Institute of Education

UCL

Maximising the plausibility of the Missing At Random assumption: Results from the 1958 British birth cohort

CENTRE FOR LONGITUDINAL STUDIES

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Outline

- Missing data theory – Rubin’s classification
- Data driven approach for identifying predictors of non response
- Results from the 1958 British birth cohort

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

CENTRE FOR
LONGITUDINAL
STUDIES

Missing Completely At Random - MCAR

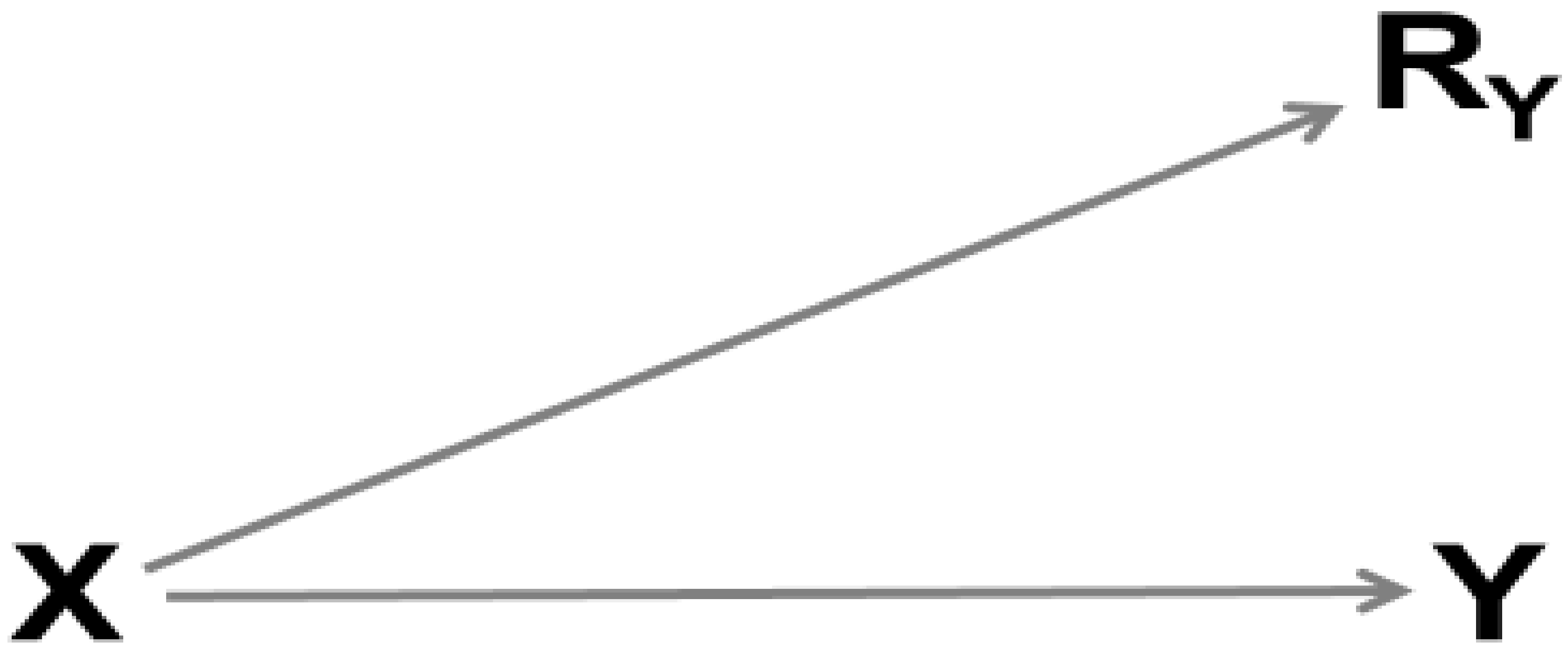
Rubin's framework in the context of longitudinal surveys

- <u>Missing Completely At Random (MCAR):</u> There are no systematic differences between the missing values and the observed values
- <u>Missing At Random (MAR):</u> Systematic differences between the missing values and the observed values can be explained by observed data
- <u>Missing Not At Random (MNAR):</u>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

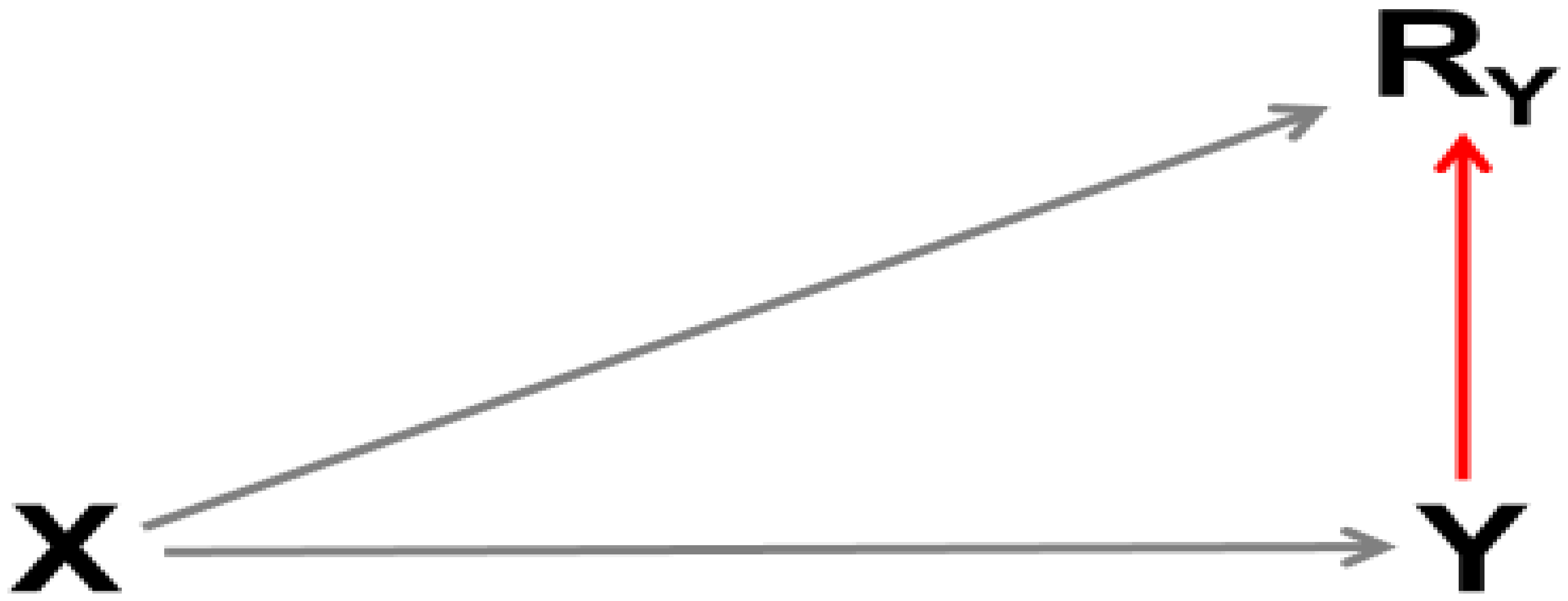


CENTRE FOR
LONGITUDINAL
STUDIES

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



CENTRE FOR
LONGITUDINAL
STUDIES

Rubin's framework in the context of longitudinal surveys

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

CENTRE FOR
LONGITUDINAL
STUDIES

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)

MAR vs MNAR in UK 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

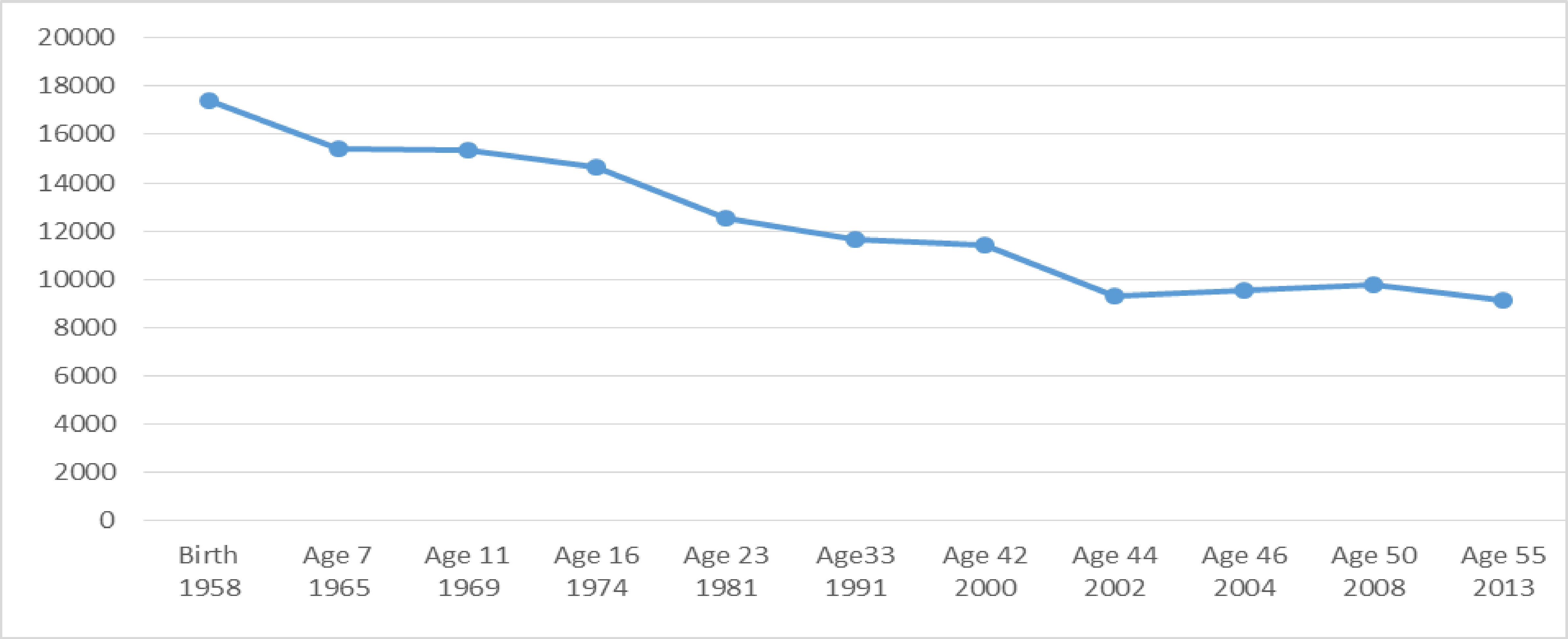
	1958 Birth	1965 7	1969 11	1974 16	1981 23	1991 33	2000 42	2003 45	2004 46	2008 50	2013 55
main respondent	mother	parent	parent	cohort member / parent	cohort member	cohort member	cohort member	cohort member	cohort member	cohort member	cohort member
secondary respondent	medical	school medical	school medical	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

Types of information covered

CENTRE FOR LONGITUDINAL STUDIES

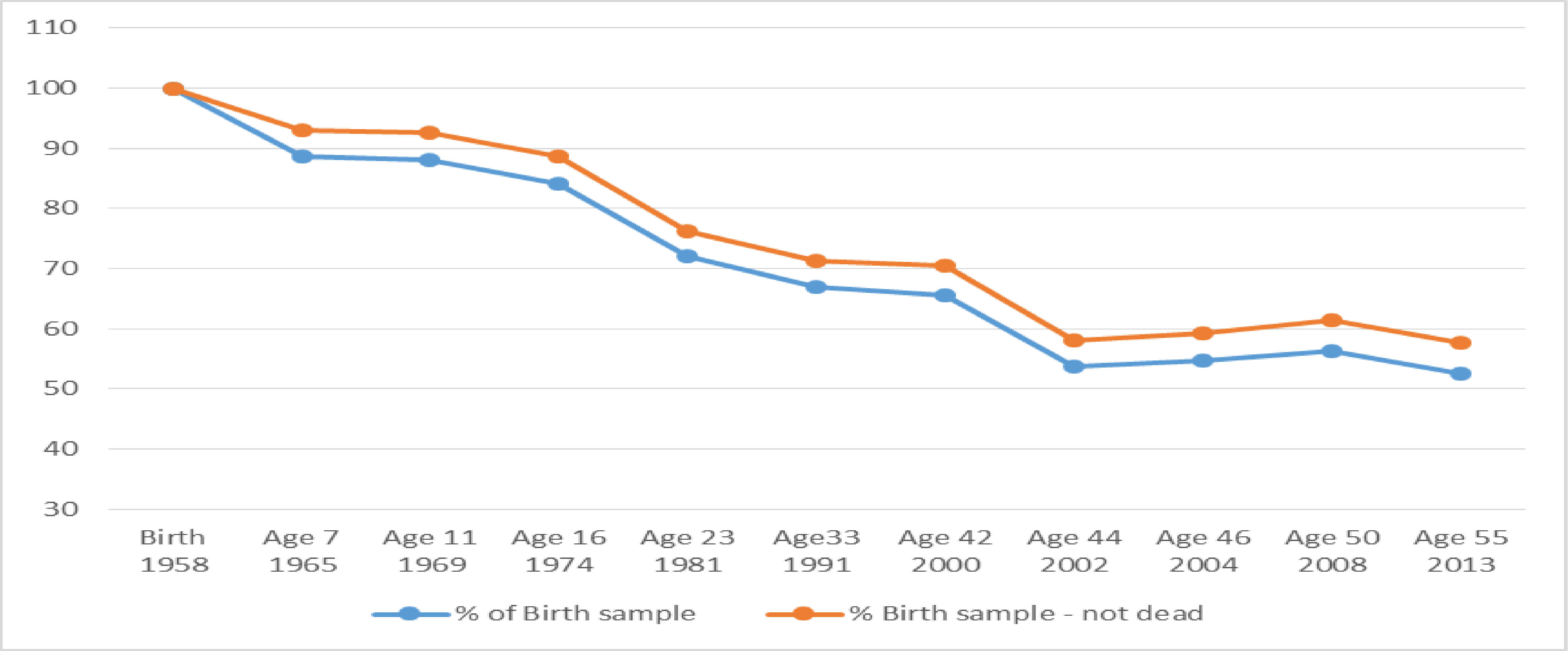
Birth	School years	Adult
Household composition	Household composition	Household composition
Parental social class	Parental social class	Employment
Obstetric history	Parental employment	Social class
Smoking in pregnancy	Financial circumstances	Income
Pregnancy (problems, antenatal care)	Housing	Housing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

Response in NCDS

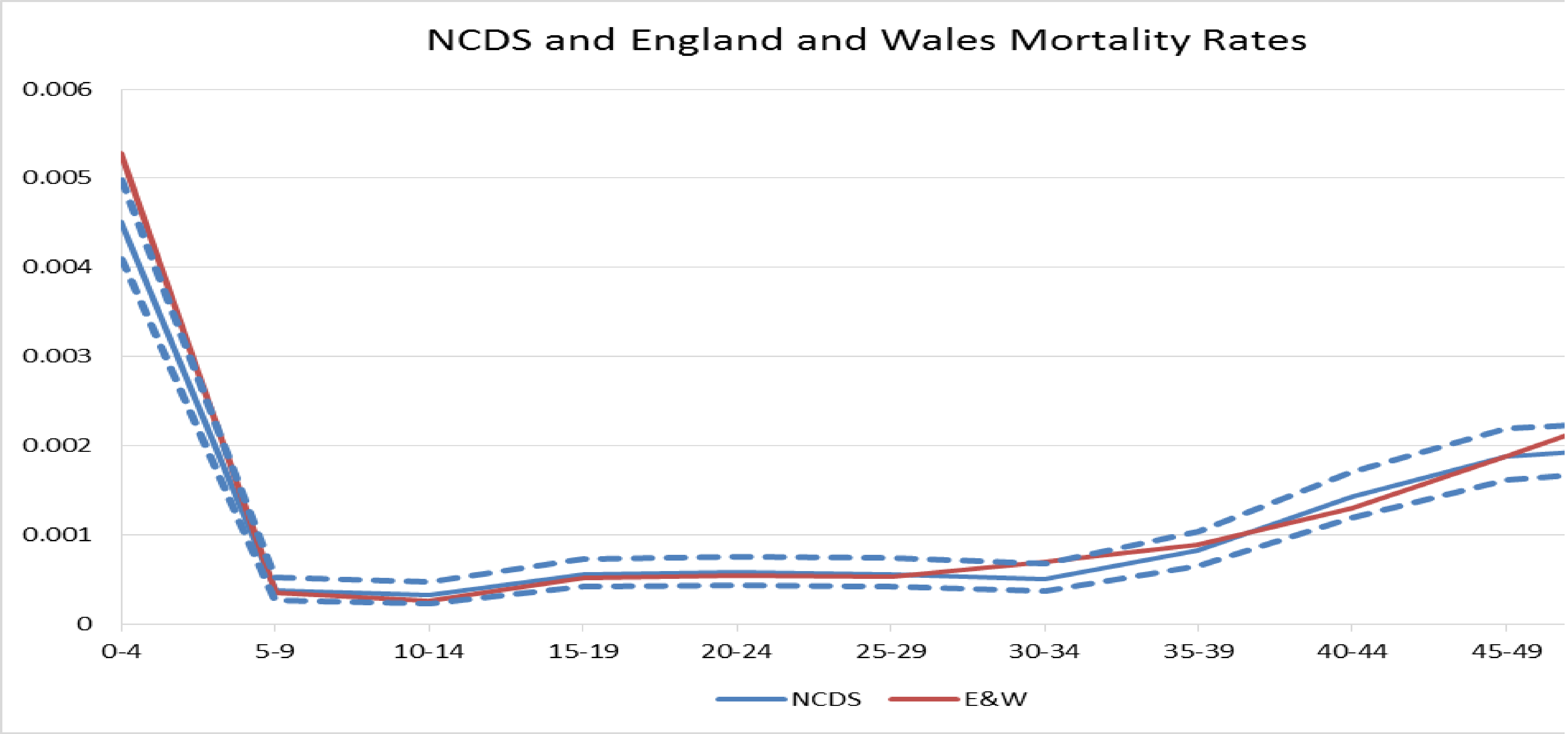


CENTRE FOR
LONGITUDINAL
STUDIES

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

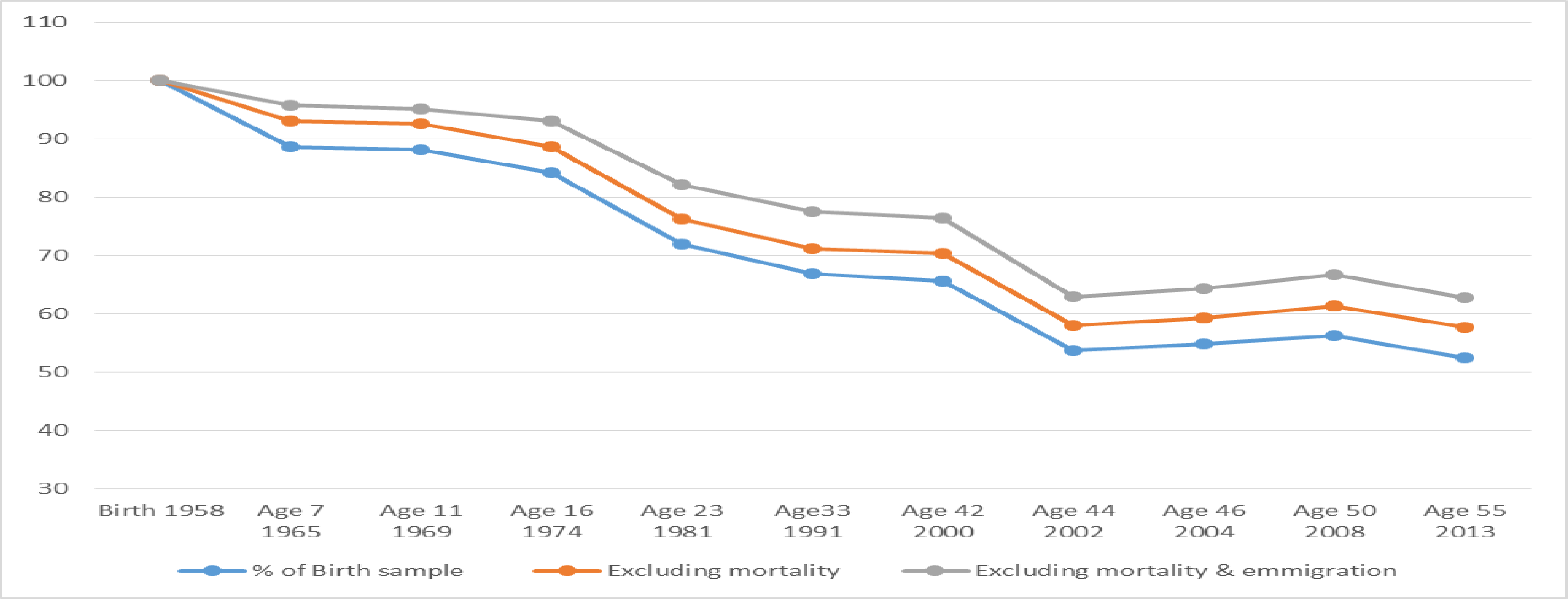


CENTRE FOR
LONGITUDINAL
STUDIES



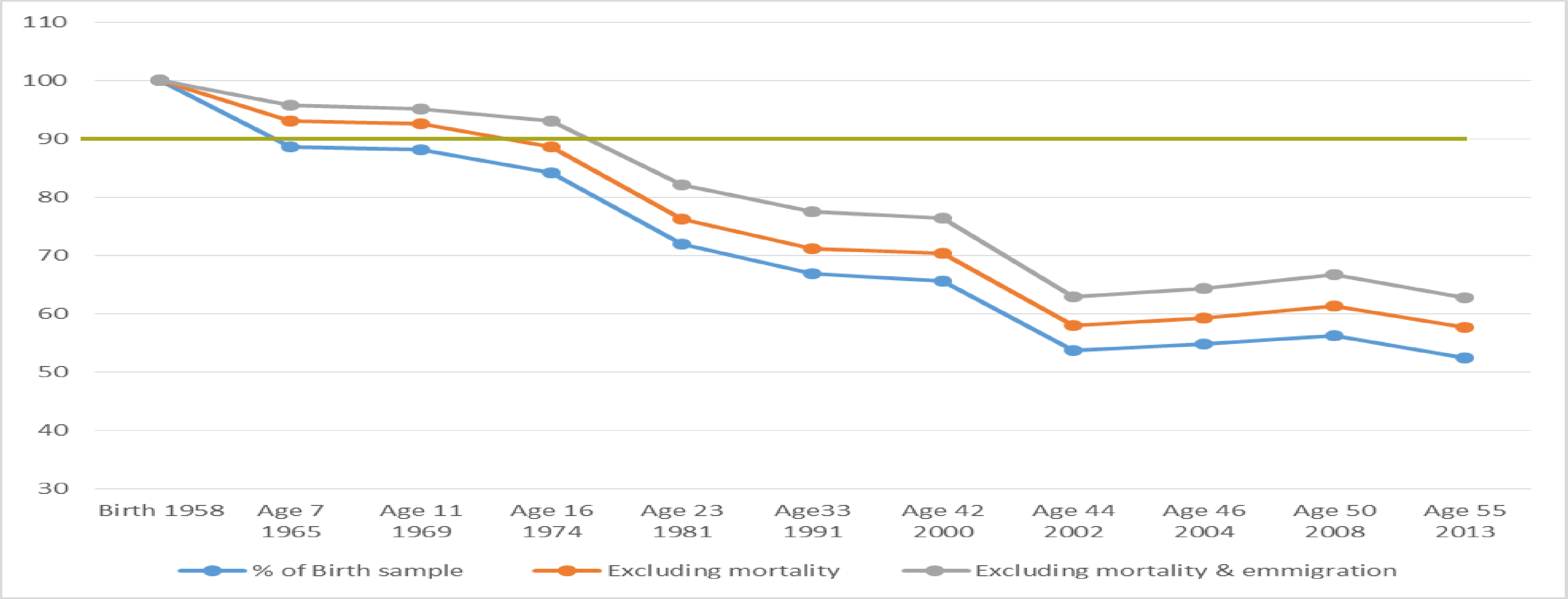
CENTRE FOR
LONGITUDINAL
STUDIES

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



CENTRE FOR
LONGITUDINAL
STUDIES

The 10% rule (of thumb)

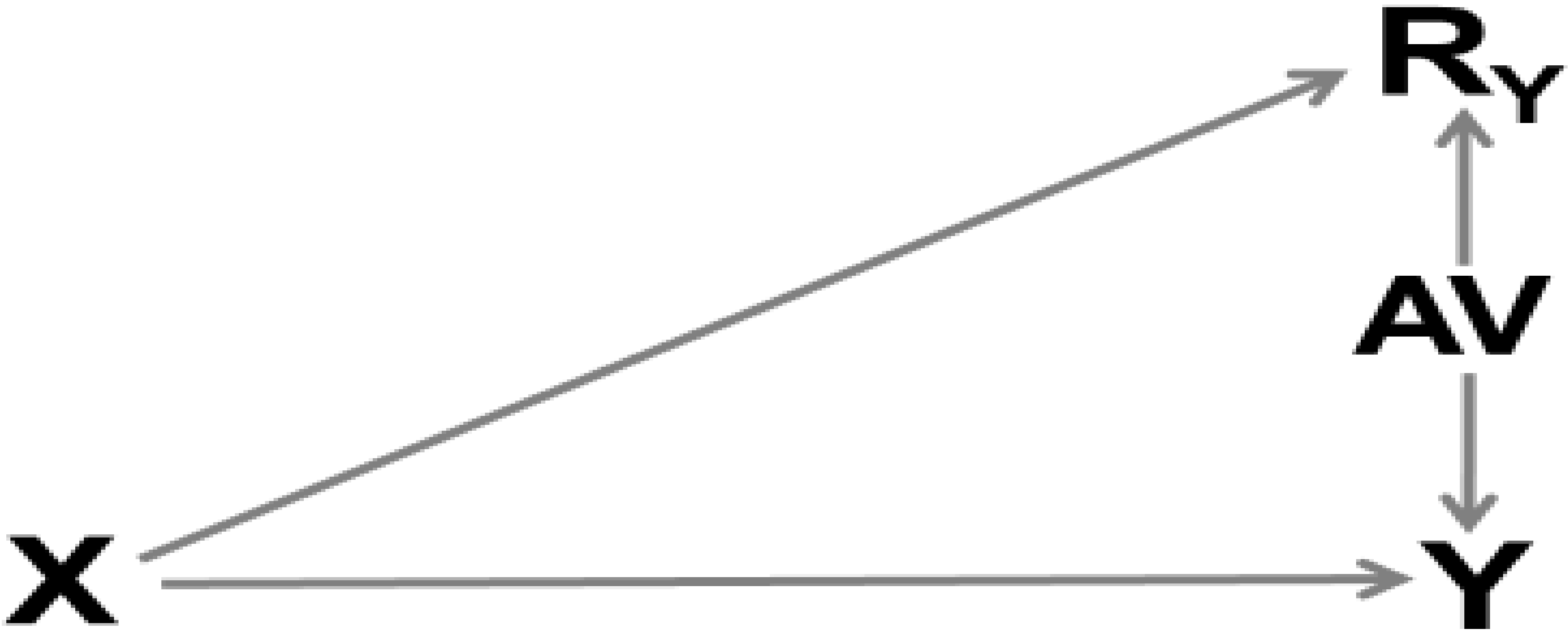


CENTRE FOR
LONGITUDINAL
STUDIES

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 within and between cohorts

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



CENTRE FOR
LONGITUDINAL
STUDIES

A data driven approach to maximise the plausibility of MAR

- About 17500 variables! => Selection is done in three stages

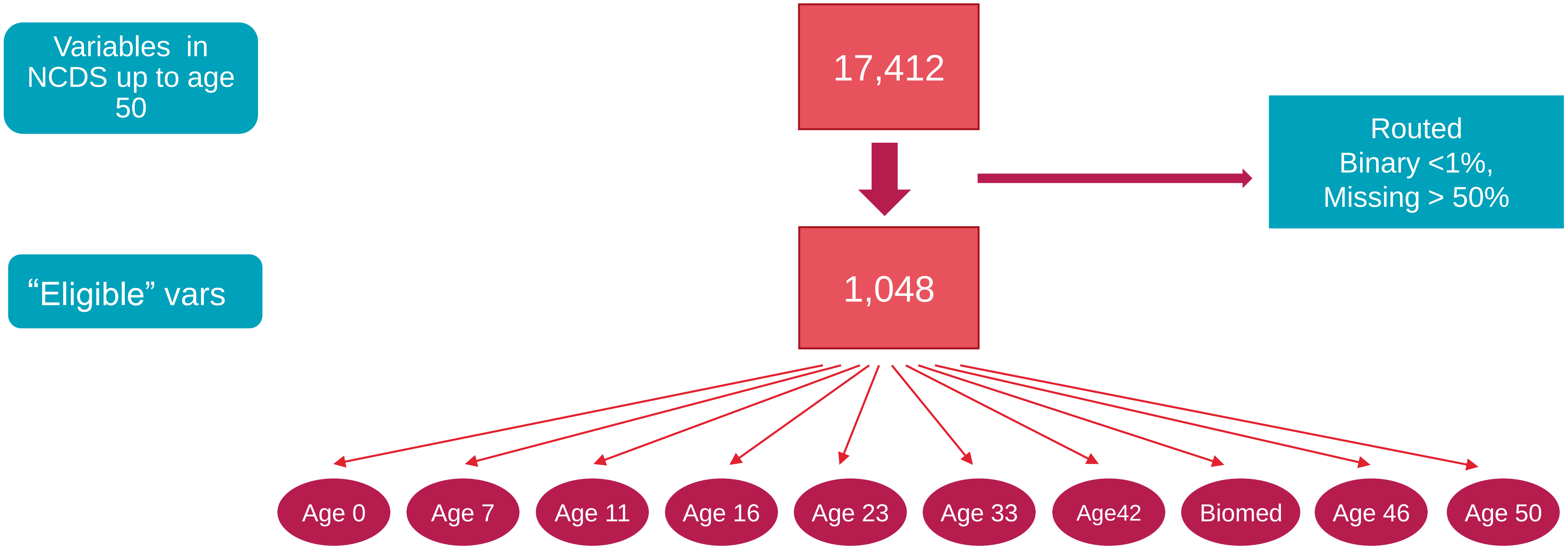
Pre – selection

- 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
- LASSO & Forward Stepwise

CENTRE FOR
LONGITUDINAL
STUDIES



	Age 0	Age 7	Age 11	Age 16	Age 23	Age 33	Age42	Biomed	Age 46	Age 50
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

CENTRE FOR LONGITUDINAL STUDIES

Outputs

- We will **not** make available imputed datasets
- List of auxiliary variables for users to adapt to their analysis
- Transparent assumptions so analysts can make an informed choice
- Straightforward approach, applicable to UK longitudinal surveys
- User guide & peer reviewed papers
- Stata, R and Mplus code
- Dynamic process, the results will be updated when new waves or other forms of data become available (paradata, data linkages)

CENTRE FOR
LONGITUDINAL
STUDIES

Thank you for your attention!

G.Ploubidis@UCL.ac.uk
@GeorgePloubidis

CENTRE FOR
LONGITUDINAL
STUDIES

Institute of Education

Maintaining representativeness by maximising the plausibility of the MAR assumption: Evidence from the 1958 British birth cohort

George B. Ploubidis, Benedetta Pongiglione,

Martina Narayanan & Brian Dodgeon

CENTRE FOR LONGITUDINAL STUDIES

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Rationale

- Occurrence of missing data threatens sample representativeness
- Direct link between missing data mechanisms and representativeness
- Does missing data handling (with principled methods) restore sample representativeness?
- Does maximising the plausibility of MAR (as we did in the 1958 British birth cohort) help?

CENTRE FOR
LONGITUDINAL
STUDIES

Today

- Missing data mechanisms and sample representativeness
- Two “experiments” - Results from the 1958 cohort

CENTRE FOR
LONGITUDINAL
STUDIES

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 vs MNAR in longitudinal surveys

- Some missing data patterns/variables may be MNAR even after the introduction of auxiliary variables (that is even after the plausibility of MAR is maximised).
- 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 observables restore/maintain representativeness?
- Does maximising MAR help with representativeness?

Two experiments

- In all surveys the data are MAR or MNAR
- 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
- In longitudinal studies we can also “travel back in time” and test whether the statistical properties of the baseline sample can be replicated
- Even when distributions are similar the target variables can <u>still be MNAR</u>, 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
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

Types of information covered

CENTRE FOR LONGITUDINAL STUDIES

Birth

Household composition
Parental social class
Obstetric history
Smoking in pregnancy
Pregnancy
(problems, antenatal care)
Labour
(length, pain relief, problems)
Birthweight, length

School years

Household composition
Parental social class
Parental employment
Financial circumstances
Housing
Health
Cognitive tests
Emotions and behaviour
School
Views and expectations
Attainment

Adult

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

Does maximising MAR help with representativeness?

- In the 1958 cohort we have identified all predictors of response ("auxiliary variables") using a data driven approach
- Plausibility of MAR maximised
- 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 of a target variable despite attrition?

“Experiment” 1

	Complete at birth			
Variable	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

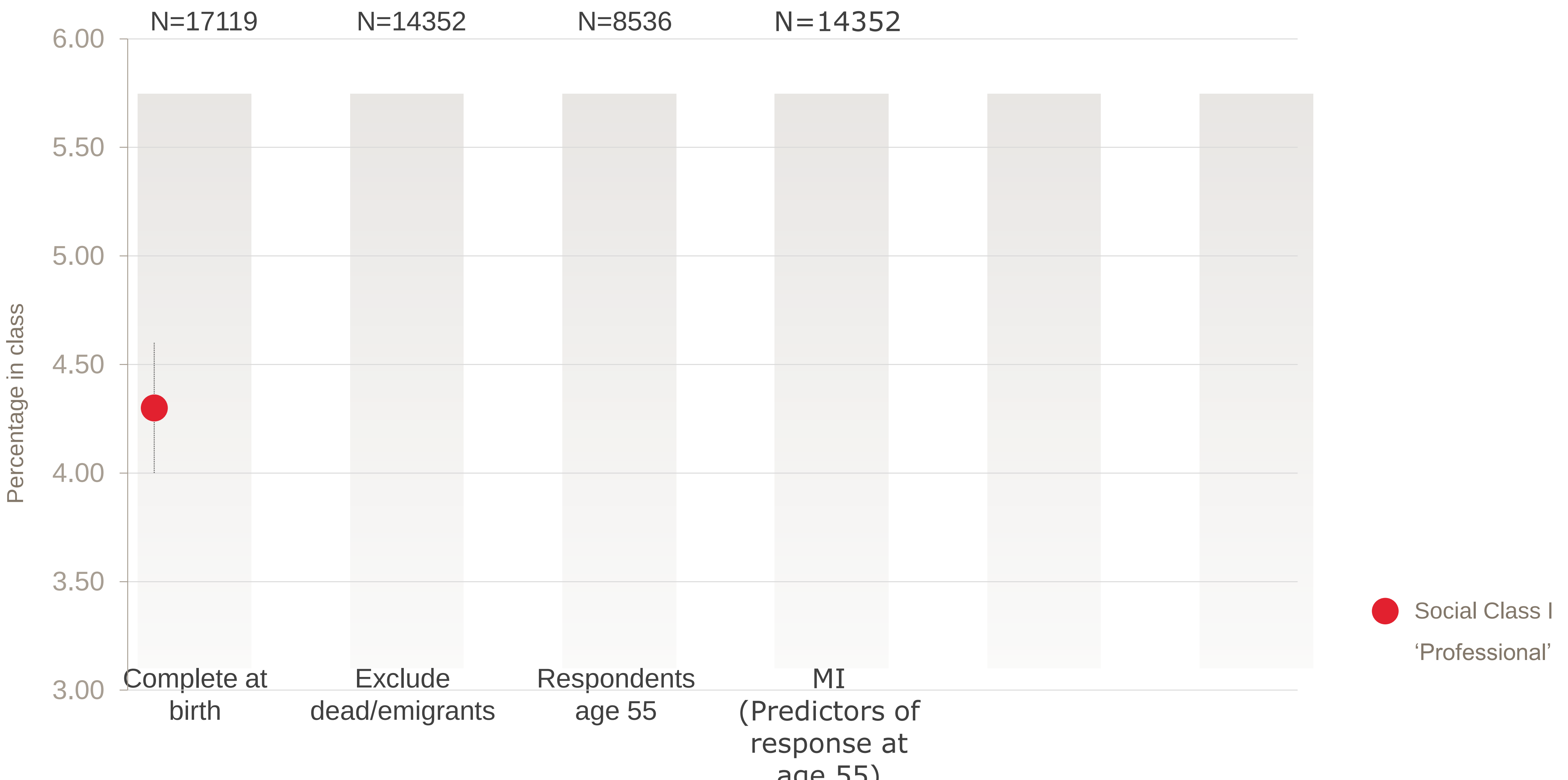
- Can we replicate the composition of Social Class at birth (N = 17119) with participants at age 55 (N = 8536)?
- Multiple Imputation with chained equations, 20 imputations using auxiliary variables

CENTRE FOR
LONGITUDINAL
STUDIES

Social Class of mother's husband 1958

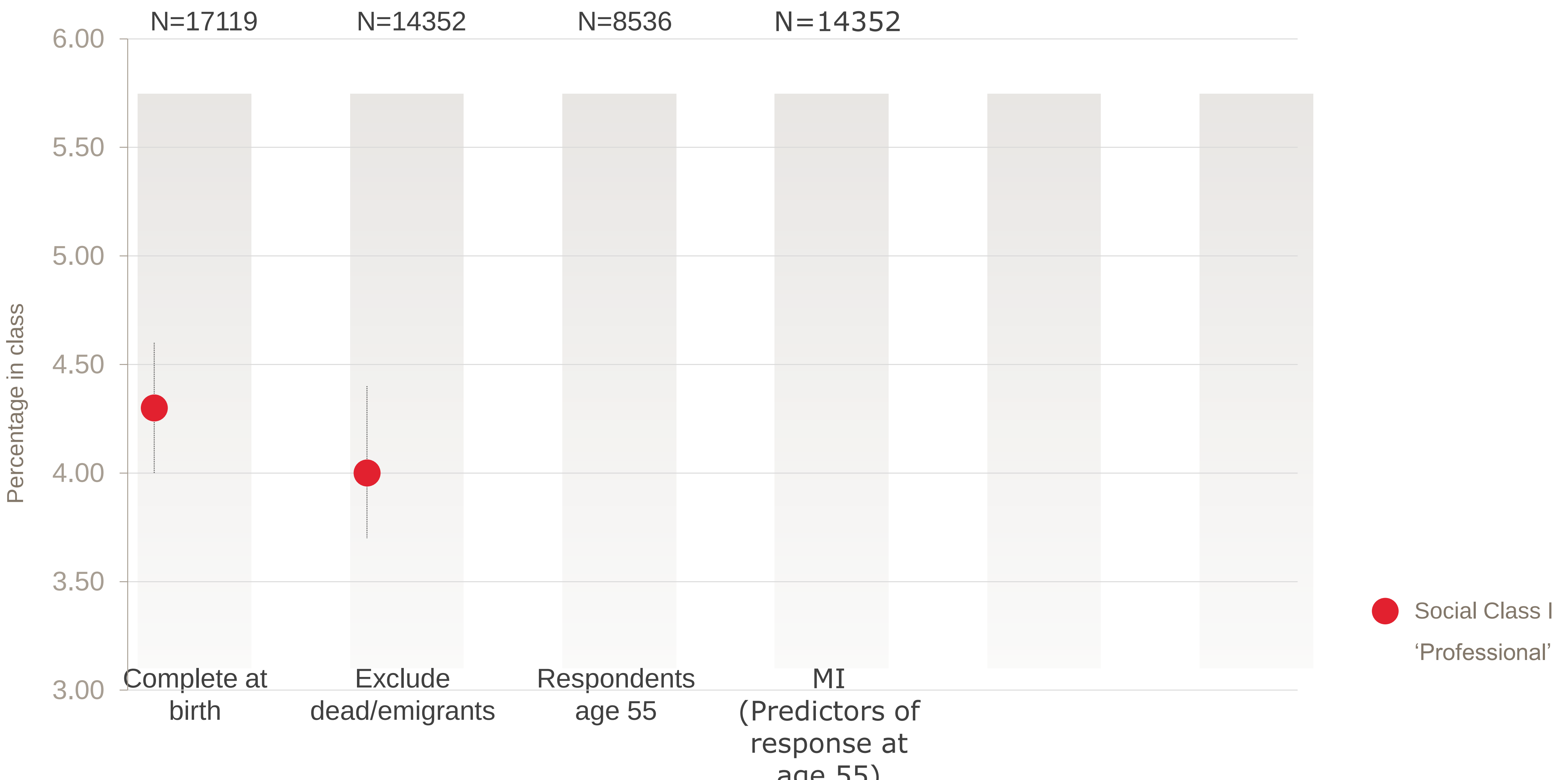
CENTRE FOR
LONGITUDINAL
STUDIES

Percentage in Social Class I



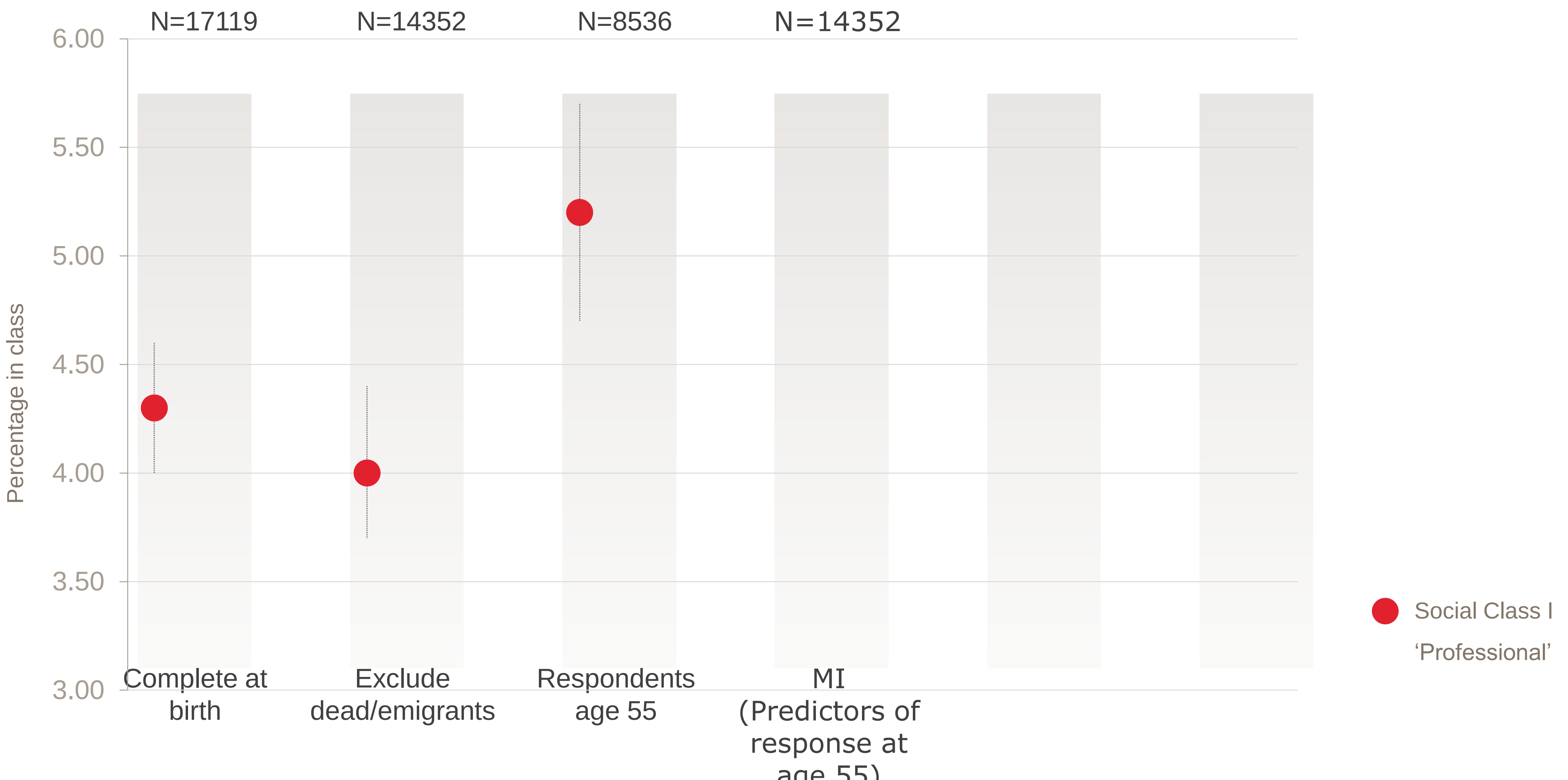
Social Class of mother’s husband 1958

Percentage in Social Class I



Social Class of mother’s husband 1958

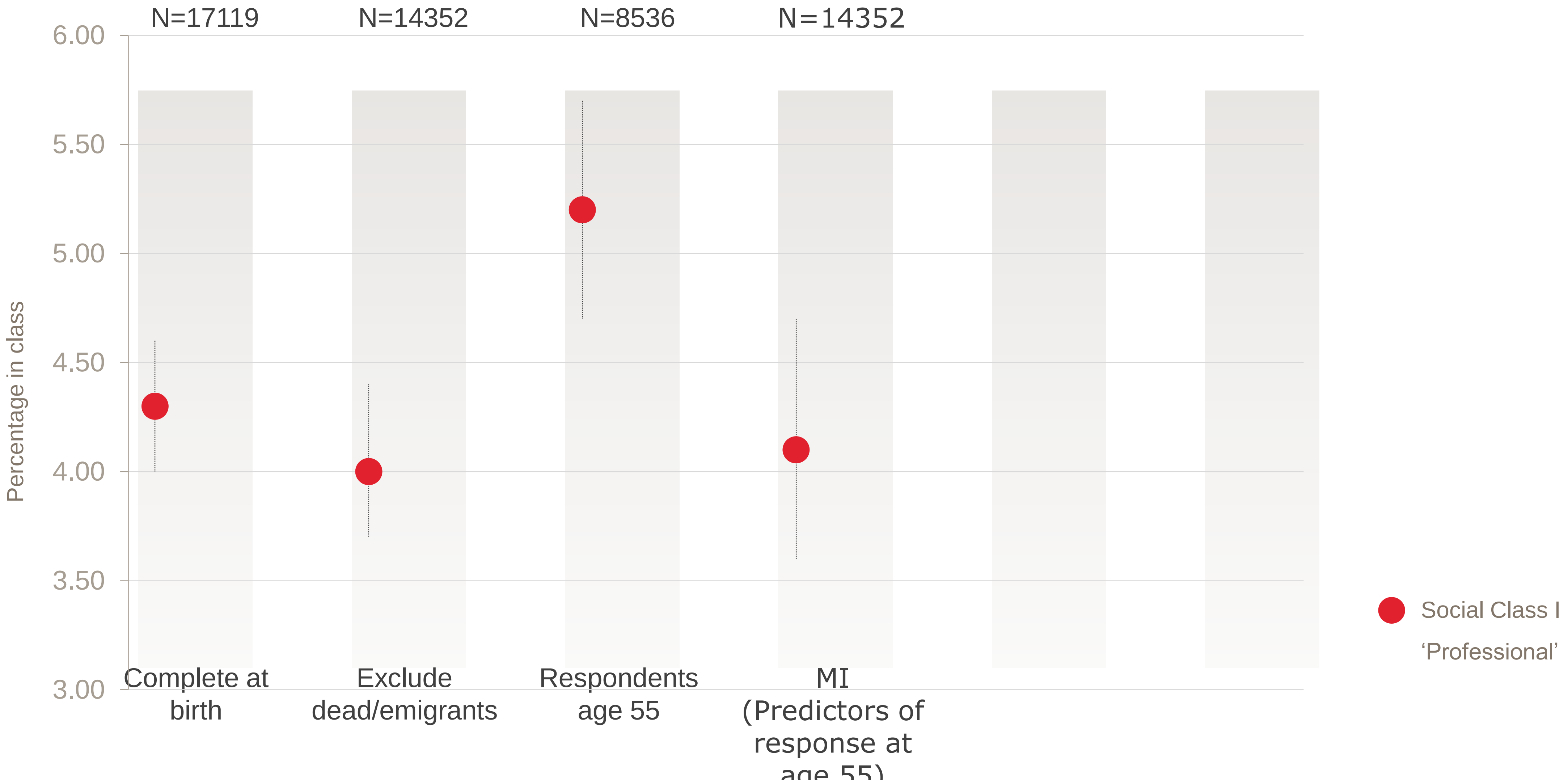
Percentage in Social Class I



Social Class of mother's husband 1958

CENTRE FOR LONGITUDINAL STUDIES

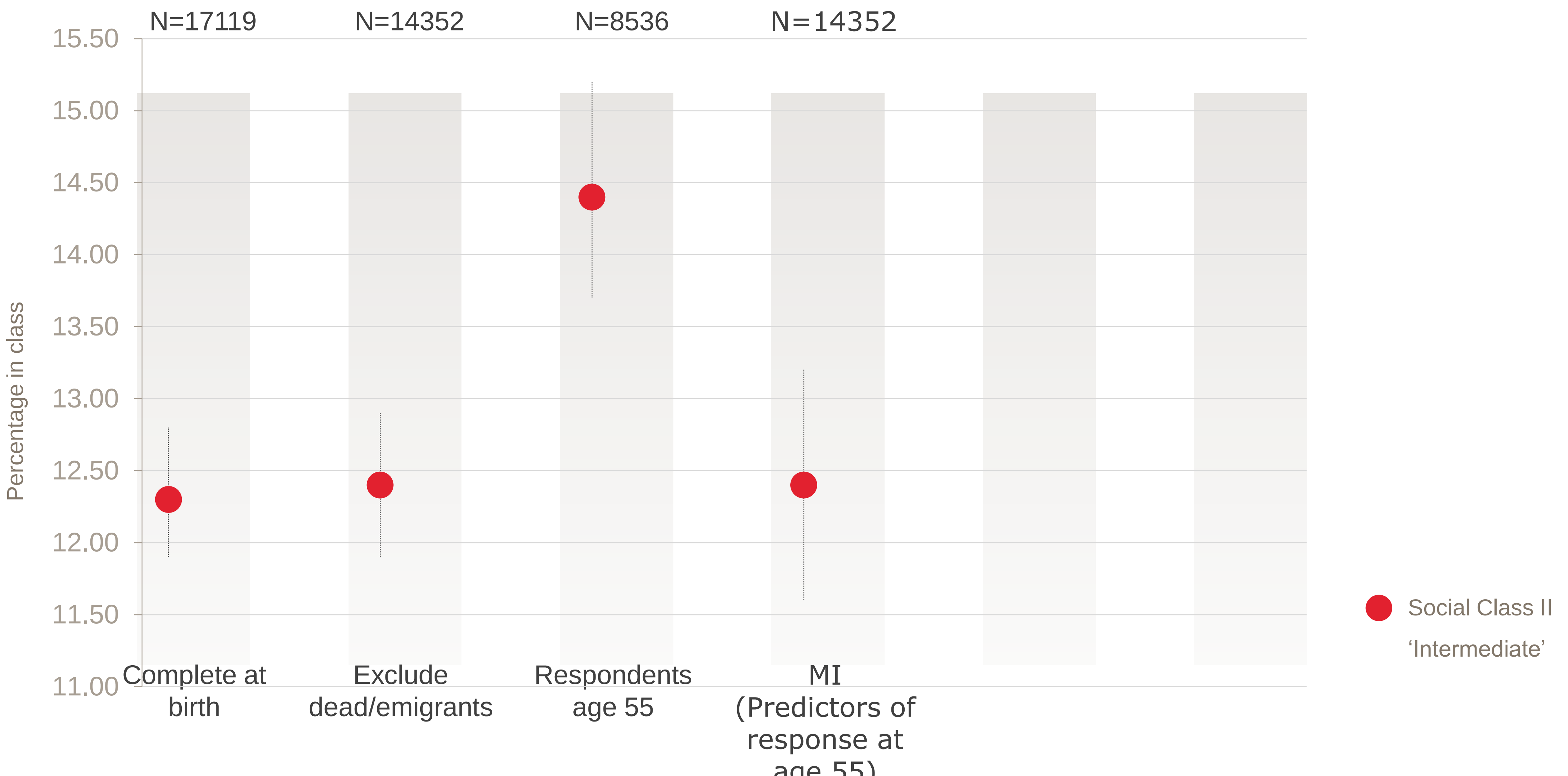
Percentage in Social Class I



Social Class of mother’s husband 1958

CENTRE FOR
LONGITUDINAL
STUDIES

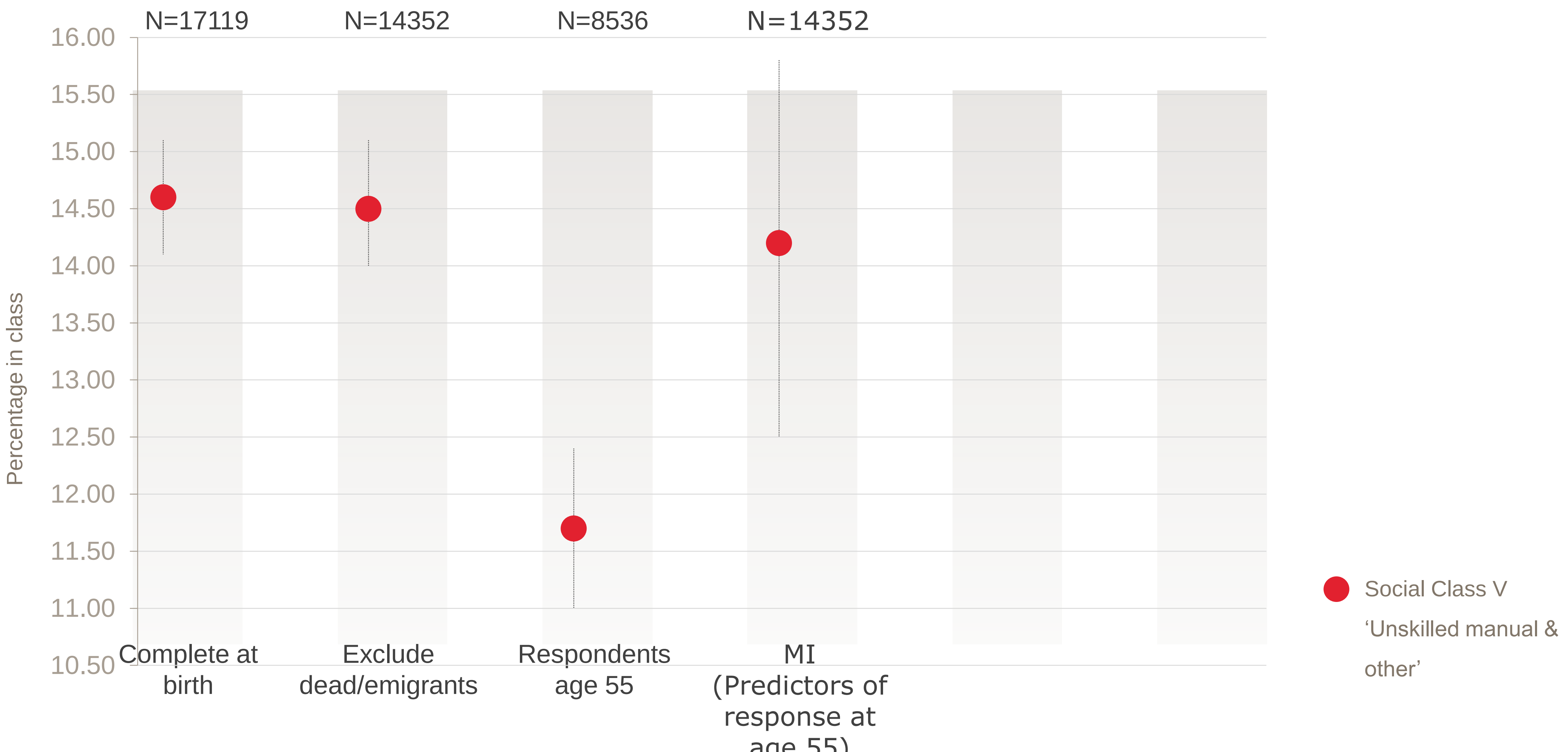
Percentage in Social Class II



Social Class of mother’s husband 1958

CENTRE FOR LONGITUDINAL STUDIES

Percentage in Social Class V



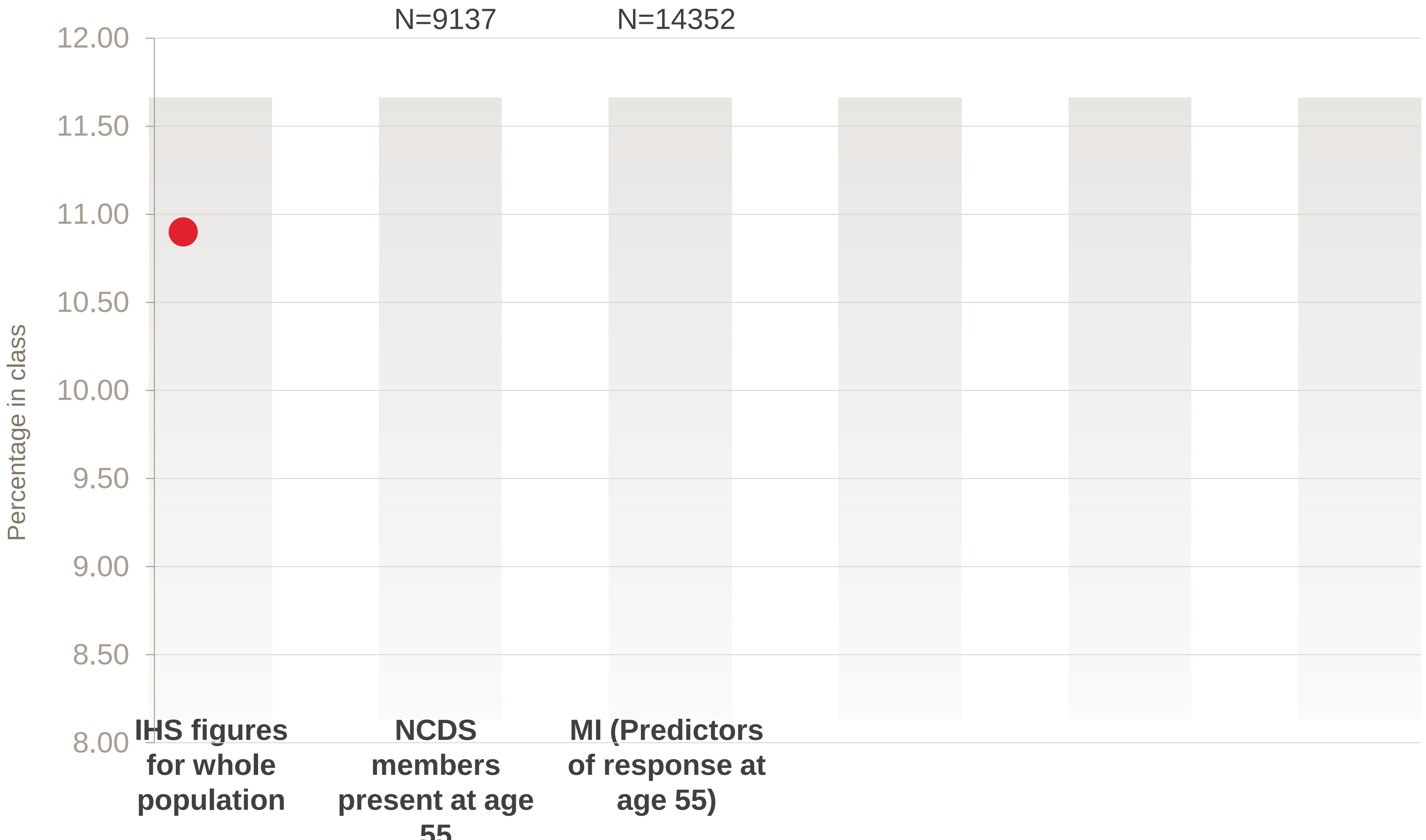
“Experiment” 2 - Marital status at age 55

- Recoded in three categories
  - Single, that is never married
  - Married/in a legal partnership and living with spouse
  - Separated/Divorced/Widowed
- Multiple Imputation (MI) with chained equations in Stata 14, 20 imputations
- “Known” population distribution form the Integrated Household Survey – Office for National Statistics
- Can we replicate the “known” population distribution after handling missing data with MI?

Marital Status at age 55

CENTRE FOR
LONGITUDINAL
STUDIES

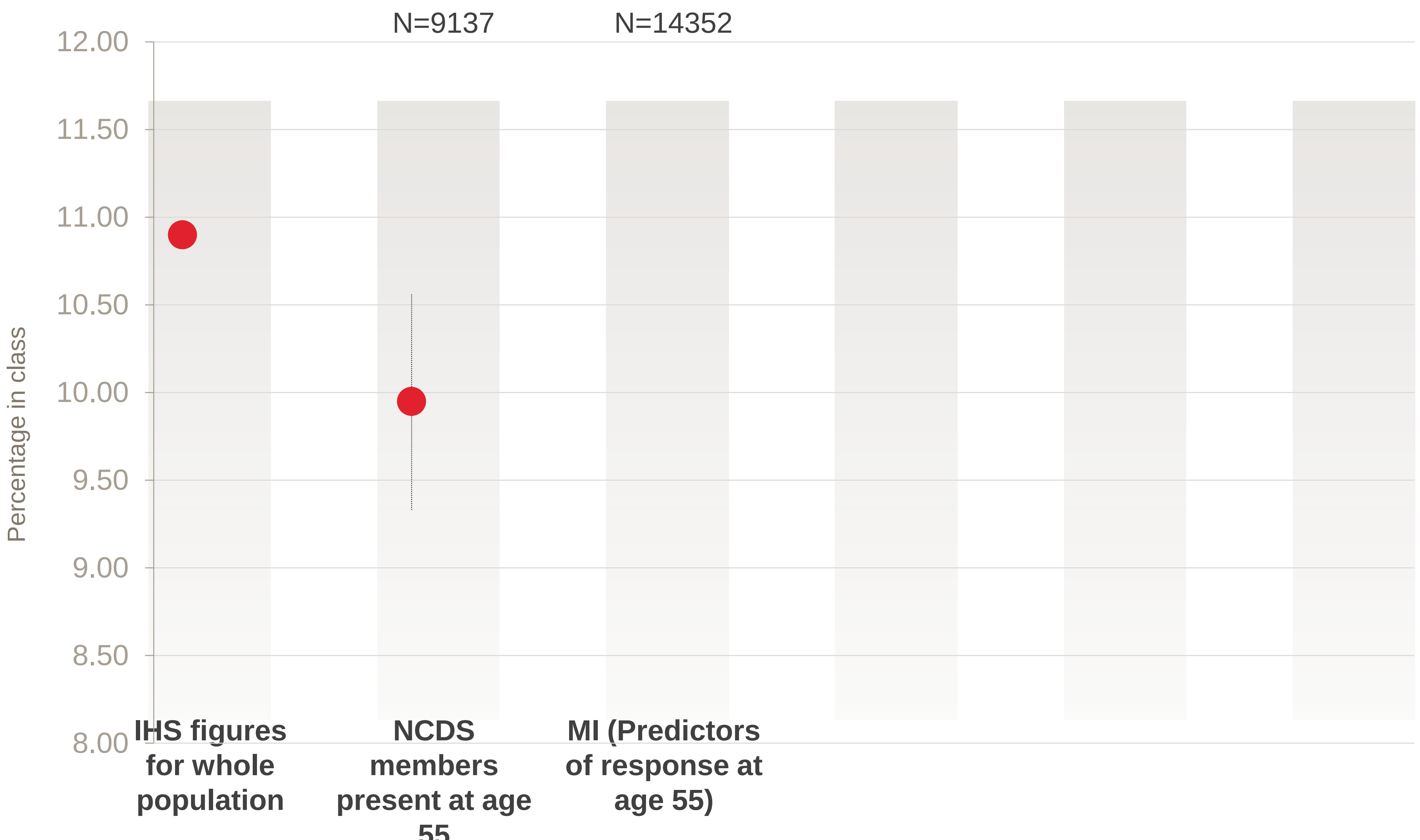
Percentage single (never-married)



Marital Status at age 55

CENTRE FOR LONGITUDINAL STUDIES

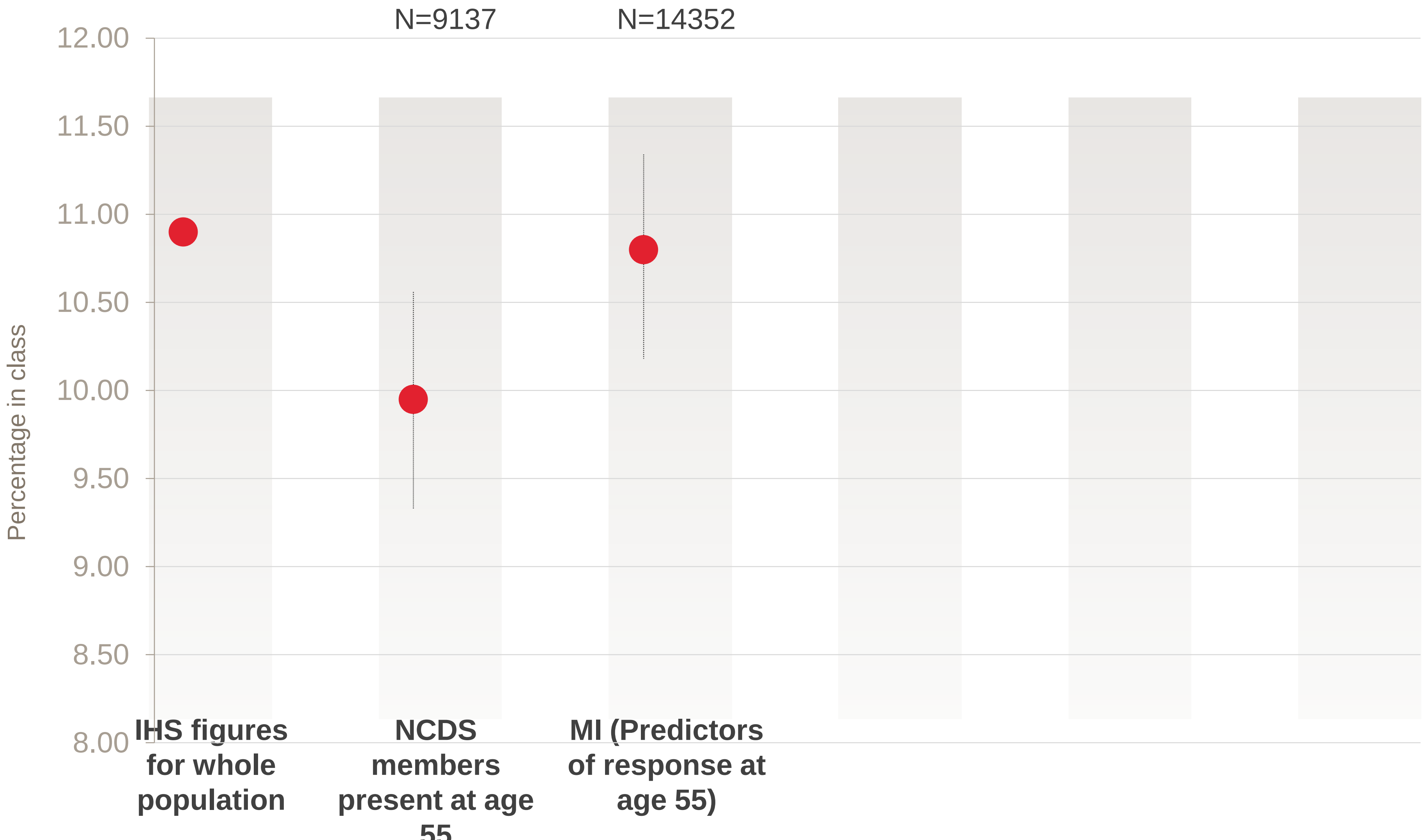
Percentage single (never-married)



Marital Status at age 55

CENTRE FOR LONGITUDINAL STUDIES

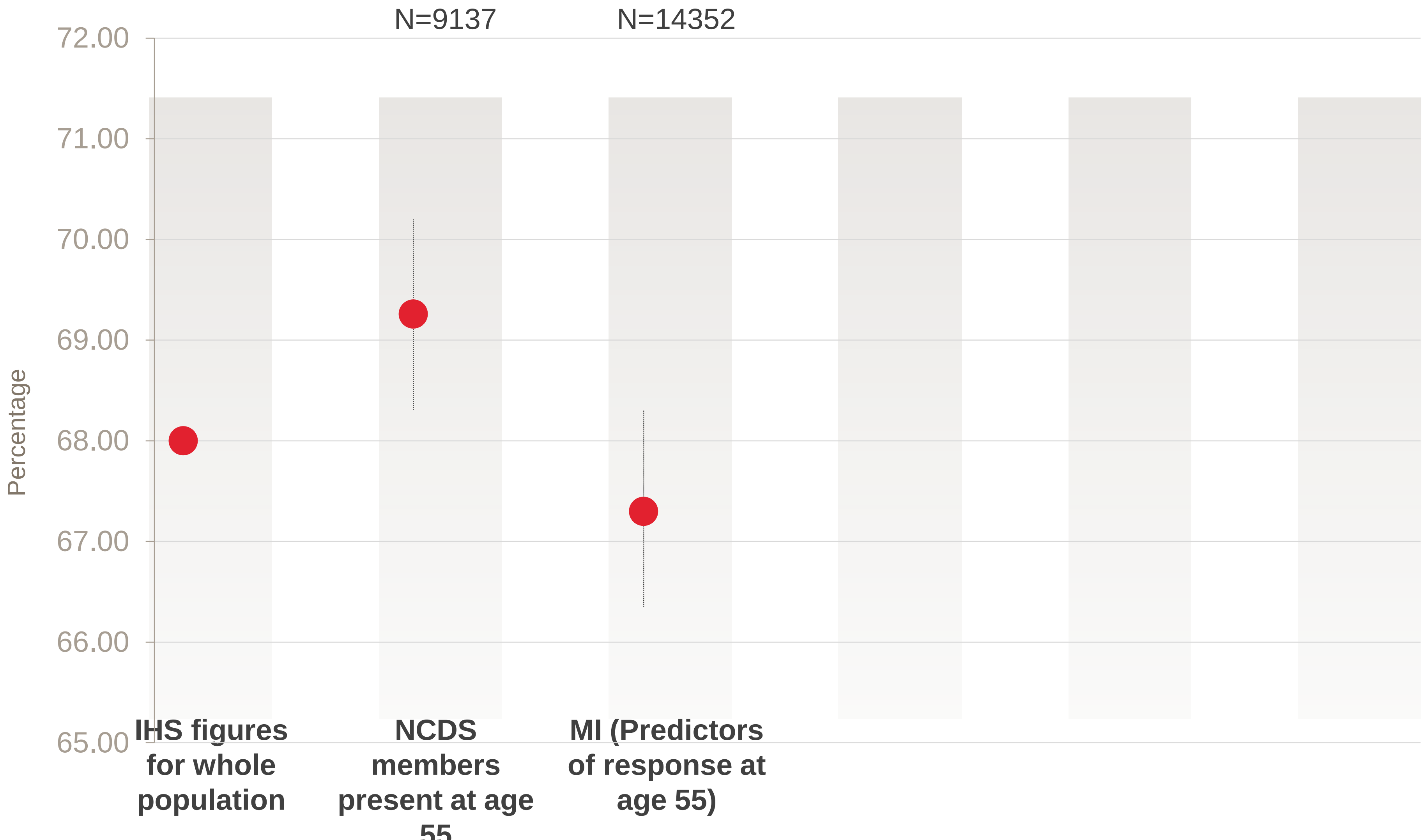
Percentage single (never-married)



Marital Status at age 55

CENTRE FOR
LONGITUDINAL
STUDIES

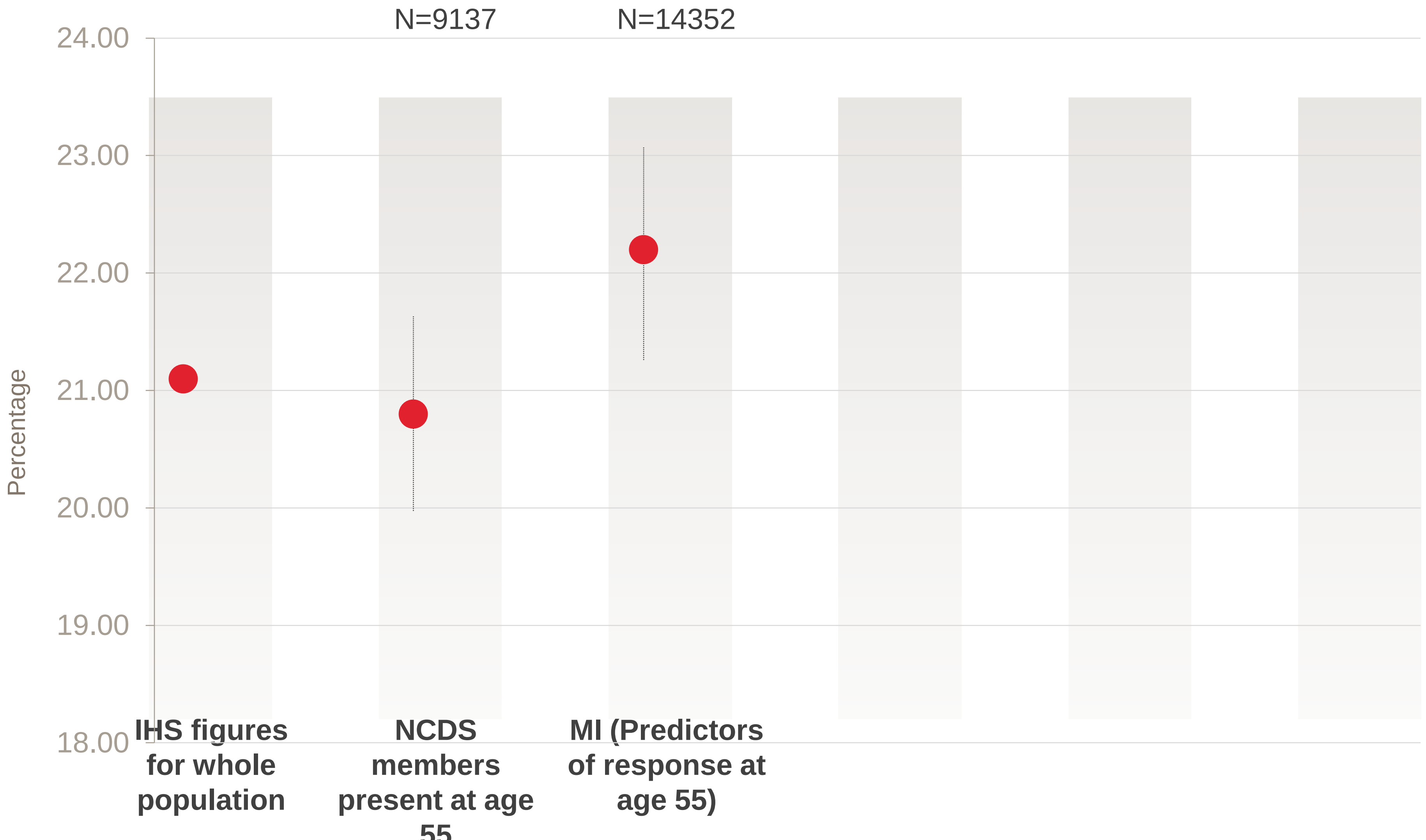
Percentage married and living with spouse/legal partner



Marital Status at age 55

CENTRE FOR LONGITUDINAL STUDIES

Percentage separated/divorced/widowed



More ‘known population distributions’

- Figures have been obtained from Office for National Statistics to enable us to compare distributions of key variables in the Integrated HOUshod Survey (IHS), the Annual Population Survey (APS) and Labour Force Survey (LFS) with the corresponding variables in NCDS.
- This enables us to check sample representativeness further
- May also be able to compare distributions with Census-based longitudinal datasets like the ONS Longitudinal Study (ONS-LS)
- Plans are in place to receive income distributions from HMRC

More ‘known population distributions’

Variable	Source	Age 42 in 2000	Age 50 in 2008	Age 55 in 2013
Marital Status	IHS			✓
	LFS	✓	✓	
	APS		✓	✓
No. Dep. Children in Hhold	LFS	✓	✓	
	APS		✓	✓
Social Class	LFS	✓	✓	
	APS		✓	✓
Highest Qualification	LFS	✓	✓	
	APS		✓	✓
Income	HMRC (soon to be acquired)	✓	✓	✓

IHS= Integrated Household Survey

APS= Annual Population Survey

LFS= Labour Force Survey

HMRC=Her Majesty’s Revenue & Customs

Conclusion

- Maximising the plausibility of the MAR assumption with observed data has the potential to restore/maintain sample representativeness in longitudinal surveys
- Reassuring for substantive research
- Not a test for MAR vs MNAR
- Bias due to missing data still possible

Thank you for your attention!

Variables in the imputation model

CENTRE FOR
LONGITUDINAL
STUDIES

14 variables

"Outcome" (1)

1. Marital status

Baseline "complete" (5)

1. Birthweight
2. Smoking prior to pregnancy
3. Social class mother's husband at birth
4. Mother worked before age 5
5. Reading test score at age 7

Predictors of marital status (2)

1. Income at age 42
2. BMI at age 23

Auxiliaries associated with marital status and response (5)

1. Special education age 11
2. Ask partner for help age 46
3. Education/apprent./training age 23
4. Legal status age 42
5. Member of organisation age 42

Strongest predictor of response at age 55 (1)

1. Accommodation rented or owned age 42

Institute of Education

Missing data strategy:
A step by step guide

Martina Narayanan, Benedetta Pongiglione, Brian Dodgeon & George Ploubidis

NCDS conference, March 9th 2018

CENTRE FOR LONGITUDINAL STUDIES

Using a concrete example for our step by step guide:

- “Does age at first birth predict biomarkers at age 44?”
- NCDS Paper by Maria Sironi, George B. Ploubidis & Emily Grundy
- Complete case analyses including all covariates would comprise only 2,506 individuals
- Multiple Imputation: Sample of 11,754 respondents

Substantive model

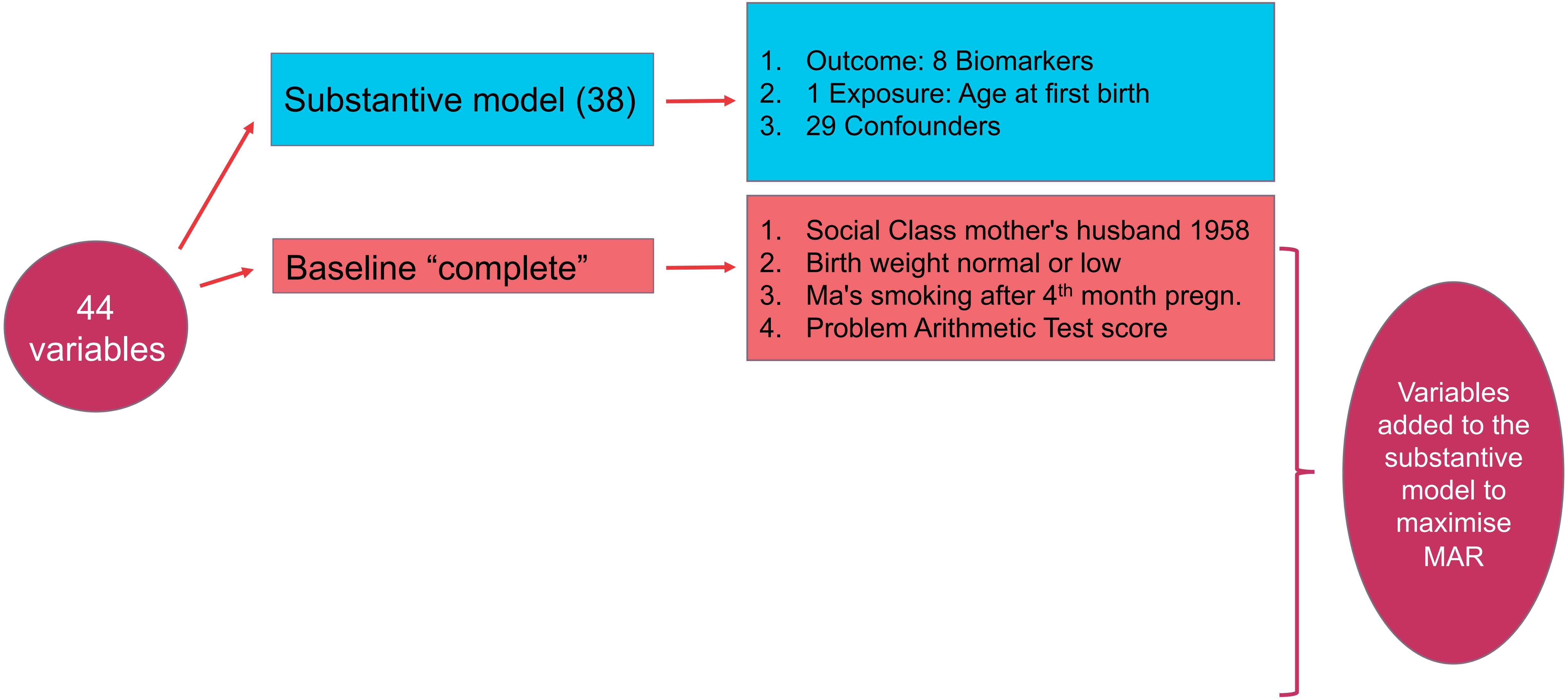
44 variables

Substantive model (38)

1. Outcome: 8 Biomarkers
2. 1 Exposure: Age at first birth
3. 29 Confounders

Add baseline “complete”

CENTRE FOR LONGITUDINAL STUDIES



Add predictors of outcome

44 variables

Substantive model (38)

1. Outcome: 8 Biomarkers
2. 1 Exposure: Age at first birth
3. 29 Confounders

Baseline “complete”

1. Social Class mother's husband 1958
2. Birth weight normal or low
3. Ma's smoking after 4th month pregn.
4. Problem Arithmetic Test score

Predictors of biomarkers (1)

1. Self rated health (age 33)

Variables added to the substantive model to maximise MAR

Add auxiliary variables

44 variables

Substantive model (38)
1. Outcome: 8 Biomarkers
2. 1 Exposure: Age at first birth
3. 29 Confounders

Baseline "complete"
1. Social Class mother's husband 1958
2. Birth weight normal or low
3. Ma's smoking after 4th month pregn.
4. Problem Arithmetic Test score

Predictors of biomarkers (1)
1. Self rated health (age 33)

Auxiliaries also associated with biomarkers (4)
1. Special Educ Treatment (age 11)
2. General Motor Handicap (age 16)
3. Current main activity (age 42)
4. Own house (age 23)

Variables added to the substantive model to maximise MAR

How do we identify good auxiliary variables?

CENTRE FOR
LONGITUDINAL
STUDIES

Auxiliary variable = predictive of non-response **as well as** of our specific outcome variable(s)

1. Look at the list of predictors of non-response at age 44
2. See how those variables predict biomarkers at age 44 (using regression, correlation etc.)
3. Choose variables that are strong predictors of both non-response **and** biomarkers

Add auxiliary variables

CENTRE FOR LONGITUDINAL STUDIES

44 variables

Substantive model (38)

1. Outcome: 8 Biomarkers
2. 1 Exposure: Age at first birth
3. 29 Confounders

Baseline "complete"

1. Social Class mother's husband 1958
2. Birth weight normal or low
3. Ma's smoking after 4th month pregn.
4. Problem Arithmetic Test score

Predictors of biomarkers (1)

1. Self rated health (age 33)

Auxiliaries also associated with biomarkers (4)

1. Special Educ Treatment (age 11)
2. General Motor Handicap (age 16)
3. Current main activity (age 42)
4. Own house (age 23)

Variables added to the substantive model to maximise MAR

Add strongest predictor of non-response

CENTRE FOR LONGITUDINAL STUDIES

44 variables

Substantive model (38)

1. Outcome: 8 Biomarkers
2. 1 Exposure: Age at first birth
3. 29 Confounders

Baseline “complete”

1. Social Class mother's husband 1958
2. Birth weight normal or low
3. Ma's smoking after 4th month pregn.
4. Problem Arithmetic Test score

Predictors of biomarkers (1)

1. Self rated health (age 33)

Auxiliaries also associated with biomarkers (4)

1. Special Educ Treatment (age 11)
2. General Motor Handicap (age 16)
3. Current main activity (age 42)
4. Own house (age 23)

Strongest predictor of response at age 44 (1)

1. Currently member of Union/Staff Assoc (age 42)

Variables added to the substantive model to maximise MAR

Multiple Imputation

1. Setup
2. Run
3. Analyse

Multiple Imputation: Setup


```
mi set wide
/* tells Stata that data are in wide format */

mi register imputed V1 V2 V3 V4 V5 V6 V7 V8
/* specify which variables to include in the imputation */

mi set M=50
/* specify how many imputations to run */
```

Multiple Imputation: Setup for specific example


```
mi set wide

mi register imputed logfib logcrp hba1c cholratio highbp obese whr_10 htfev ///
age_child1_y_cat age_lastchild_y_cat num_child46_4 ///
age_int edu_lvl3 num_marr_cat num_cohab_cat months_unemp7801 sc0_manual ///
finhard11 overcrowd11 housing11 birthweight smoke_preg bsga_tot11 ///
parint01_edu teen_smoke rutterpd7 rutterpd11 num_hsptl11 offschool11_1m ///
genability11 female eduyrs_p stayschool_m fam_diff7 a10_divorce enuresis7 ///
enuresis11 phycoord11 ///
srh_33 n1476_01 unionmem_2 mothandicap curract type_acc

mi set M=50
```

Multiple Imputation: Run

CENTRE FOR
LONGITUDINAL
STUDIES

```
mi impute chained /* this will run the imputation using chained equations */ ///
(regress) V1 V2 /* specify (regress) in front of all continous variables */ ///
(logit) V5 V6 /* specify (logit) for dichotomous variables */ ///
(ologit) V3 V7 /* specify (ologit) for ordinal variables */ ///
(mlogit) V4 V8 /* specify (mlogit) for nominal variables */ ///
, replace noisily augment /* replaces old imputations (if there are any) */
                          /* noisily returns output for each imputation */
                          /* augment adds variance in imputation phase */

save "MIdatafile.dta", replace /* Save imputed data for later use */
```

Multiple Imputation: Run on specific example

CENTRE FOR
LONGITUDINAL
STUDIES

```
mi impute chained (regress) logfib logcrp hba1c cholratio whr_10 htfev ///
age_int months_unemp7801 birthweight bsga_tot11 rutterpd7 rutterpd11 ///
genability11 (ologit) num_child46_4 eduyrs_p num_hsptl11 age_child1_y_cat ///
age_lastchild_y_cat edu_lvl3 num_marr_cat num_cohab_cat srh_33 ///
(mlogit) curract (logit) obese highbp sc0_manual finhard11 overcrowd11 ///
housing11 smoke_preg parint01_edu offschool11_1m teen_smoke female ///
stayschool_m fam_diff7 a10_divorce enuresis7 enuresis11 phycoord11 n1476_01 ///
unionmem_2 mothandicap type_acc, replace noisily augment

save "MIdatafile.dta", replace
```

Multiple Imputation: What happens after running it?

Variables

Enter filter tex

Name	Label
V1	Variable 1
V2	Variable 2
V3	Variable 3
V4	Variable 4
_mi_miss	
_1_V1	
_2_V1	
_3_V1	
_4_V1	
_5_V1	
_1_V2	
_2_V2	
_3_V2	
_4_V2	
_5_V2	
_1_V3	
_2_V3	

mi impute creates a variable '_mi_miss' which is needed to run analyses on the imputed data!

In addition you will find the imputed values for each variable and each imputation.

Multiple Imputation: Analyse

CENTRE FOR
LONGITUDINAL
STUDIES

```
/* mi estimate: before an analysis tells Stata to use imputed data */

mi estimate: reg V1 V2 V3

mi estimate: logit V4 V2 V3
```

Multiple Imputation: Analyse for specific example

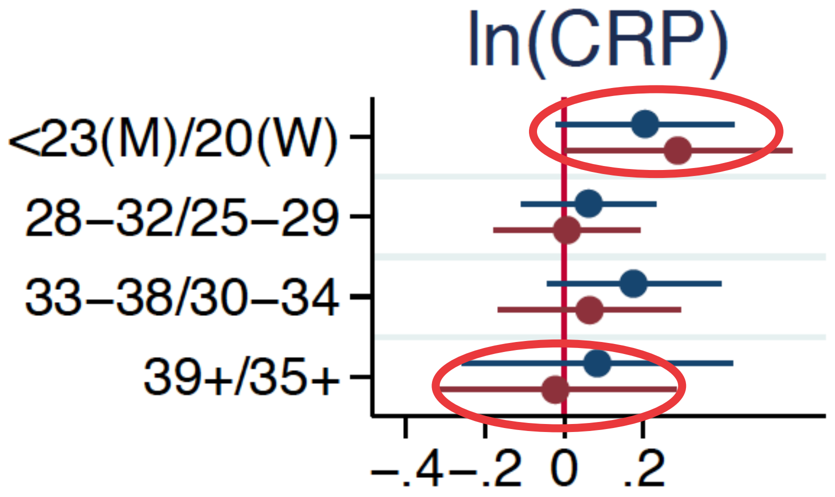
CENTRE FOR LONGITUDINAL STUDIES

```
mi estimate: reg cholratio ib2.age_child1_y_cat i.num_child46_4 age_int ///
i.edu_lvl3 i.num_marr_cat i.num_cohab_cat months_unemp7801 sc0_manual ///
finhard11 overcrowd11 housing11 birthweight smoke_preg bsga_tot11 ///
parint01_edu teen_smoke rutterpd7 rutterpd11 eduyrs_p stayschool_m fam_diff7 ///
a10_divorce num_hsptl11 offschool11_1m genability11 enuresis7 enuresis11 ///
phycoord11 if female==0
```

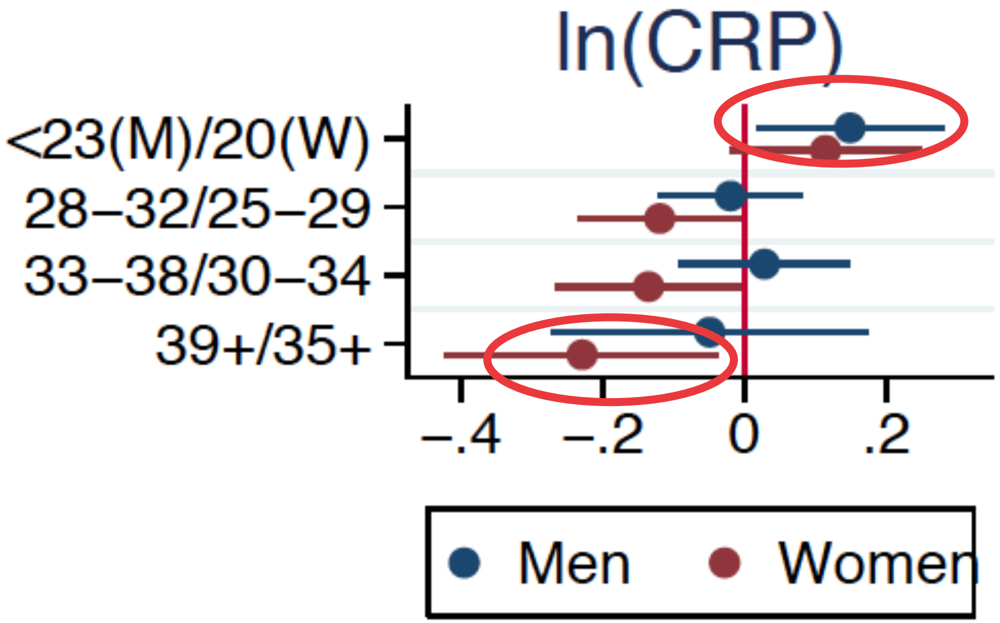
Comparing results: C-reactive Protein
(indicator for inflammation in body)

CENTRE FOR
LONGITUDINAL
STUDIES

Complete case analysis



Multiple imputed data



Conclusion

CENTRE FOR LONGITUDINAL STUDIES

- Multiple imputation is feasible
- Differing results between complete case analyses and multiple imputation

Institute of Education

THANK YOU FOR YOUR ATTENTION

m.narayanan@ucl.ac.uk

CENTRE FOR LONGITUDINAL STUDIES

Comparability of characteristics 1958 and 1970 birth cohort data with linked census ONS Longitudinal Study respondents born in 1958 and 1970

Nicola Shelton, Rachel Stuchbury,

Gemma Archer, Wei Xun: CeLSIUS, UCL

Research Questions

- 1) Is the most recent sweep of 1970/1958 cohort comparable to an equivalent ONS LS sample?
- 2) Are longitudinal associations between key sociodemographic factors and general health outcomes comparable between the two LS and 1958 datasets?

Samples

- ‘Matched’
  - 1958 – Excluding Scotland and those non-resident in UK
  - ONS LS – Including only those born in 1958 in England or Wales

- ‘Max’
  - 1958/1970 – All included
  - ONS LS – Including only those born in 1958/1970

CeLSIUS
Centre for Longitudinal Study Information & User Support

The ONS Longitudinal study

Individual and household level microdata

◦ → 1% sample

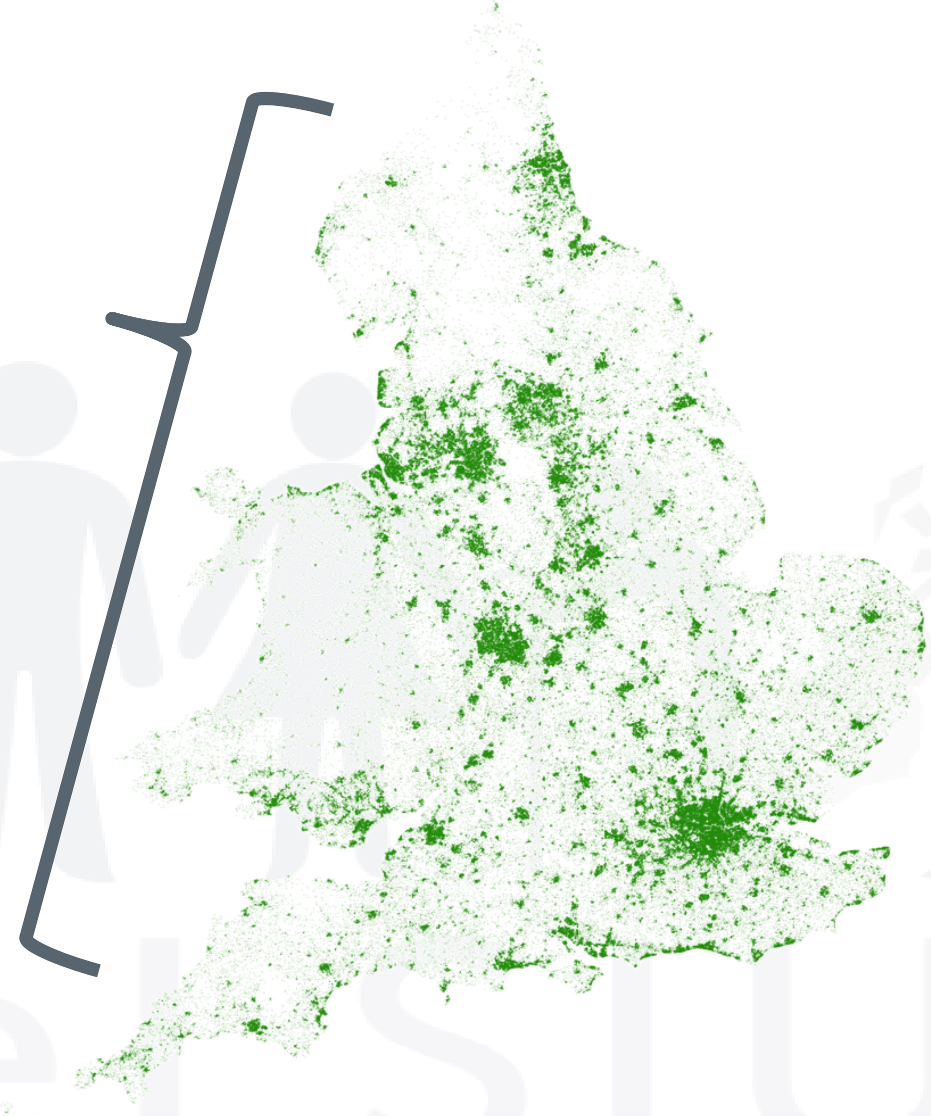
Random selection based on 4 birthdays

Longitudinal follow-up since 1971

All Census topics available

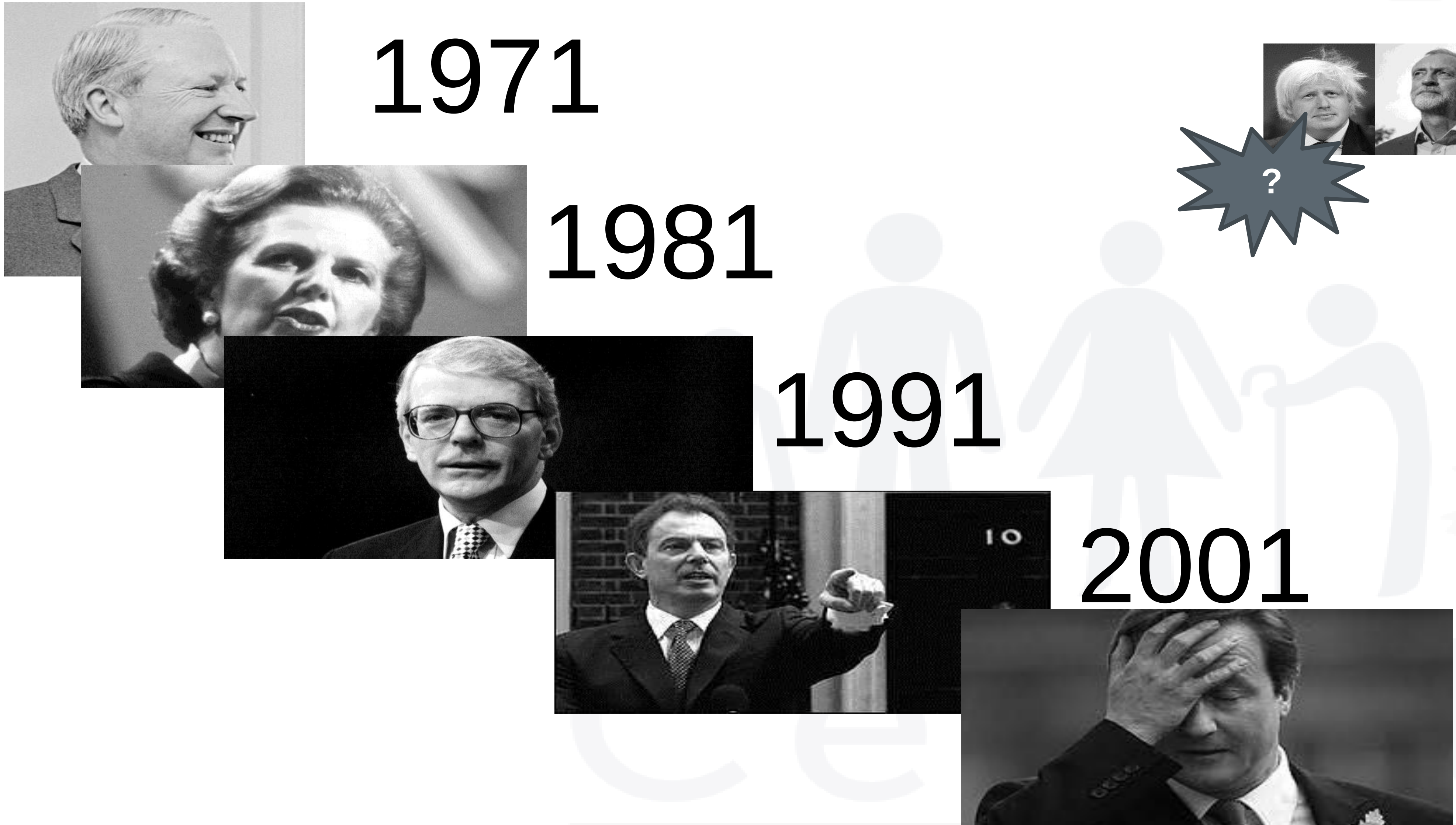
Life events data also linked to LS members, including births to sample mothers, deaths and cancer registrations

Large sample → subgroup analyses



http://www.theguardian.com/news/datablog/interactive/2013/aug/01/every-person-in-england-wales-dot-map#7/52.929/-2.571

Longitudinal data: sequence of censuses



1971

1981

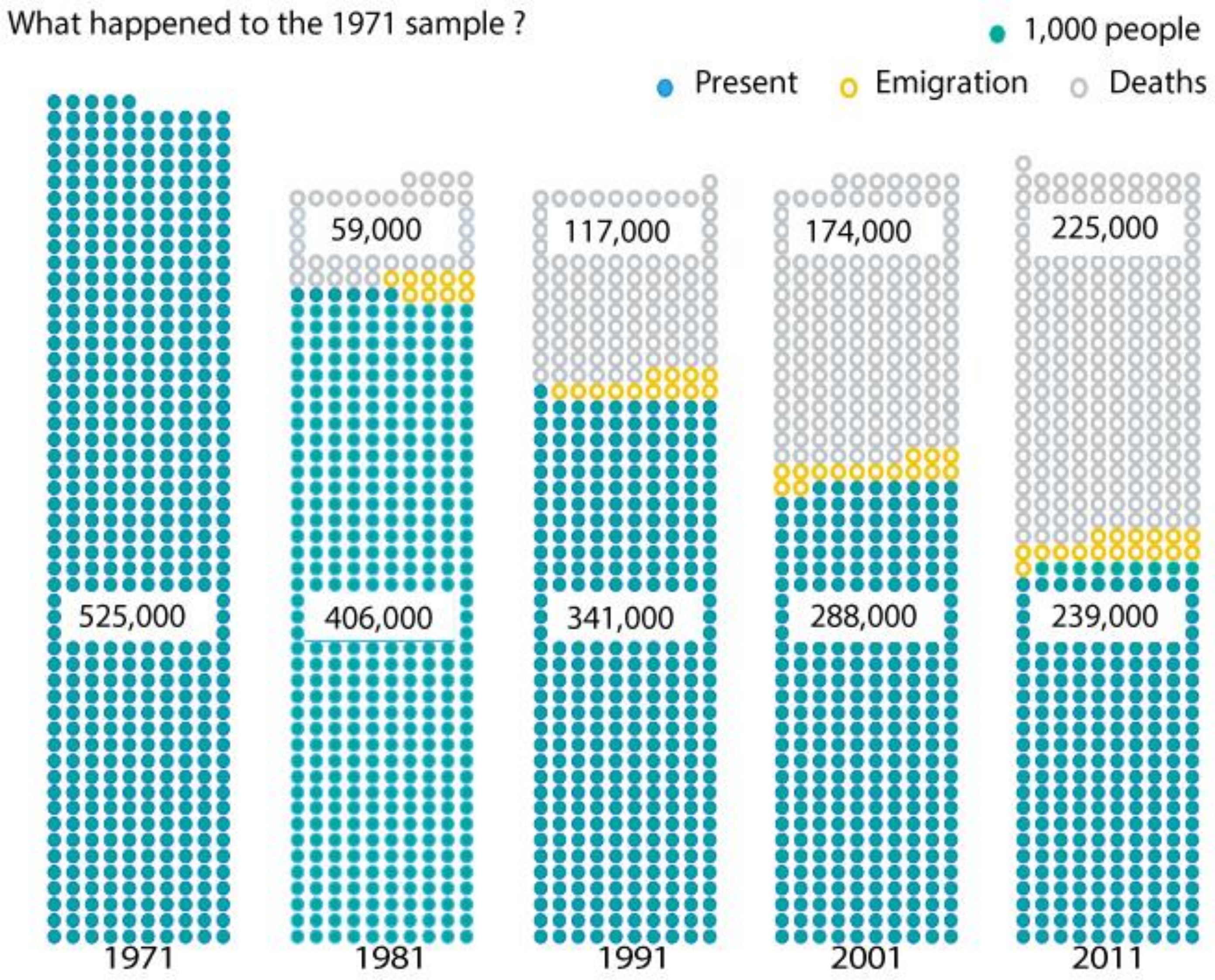
1991

2001

2011

2021

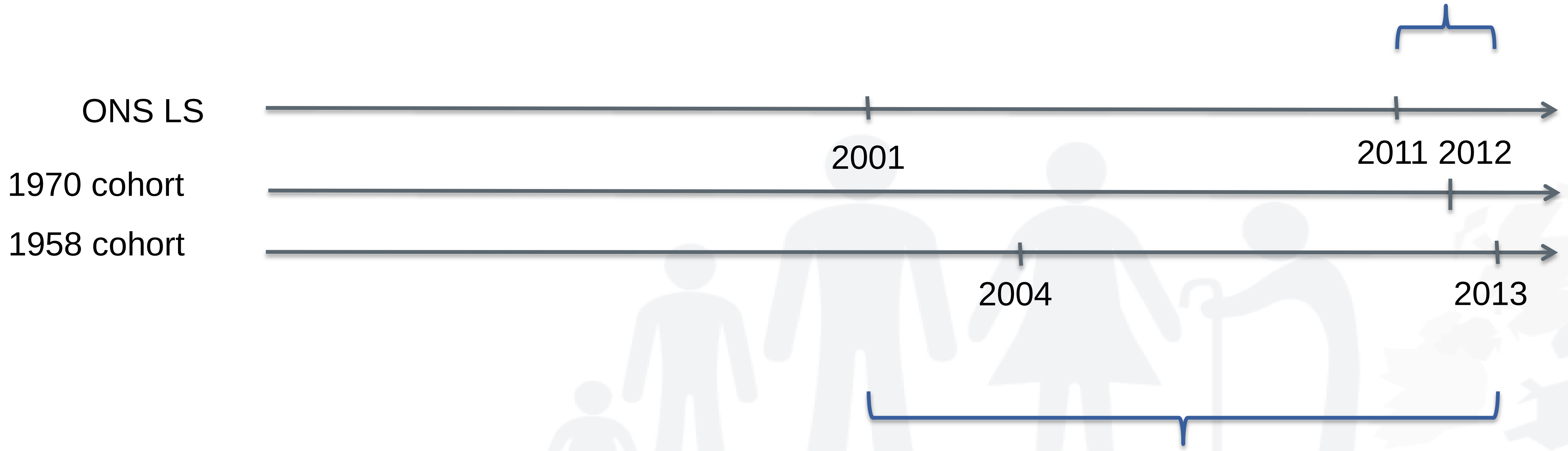
Low levels of attrition



Samples

- ‘Matched’
  - 1958 – Include only those resident in England and Wales
  - ONS LS – Include only those born in 1958 in England or Wales

1) Descriptive comparison: 2011 vs. 2012/2013



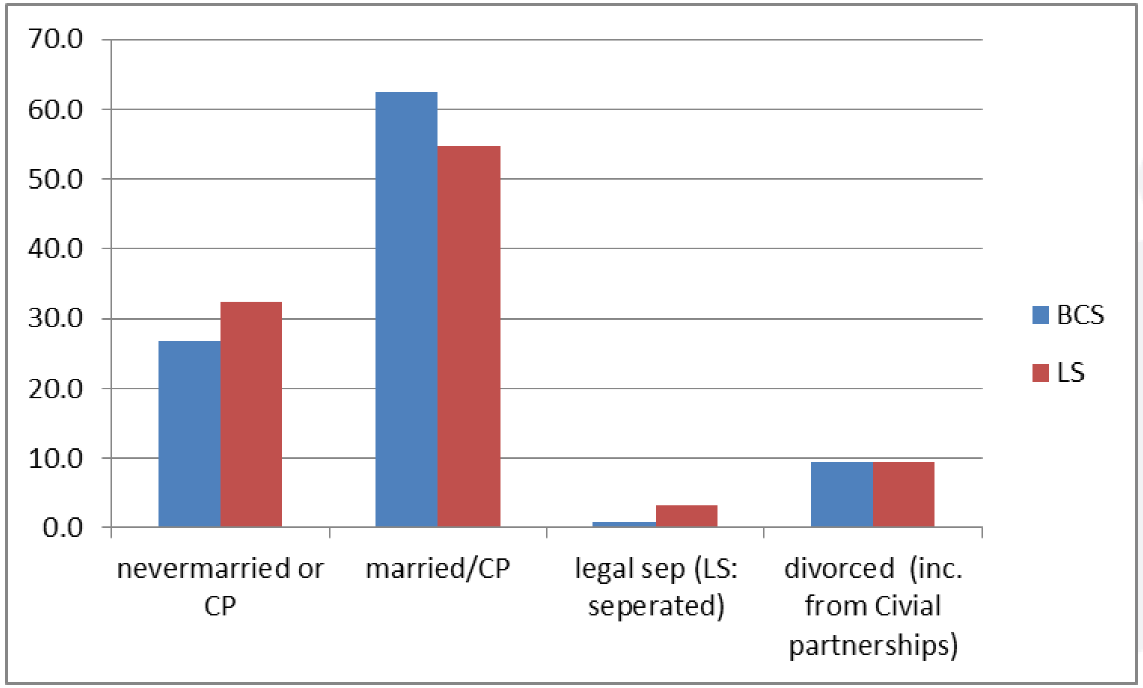
2) Longitudinal logistic regression

ONS LS: 2001-2011 (10 year)
1958 cohort: 2004-2013 (9 year)

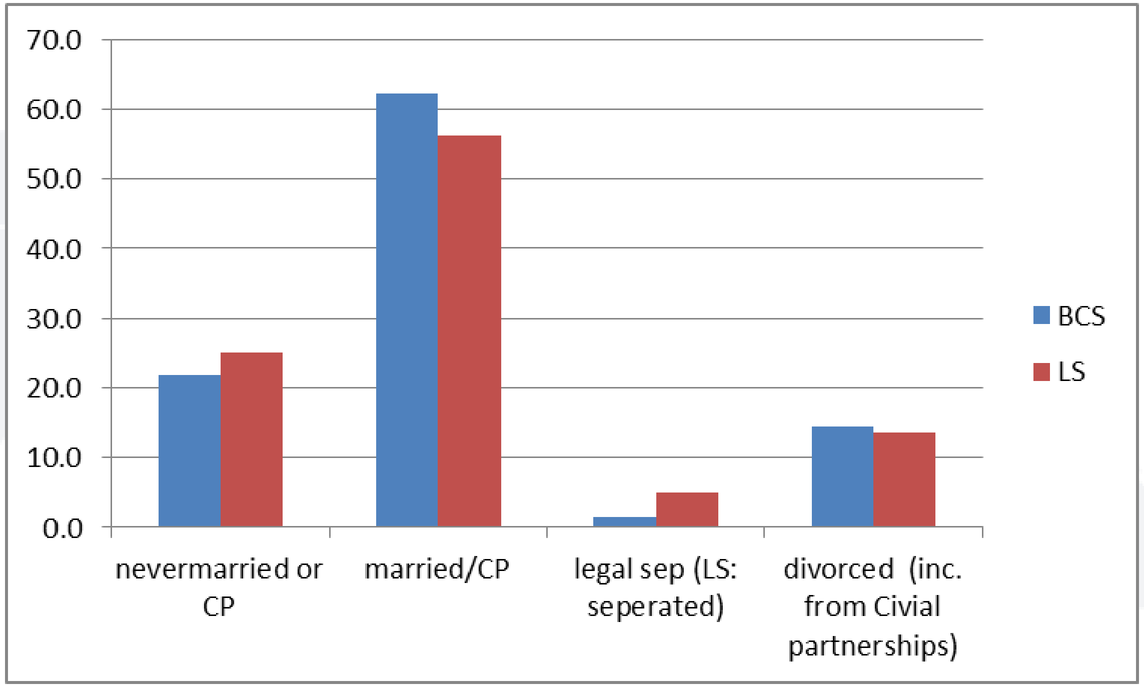
Comparability with BCS70

Marital status

Men

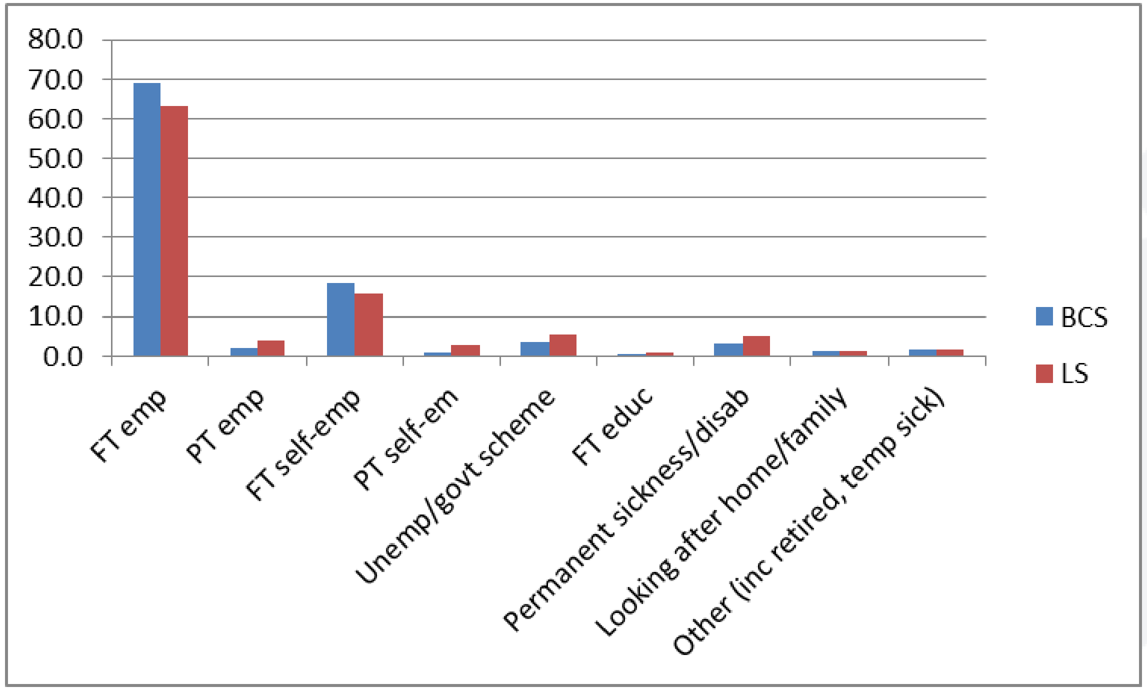


Women

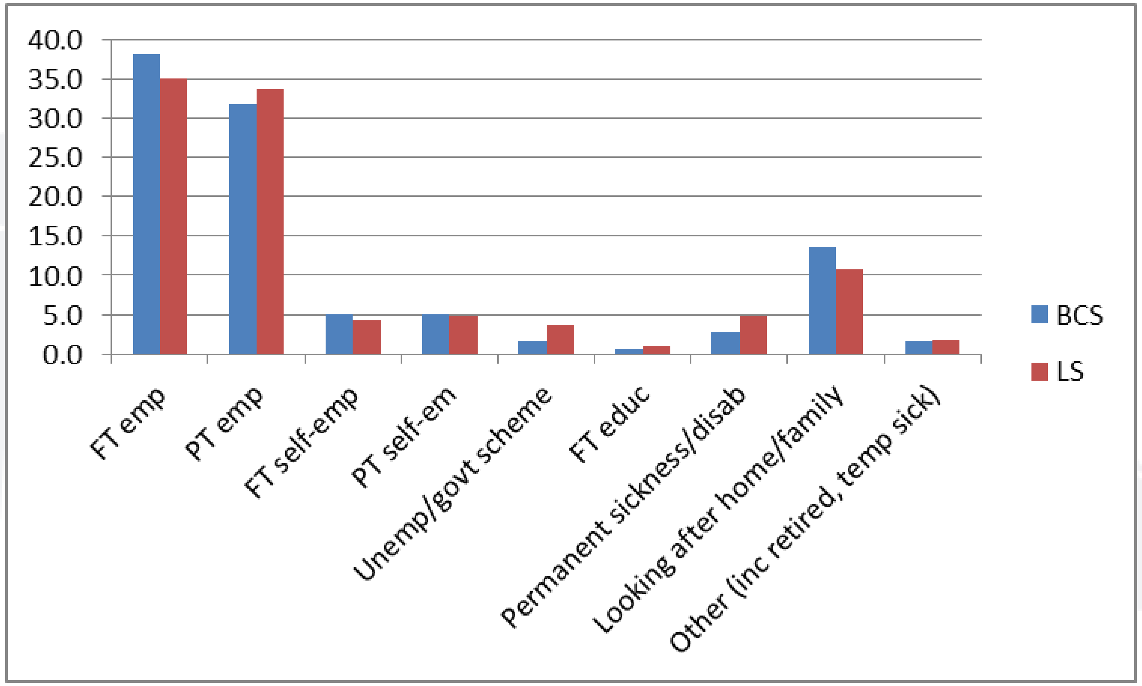


Employment status

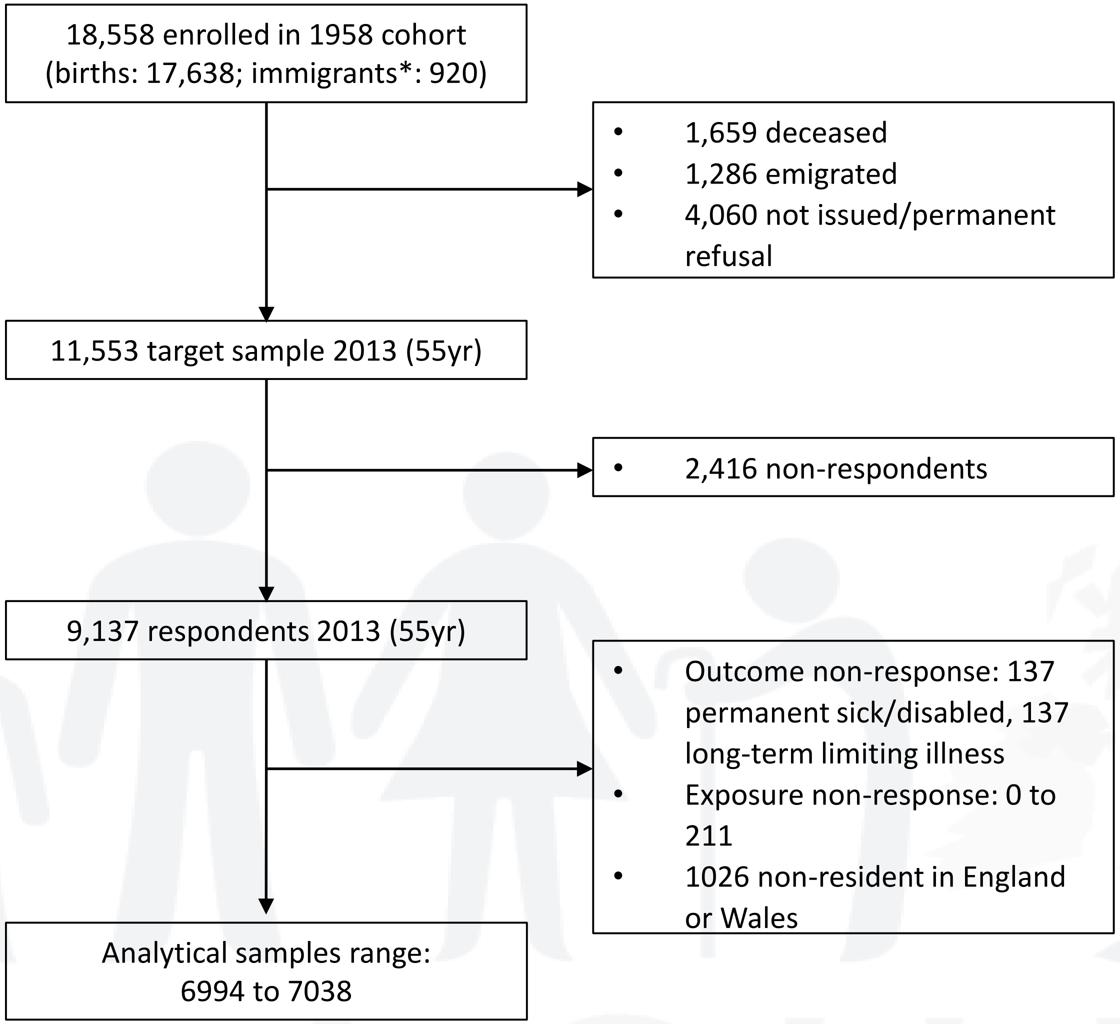
Men



Women



1958 cohort


18,558 enrolled in 1958 cohort
(births: 17,638; immigrants*: 920)
• 1,659 deceased
• 1,286 emigrated
• 4,060 not issued/permanent refusal
11,553 target sample 2013 (55yr)
• 2,416 non-respondents
9,137 respondents 2013 (55yr)
• Outcome non-response: 137 permanent sick/disabled, 137 long-term limiting illness
• Exposure non-response: 0 to 211
• 1026 non-resident in England or Wales
Analytical samples range:
6994 to 7038

Descriptive analyses

1958 cohort 2013 vs. ONS LS 2011

Variable		1958 cohort (2013) (n=8107)		ONS LS (2011) (n=7,085)	
		n	%	n	%
Sex	Male	3,931	48.5	3,545	50.0
	Female	4,176	51.5	3,540	50.0
Ethnicity	White	7,939	97.9	6,629	95.1
	Mixed	27	0.3	46	0.7
	Indian	32	0.4	93	1.3
	Pakistani and Bangladeshi	12	0.2	67	1.0
	Black or Black British	51	0.6	97	1.4
	Other ethnic group (inc. Chinese, all other)	46	0.6	37	0.5

Note: Totals may not sum due to missing cases with variable

Variable		1958 cohort (2013) (n=8107)		ONS LS (2011) (n=7,085)	
		n	%	n	%
Economic activity	Full time (30h or more)	4,890	**61.2**	4,198	**60.2**
	Part time (under 30h)	1,610	**20.2**	1,427	**20.5**
	Unemployed/seeking work	229	**2.9**	248	**3.6**
	Long-term sick/disabled	416	**5.2**	533	**7.6**
	Looking after home/family	496	**6.2**	278	**4.0**
	Other	346	**4.3**	288	**4.1**

Note: Totals may not sum due to missing cases with variable

Variable		1958 cohort (2013) (n=8107)		ONS LS (2011) (n=7,085)	
		n	%	n	%
Region	South: (London, SW, SE, EE)	3728	**46.0**	3,323	**46.9**
	North: (NW, NE, Y&H,EM, WM)	3,897	**48.1**	3,365	**47.5**
	Wales	482	**6.0**	397	**5.6**
Marital status	Married or civil partnership	5,796	**71.5**	4,785	**68.1**
	Divorced or former civil partner, separated, widowed	1,507	**18.6**	1,360	**19.4**
	Single and never married	799	**9.9**	882	**12.6**

Note: Totals may not sum due to rounding

Longitudinal Analyses

1958 cohort 2004-2013 vs. ONS LS 2001-2011

Exposure variables

- Region (South: London, SW, SE; North: NW,NE, Midlands, Yorkshire & Humberside; Wales)
- Marital status (married, divorced/separated/widowed, single)

Outcome variables

Long-term sick or disabled (yes/no)

1958 cohort

- **Q.** Which of the following best describes what you are currently doing?....   *'Sick or disabled'*
- **Q.** And are you temporarily sick or disabled or long-term sick or disabled?' *'Long-term sick or disabled (more than six months)'*

Census

- **Q.** Last week, were you:… *'long-term sick or disabled?'*

Outcome variables

Long-term limiting illness (yes/no)

1958 cohort

- **Q.** Do you have any physical or mental health conditions or illnesses lasting or expected to last 12 months or more? *'Yes; No'*
- **Q.** [If yes] Do any of your conditions or illnesses reduce your ability to carry out day-to-day activities? *'Yes, a lot; Yes, a little; Not at all'*

Census

- **Q.** Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months? Include problems related to old age. '*Yes limited a lot; Yes limited a little; No'*

Outcome Variables

Outcome	1958 cohort (2013) (n=8107)		ONS LS (2011) (n=7,085)	
	n	**%**	**n**	**%**
Long-term limiting illness (yes)	1575	**19.7**	1,317	**19.0**
Missing	115		156	
Long-term sickness or disability (yes)	416	**5.2**	533	**7.6**
Missing	137		113	

Longitudinal associations: Long-term illness

		NCDS 1958 (2004-2013)		ONS LS (2001-2011)	
Variable		**OR (95% CI)**	**p**	**OR (95% CI)**	**p**
Region	South	ref		ref	
	North	1.24 (1.10,1.40)	<0.001	1.40 (1.22, 1.60)	<0.001
	Wales	1.41 (1.10,1.81)	0.01	1.35 (1.01, 1.79)	0.04
Marital status	Married	ref		ref	
	Divorced, separated, widowed	1.39 (1.19,1.63)	<0.001	1.69 (1.44, 1.99)	<0.001
	Single	1.66 (1.39,1.98)	<0.001	1.93 (1.62, 2.30)	<0.001

Longitudinal associations: Long-term sick/disabled

		NCDS 1958 (2004-2013)		ONS LS (2001-2011)	
Variable		**OR (95% CI)**	**p**	**OR (95% CI)**	**p**
Region	South	ref		ref	
	North	1.64 (1.29,2.07)	<0.001	1.62 (1.31, 1.99)	<0.001
	Wales	2.60 (1.75,3.87)	<0.001	1.68 (1.11, 2.54)	0.01
Marital status	Married	ref		ref	
	Divorced, separated, widowed	2.01 (1.53,2.64)	<0.001	2.42 (1.91, 3.08)	<0.001
	Single	2.78 (2.10,3.70)	<0.001	3.69 (2.90, 4.69)	<0.001

Longitudinal associations: Long-term limiting illness (Inc. / Excl. Scotland)

		NCDS 1958 (2004-2013) Resident E & W		NCDS 1958 (2004-2013) All	
Variable		**OR (95% CI)**	**p**	**OR (95% CI)**	**p**
Region	South	ref		ref	
	North	1.24 (1.10,1.40)	<0.001	1.24 (1.10,1.40)	<0.001
	Wales	1.41 (1.10,1.81)	0.01	1.42 (1.11,1.83)	0.006
	Scotland	-	-	1.22 (0.99,1.49)	0.063
Marital status	Married	ref		ref	
	Divorced, separated, widowed	1.39 (1.19,1.63)	<0.001	1.34 (1.15,1.56)	<0.001
	Single	1.66 (1.39,1.98)	<0.001	1.55 (1.31,1.83)	<0.001

Summary / future work

- Decide whether to max sample or make as comparable as possible
- Consider which variables are most similar
- Develop longitudinal weights

Acknowledgment

- The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is supported by the ESRC Census of Population Programme under project ES/K000365/1. The authors alone are responsible for the interpretation of the data.
- This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates