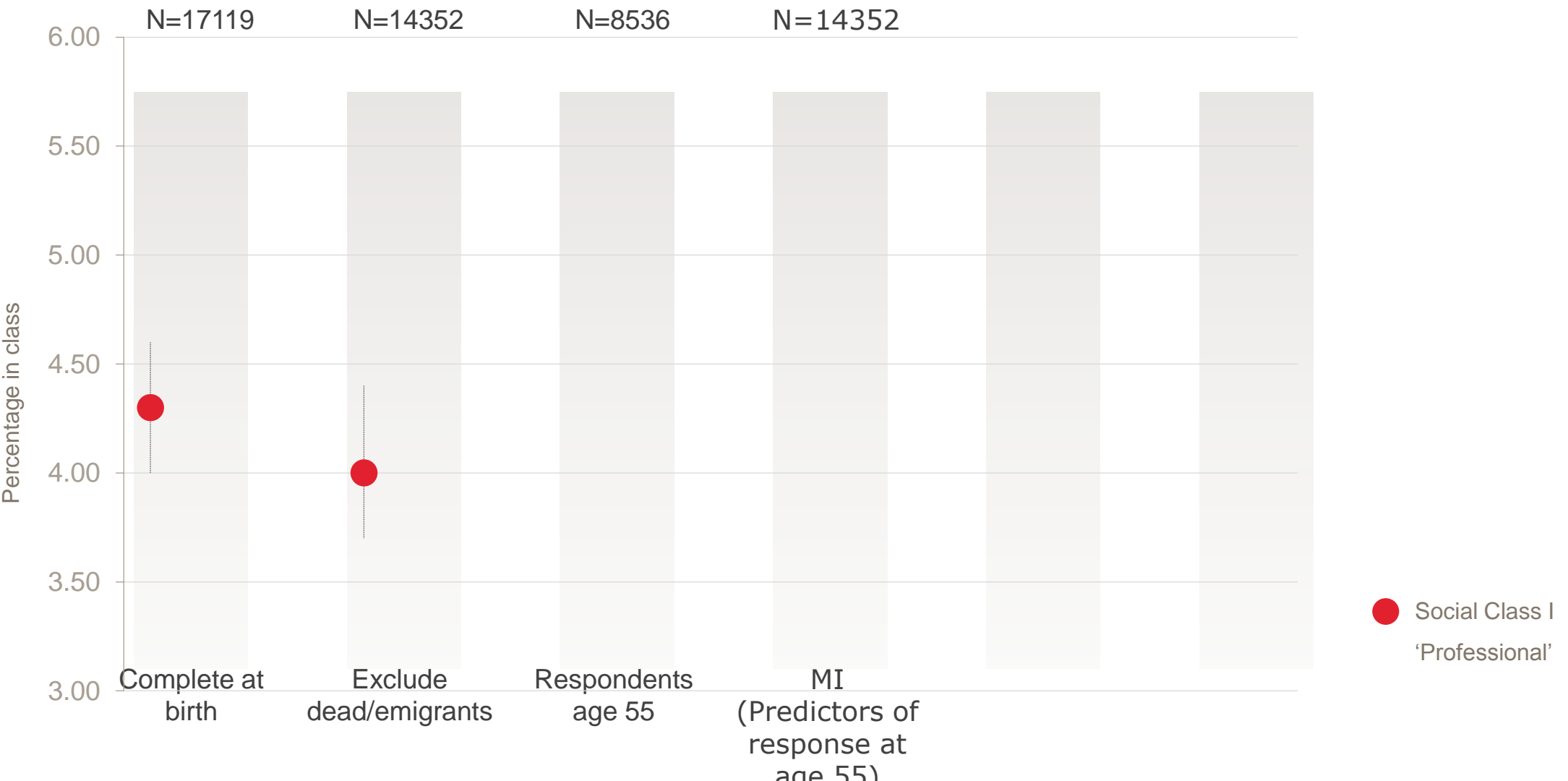
Social Class of mother's husband 1958

Percentage in Social Class I



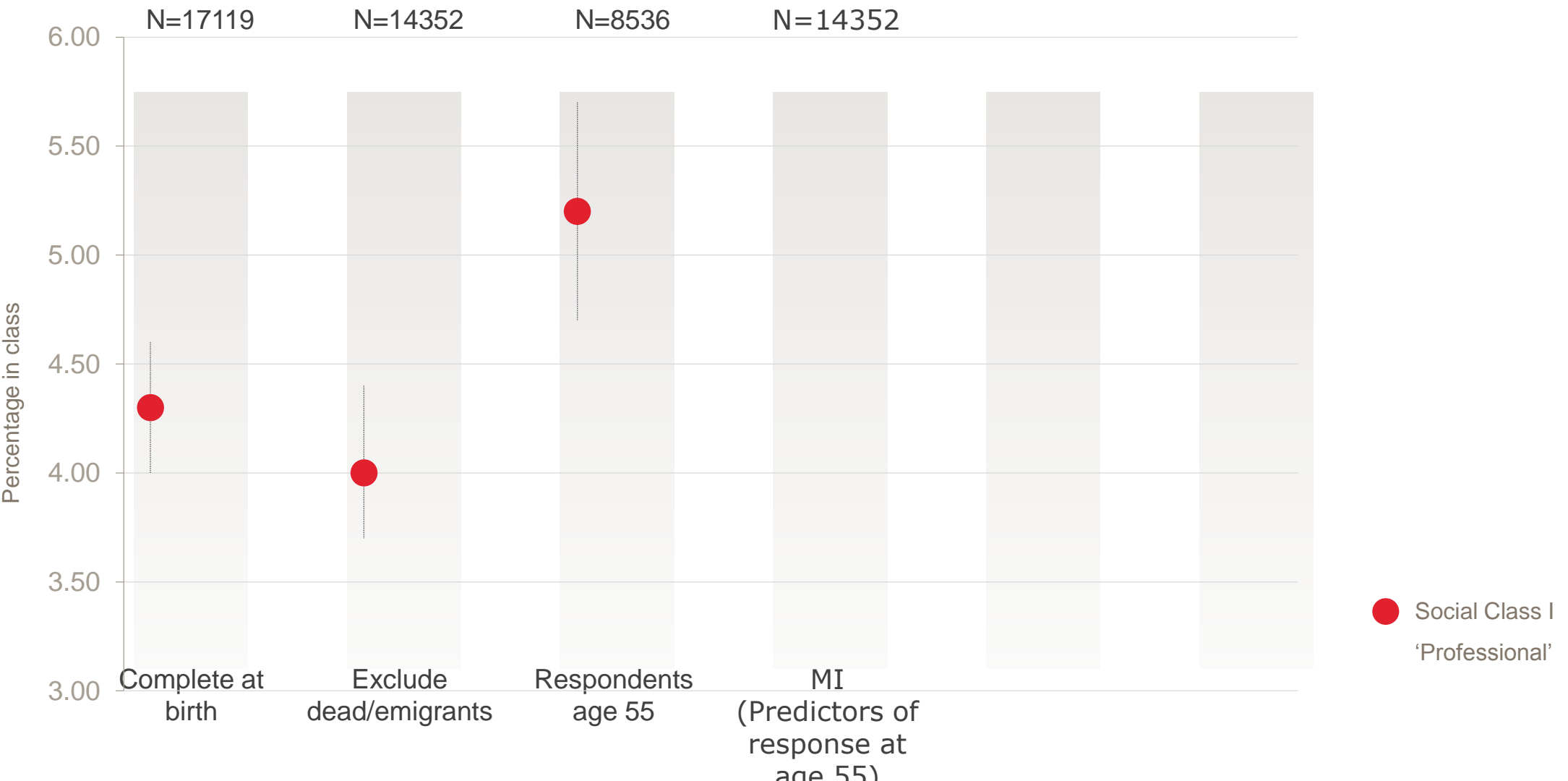
Social Class of mother's husband 1958

Percentage in Social Class I



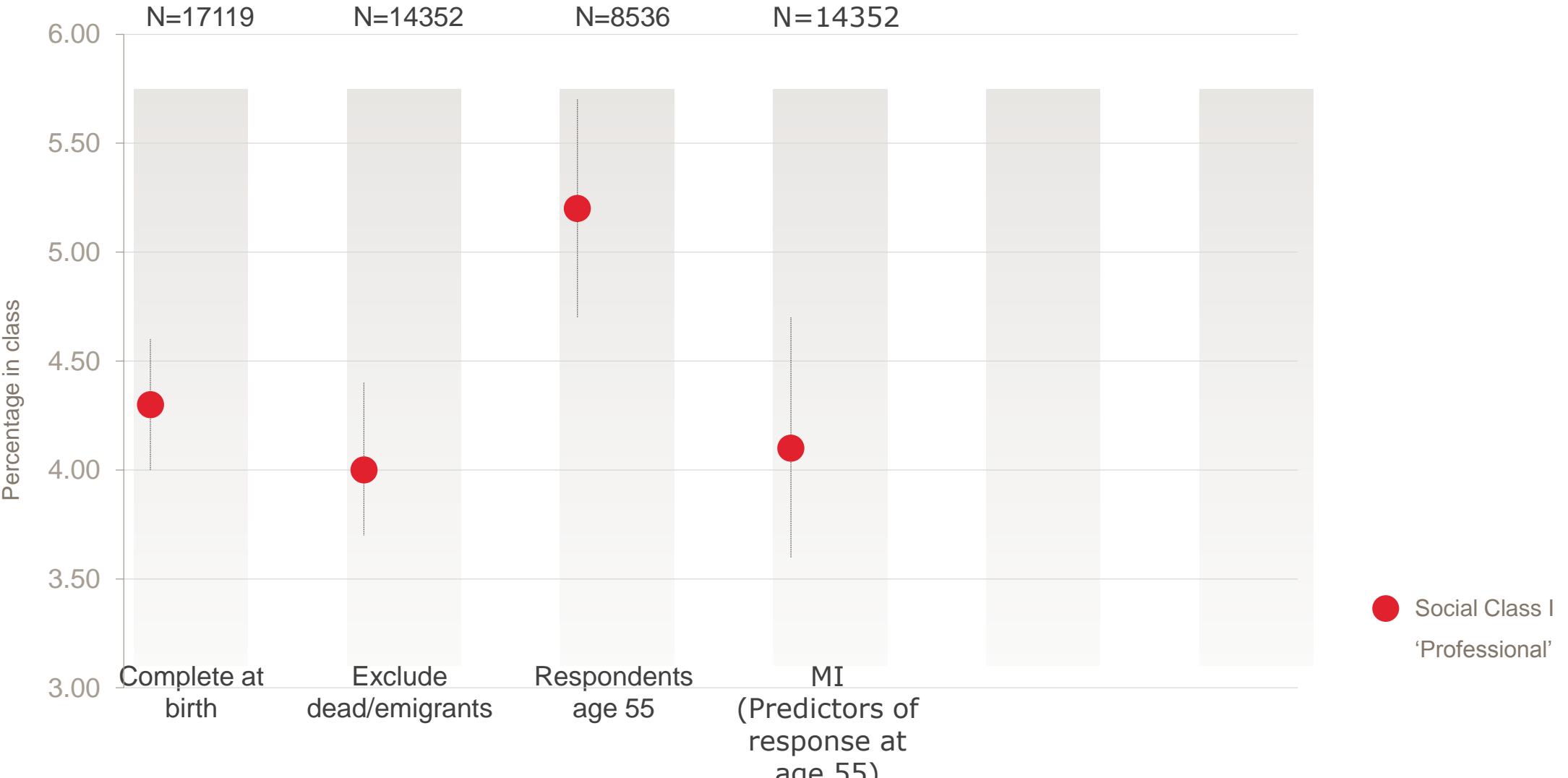
Social Class of mother's husband 1958

Percentage in Social Class I



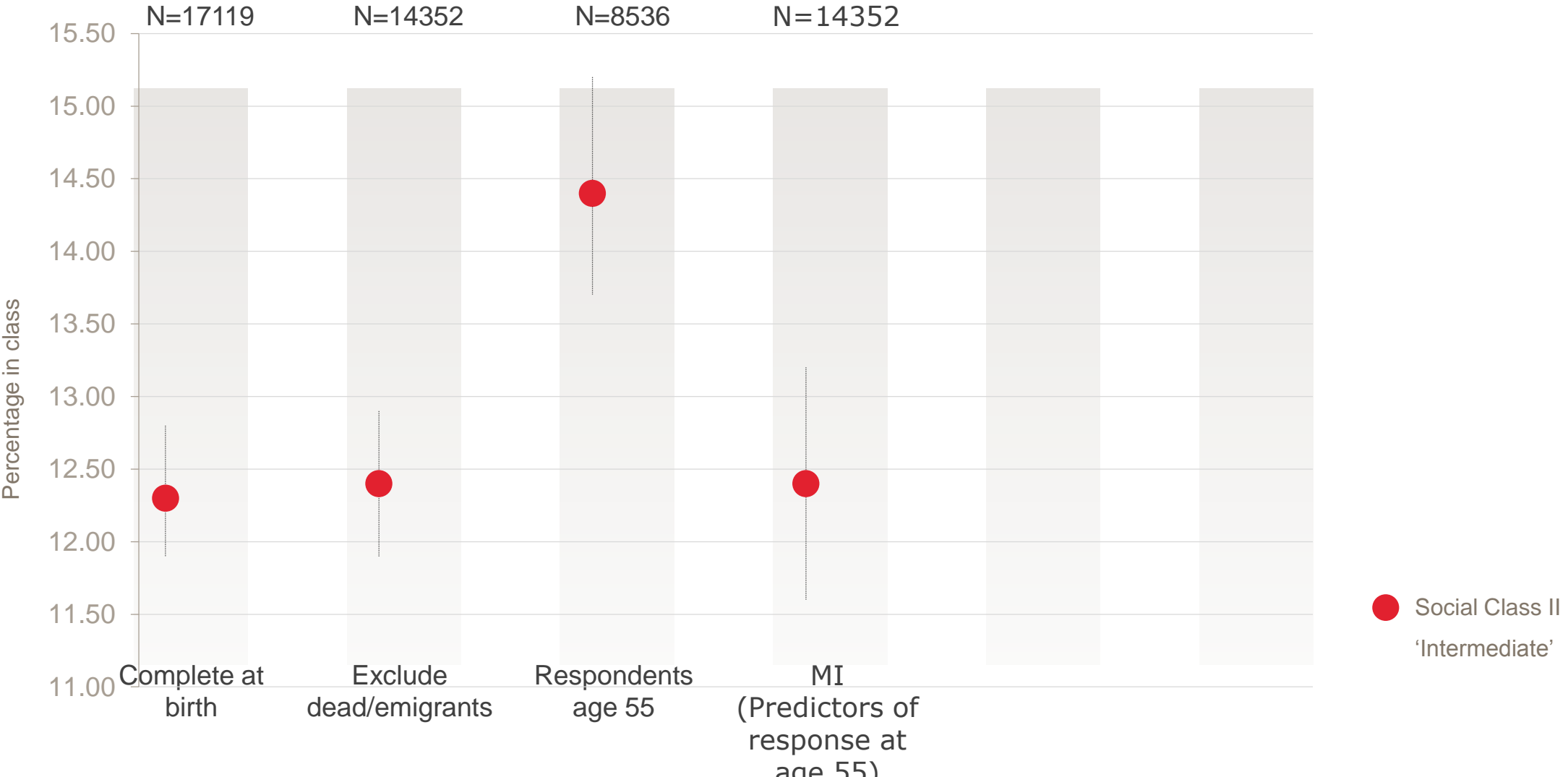
Social Class of mother's husband 1958

Percentage in Social Class I



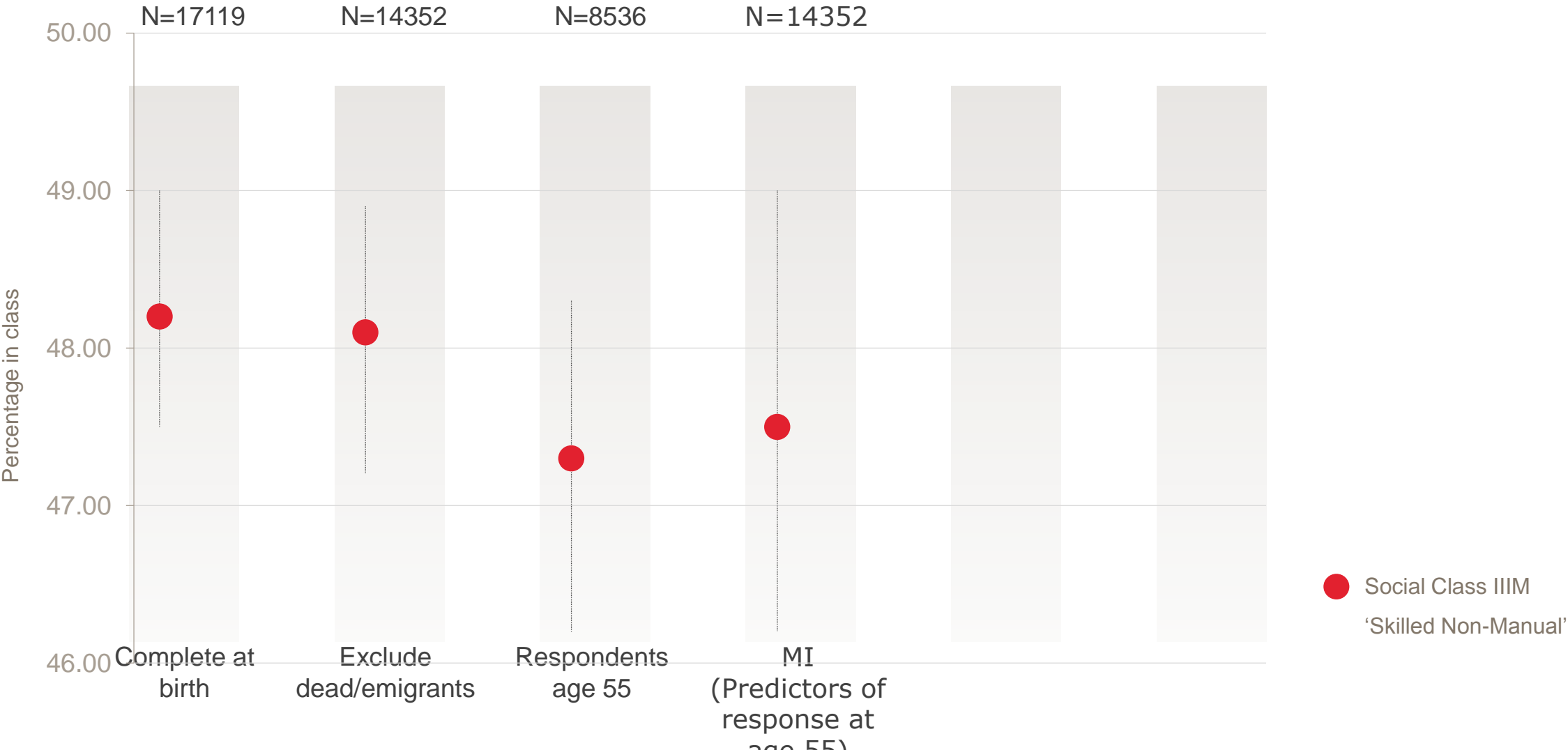
Social Class of mother's husband 1958

Percentage in Social Class II



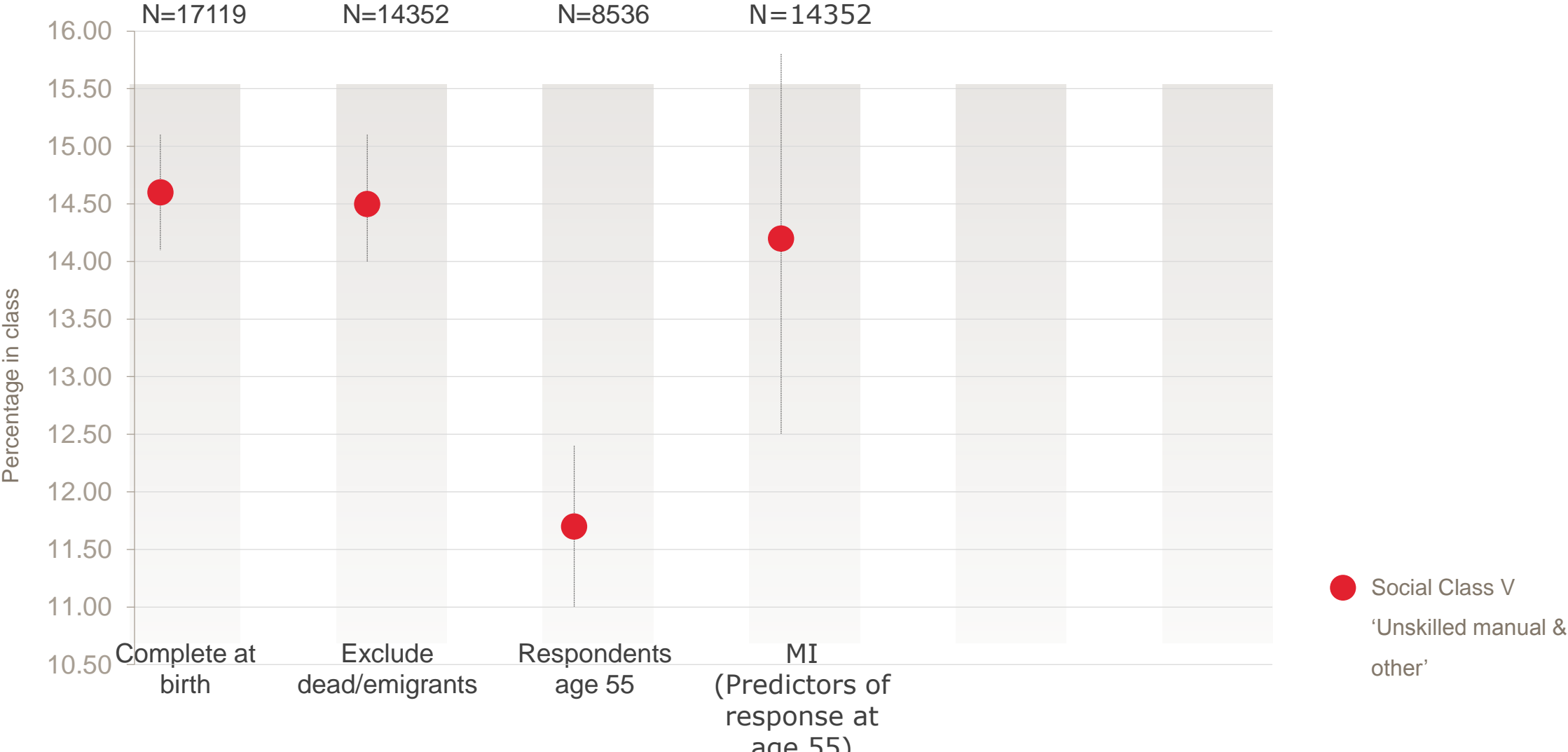
Social Class of mother's husband 1958

Percentage in Social Class IIIM



Social Class of mother's husband 1958

Percentage in Social Class V



What we will advise CLS data users to do

- Non dynamic longitudinal models

Regression based analyses, including interactions and/or formal mediation

“Longitudinal” since data from at least two stages of the lifecourse are used

>80% of papers in NCDS

MI plausible, arguably more flexible than FIML since auxiliary variables are more easily included in the imputation phase

- Dynamic longitudinal models

Explicitly quantify change over time: Growth models, Mixtures, Latent Transitions Models, Fixed/Random effects, Multilevel models, Generalised Estimating Equations, Generalised Methods of Moments etc

Difficult to incorporate longitudinal structure to imputation model – FIML more flexible

Which variables should be used for missing data handling (with MI, FIML etc)?

- Variables in the substantive model
- “Complete” variables measured at birth (social class, birthweight, maternal smoking)
- Strong predictors of the outcome(s) (deals with item non response too)
- Auxiliary variables from our list that are also associated with the outcome
- Auxiliary variables strongly predicting non-response

Which variables should be used for missing data handling (with MI, FIML etc)?

- Variables in the substantive model
- “Complete” variables measured at birth (social class, birthweight, maternal smoking)
- Strong predictors of the outcome(s) (deals with item non response too)
- Auxiliary variables from our list that are also associated with the outcome
- Auxiliary variables strongly predicting non-response

Example for the user guide

- **Is cognitive function at age 11 associated with childlessness at age 42?**
- At age 11 we have N=14095 cohort members who did cog tests
- By age 42 we have only N=11419 cohort members for whom we know whether or not they're childless (not all present at age 11)
- We control for childhood confounders:
 - Birthweight
 - Breastfeeding
 - Parental social class & education
 - Mother smoking prior to pregnancy
 - Mother working before child aged 5

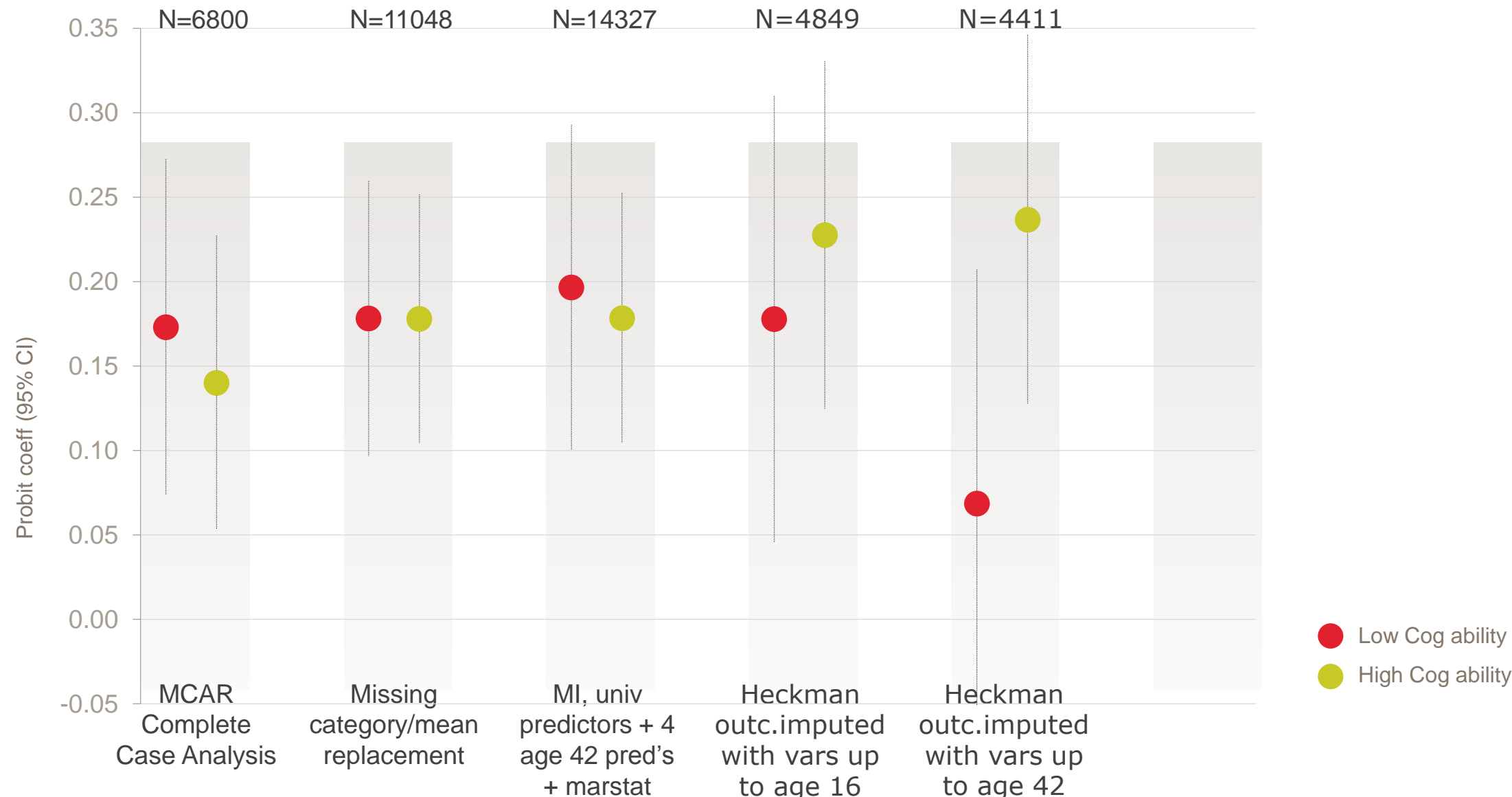
Example for CLS user guide

- Age 11 cog test recoded into 3 categories (preliminary evidence for non linear association)

Low	1SD or more below mean
Middle	mean \pm 1 SD (reference group)
High	1SD or more above mean
- MI with chained equations in Stata 14, 20 imputations
- How many variables are used in the imputation model?
- 16 variables = Substantive model (8), Baseline “complete” (3), partnership status (1), auxiliaries associated with childlessness (3), strongest predictor of response at age 42 (1)
- 8 variables added to the substantive model to maximise MAR
- 65 minutes with Intel Core i7, 4700MQ CPU @2.4GHz, 12GB processor

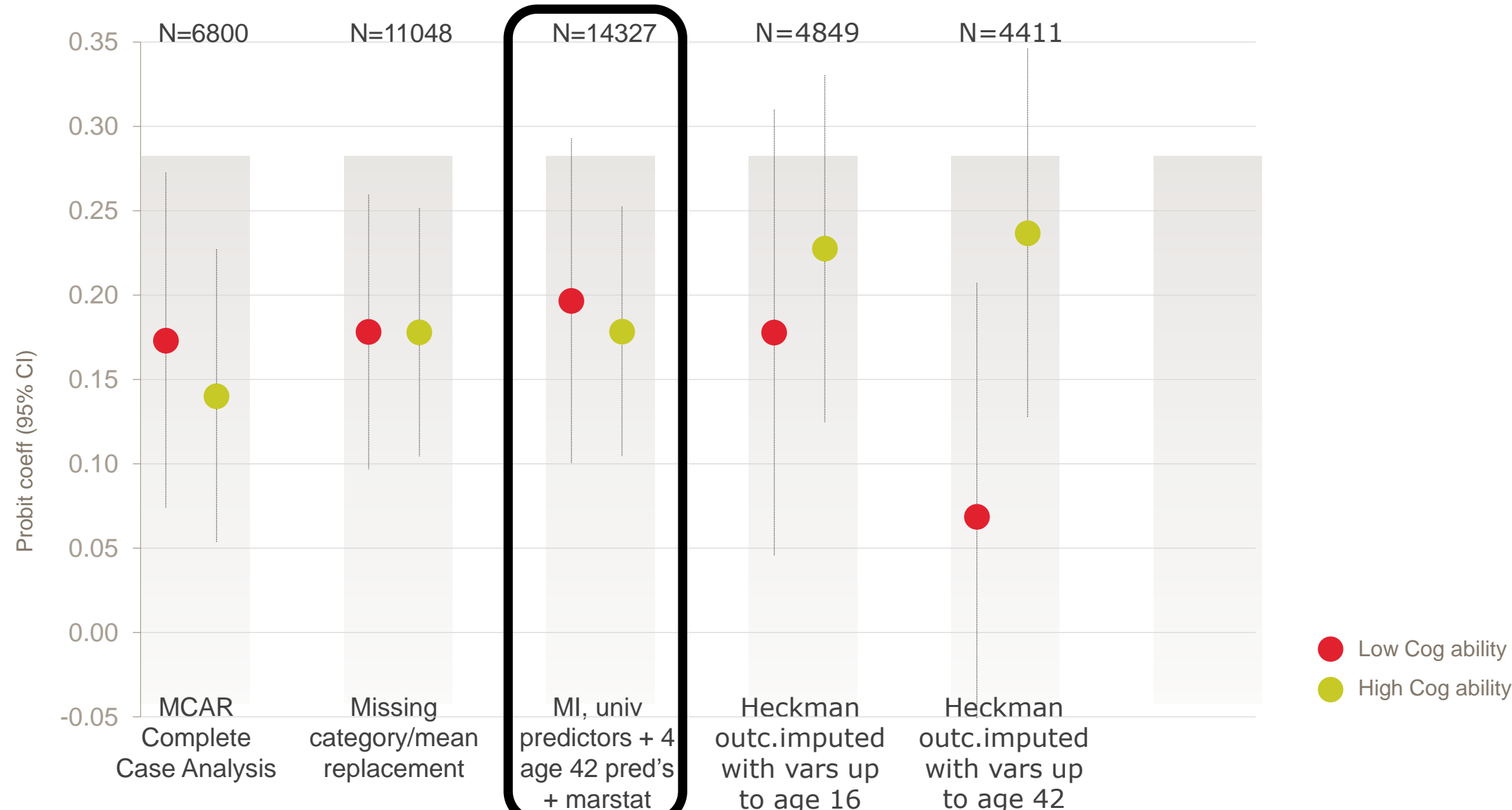
Cognition at age 11 by childlessness at age 42

Probit coefficients/95% CIs



Cognition at age 11 by childlessness at age 42

Probit coefficients/95% CIs



Outputs

- We will **not** make available imputed datasets
- Technical report, peer reviewed papers and user guide
- List of auxiliary variables for users to **adapt** to their analysis
- Stata, R and Mplus code on how to use auxiliary variables with MI
- Transparent assumptions so users can make an informed choice
- Dynamic process, the results will be updated when new waves or other data become available (paradata for example)

Acknowledgments

- Brian Dodgeon
- Tarek Mostafa
- Martina Narayanan
- Benedetta Pongiglione

Thank you for your attention!