ETHNICITY RECORDING IN PRIMARY CARE

Multiple imputation of missing data in ethnicity recording using The Health Improvement Network database

Tra Pham [1]

PhD Supervisors: Dr Irene Petersen [1], Prof James Carpenter [2], Dr Tim Morris [3]

[1]Department of Primary Care & Population Health - UCL
[2]Department of Medical Statistics - LSHTM
[3]Hub for Trials Methodology Research, MRC Clinical Trials Unit - UCL

Research funded by the Farr Institute of Health Informatics, London

OUTLINE

OUTLINE

- Background
  - Ethnicity recording in primary care
  - Aims of this project
- Study I - Impute missing ethnicity in THIN
  - Sample
  - Weighted multiple imputation
  - Comparison of missing data methods
- Study II - Simulation study
  - Motivation
  - Methods
  - Results
- Weighted multiple imputation by chained equations
  - Motivation
  - Syntax
- Summary

BACKGROUND

ETHNICITY RECORDING IN THE HEALTH IMPROVEMENT NETWORK

- THIN is one of UK's largest primary care databases
  - Contains records 12+ million patients in over 550 practices
  - Broadly representative of the UK population
  - Offers many opportunities for research
- Complete and reliable ethnicity recording is important for health research
- High level of missing ethnicity in primary care

AIMS OF THIS PROJECT

1. Explore the method of weighted multiple imputation to deal with missing ethnicity data
2. Introduce a new Stata command to implement weighted multiple imputation in multivariate missing data settings

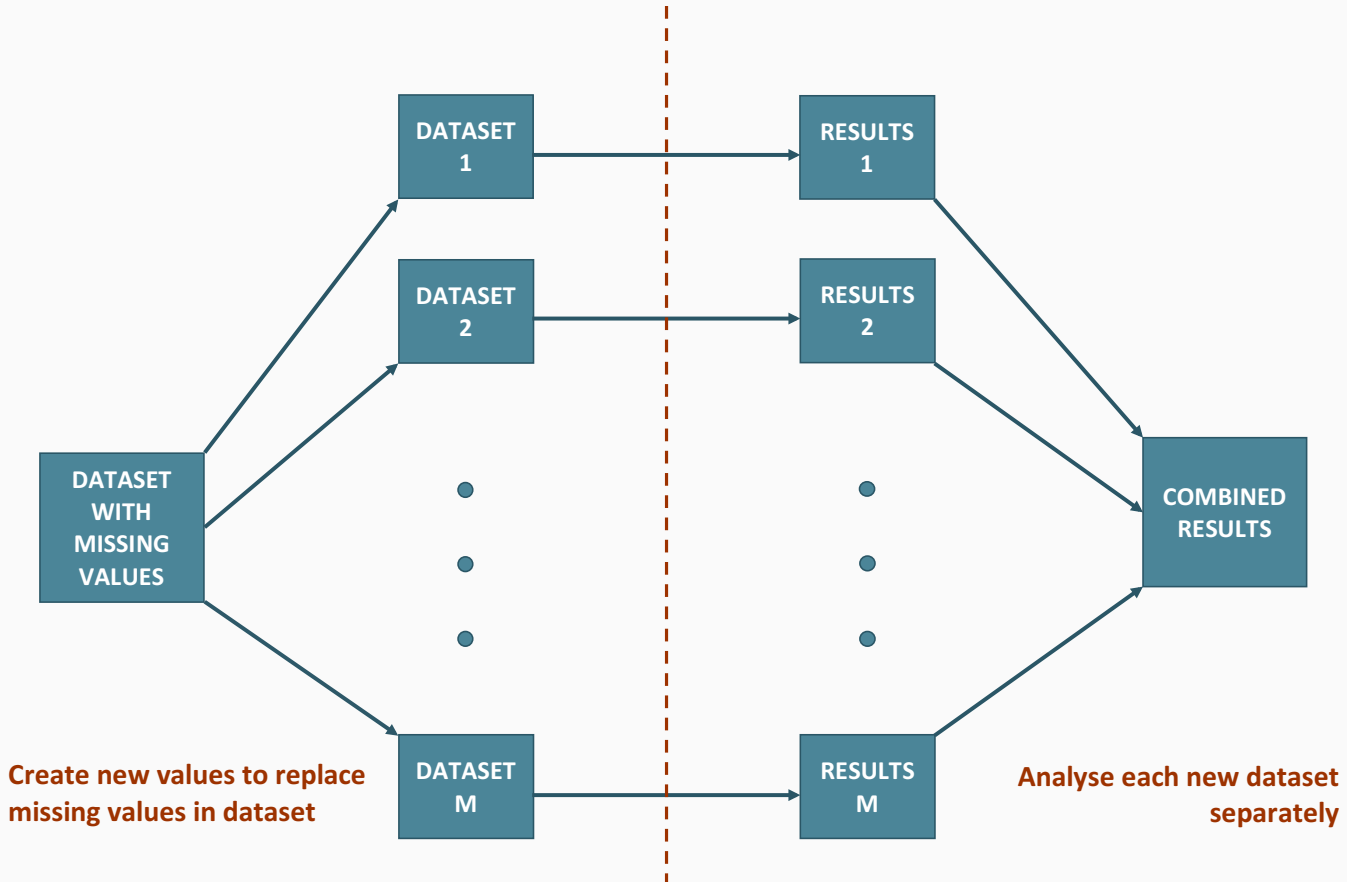
Study I - Impute missing ethnicity in THIN

SAMPLE

- 7 millions+ individuals in THIN between 2000-2013
- Ethnicity information is extracted using Read codes
- 5 categories of ethnicity: **white**, **mixed**, **Asian**, **black**, **other**
- 60% of individuals had missing ethnicity status

HOW SHOULD MISSING ETHNICITY DATA BE HANDLED?

- Studies often handle missing ethnicity data by
  - Excluding ethnicity in the analysis
  - Excluding individuals with missing ethnicity from the analysis
  - Imputing missing ethnicity as white
- A popular alternative is to use multiple imputation (MI) …



How should missing ethnicity data be handled?

... However, MI does not give plausible estimates compared with the census

$\longrightarrow$ Census data can be use to weight MI such that the correct ethnicity distribution is recovered

POST-STRATIFICATION (PS) WEIGHTS

- Deals with bias from non-response and under-represented groups in analysis of survey data

$$\text{Weight}_{ps} = 1/(p_{sample}/p_{census})$$

- Hypothetical example:

	Census Proportion	Sample Proportion	Weight_{ps}
White	0.6	0.8	0.75
Asian	0.2	0.1	2
Black	0.15	0.05	3
Other	0.05	0.05	1

- White ethnic group: $p_{weighted} = 0.8 \times 0.75 = 0.6 = p_{census}$

WEIGHTED MI OF ETHNICITY IN THIN USING PS WEIGHTS

Table 1: Ethnic proportions in one imputed dataset using PS weights

Ethnicity (%)	Observed	Imputed	Combined*	Census
White	67.1	63.2	65.7	59.8
Mixed	2.9	4.3	3.4	5.0
Asian	11.7	16.6	13.5	18.5
Black	11.7	13.2	12.2	13.3
Other	6.6	2.7	5.2	3.4
Total	100	100	100	100

* Combined data = observed data + imputed data

ADJUSTED WEIGHTS FOR MULTIPLE IMPUTATION

- Proportions required in the imputed data to get to the census proportions after imputation

Ethnicity	Observed	Imputed	Combined*
White	$p_{obs} \times N_{obs}$	$p_{req} \times N_{mis}$	$p_{census} \times N_{total}$
⋮	⋮	⋮	⋮
Total	N_{obs}	N_{mis}	N_{total}

* Combined data = observed data + imputed data

$$p_{req} = \frac{p_{census} \times N_{total} - p_{obs} \times N_{obs}}{N_{mis}}$$

$$\rightarrow \text{Weight}_{MI} = 1/(p_{obs}/p_{req})$$

WEIGHTED MI OF ETHNICITY IN THIN USING ADJUSTED WEIGHTS

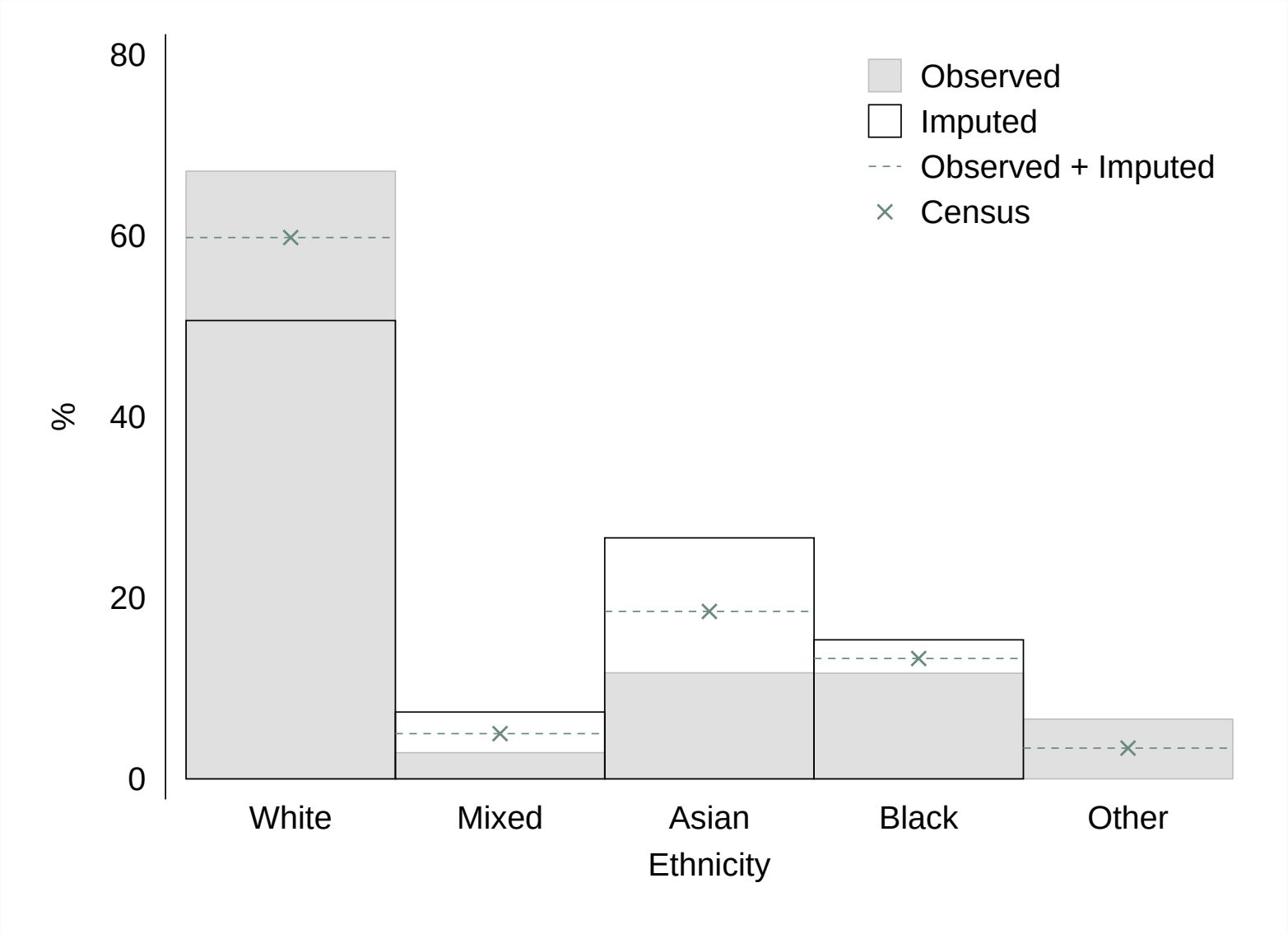
Table 2: Ethnic proportions in one imputed dataset using adjusted weights

Ethnicity	Observed	Imputed	Combined*	Census
White	67.1	50.6	61.0	59.8
Mixed	2.9	7.4	4.6	5.0
Asian	11.7	26.6	17.3	18.5
Black	11.7	15.4	13.0	13.3
Other	6.6	0	4.1	3.4
Total	100	100	100	100

* Combined data = observed data + imputed data

WEIGHTED MI OF ETHNICITY IN THIN USING ADJUSTED WEIGHTS

Figure 1: Ethnic proportions in one imputed dataset using adjusted weights



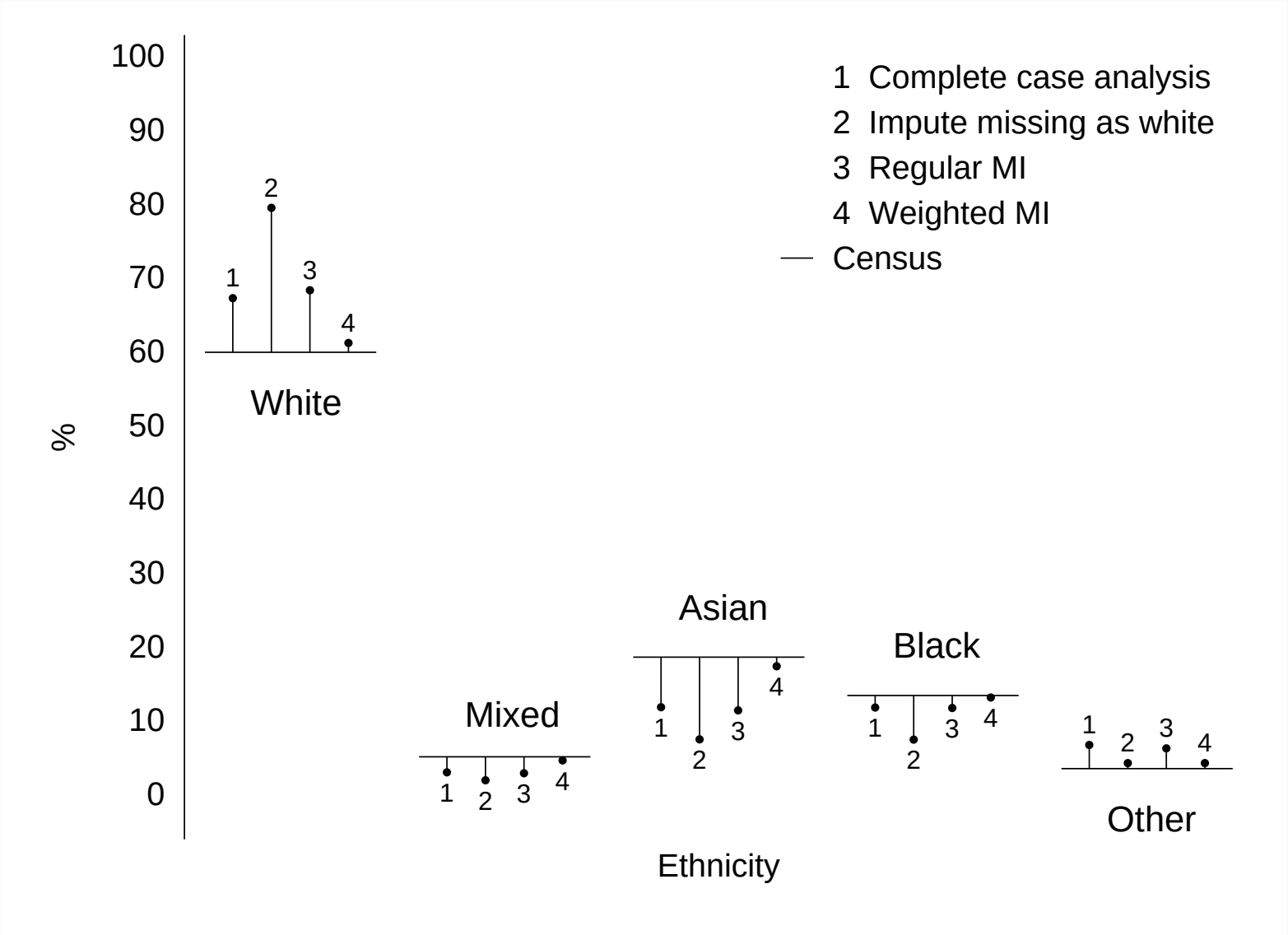
WEIGHTED MULTIPLE IMPUTATION IN STATA

- Predictors for imputation: year of birth, sex, deprivation score, indicator if first registration is between 2006-2013
- `mi impute mlogit` command allows for probability weights specification:

```
mi impute mlogit varlist [pweights] [, mi_options …]
```

COMPARISON OF MISSING DATA METHODS

Figure 2: Ethnic proportions under different missing data methods

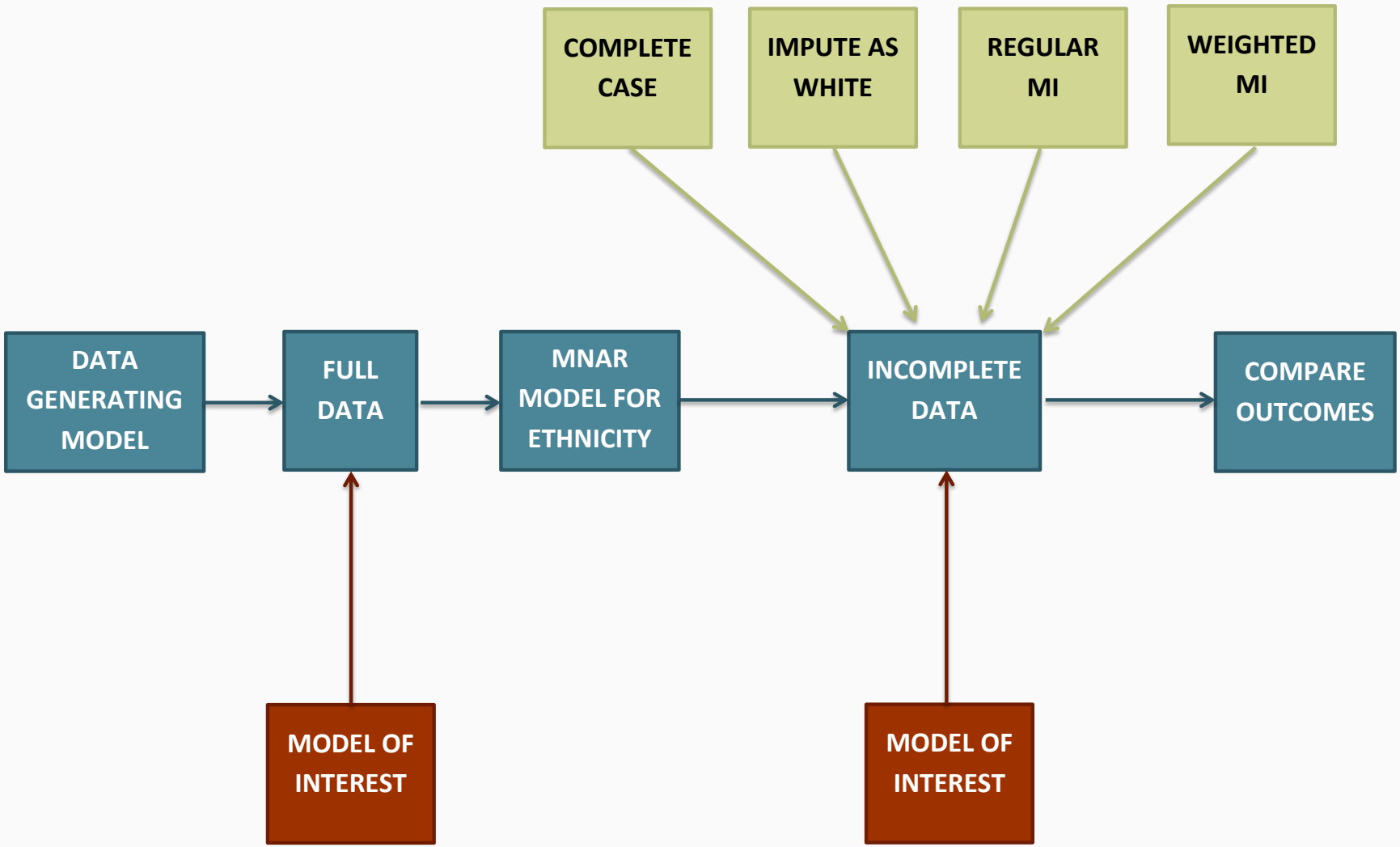


STUDY II - SIMULATION STUDY

MOTIVATION

- So far the weighted MI gives plausible estimates of ethnicity proportions
- However, studies often focus on the association between ethnicity and outcome of interest
- How to deal with missing data when ethnicity is an exposure/covariate?
- Aim: evaluate the performance of weighted MI using a simulation study

DESIGN OF SIMULATION STUDY



MODEL OF INTEREST

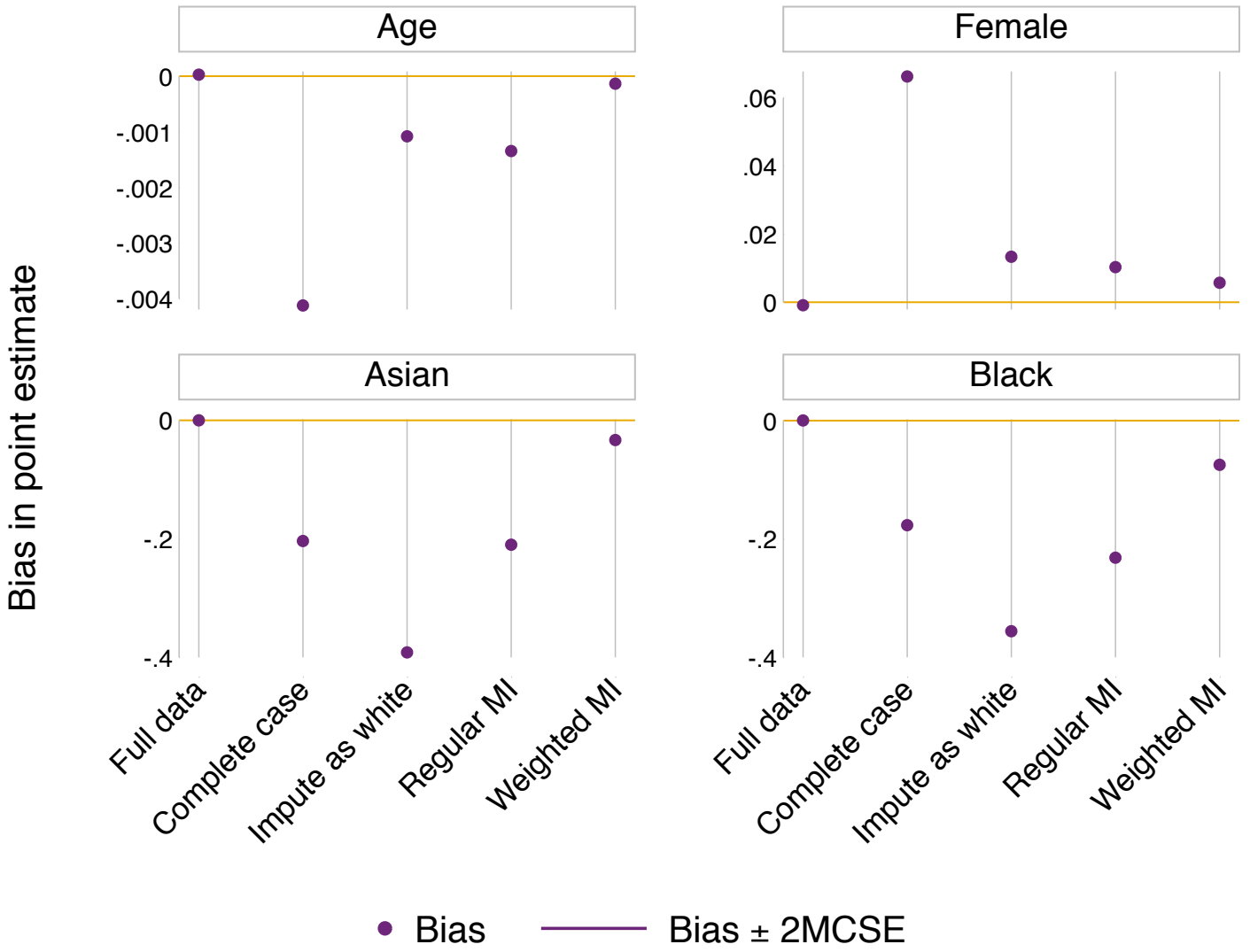
- Motivated by the association between ethnicity and myocardial infarction
- Statistical model: exponential survival model for the association between ethnicity and time to first diagnosis of myocardial infarction, adjusted for age and sex

DATA GENERATING AND MNAR MECHANISMS

1. Data generating mechanism:
   - Data are simulated that closely model a sample derived from THIN
   - Ethnicity is generated as a 3-category variable: **white**, **Asian**, **black**, using proportions from the census
2. MNAR mechanism:
   - Ethnicity is made MNAR depending on follow-up time, sex and ethnicity

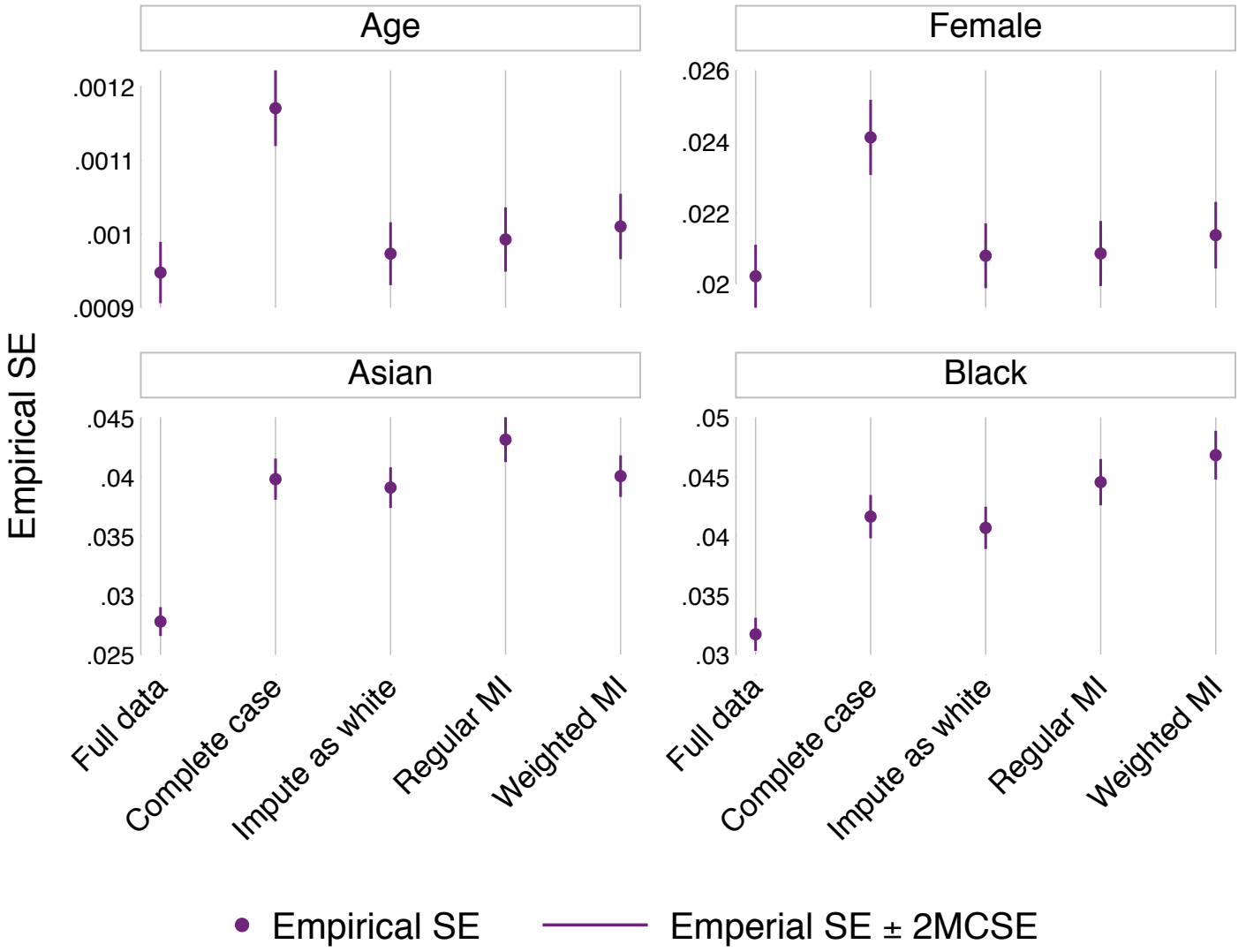
RESULTS - BIAS

- Define $\bar{\beta} = \frac{1}{K}\sum_{k}\hat{\beta}_k \longrightarrow$ Estimated bias $= \bar{\beta} - \beta$



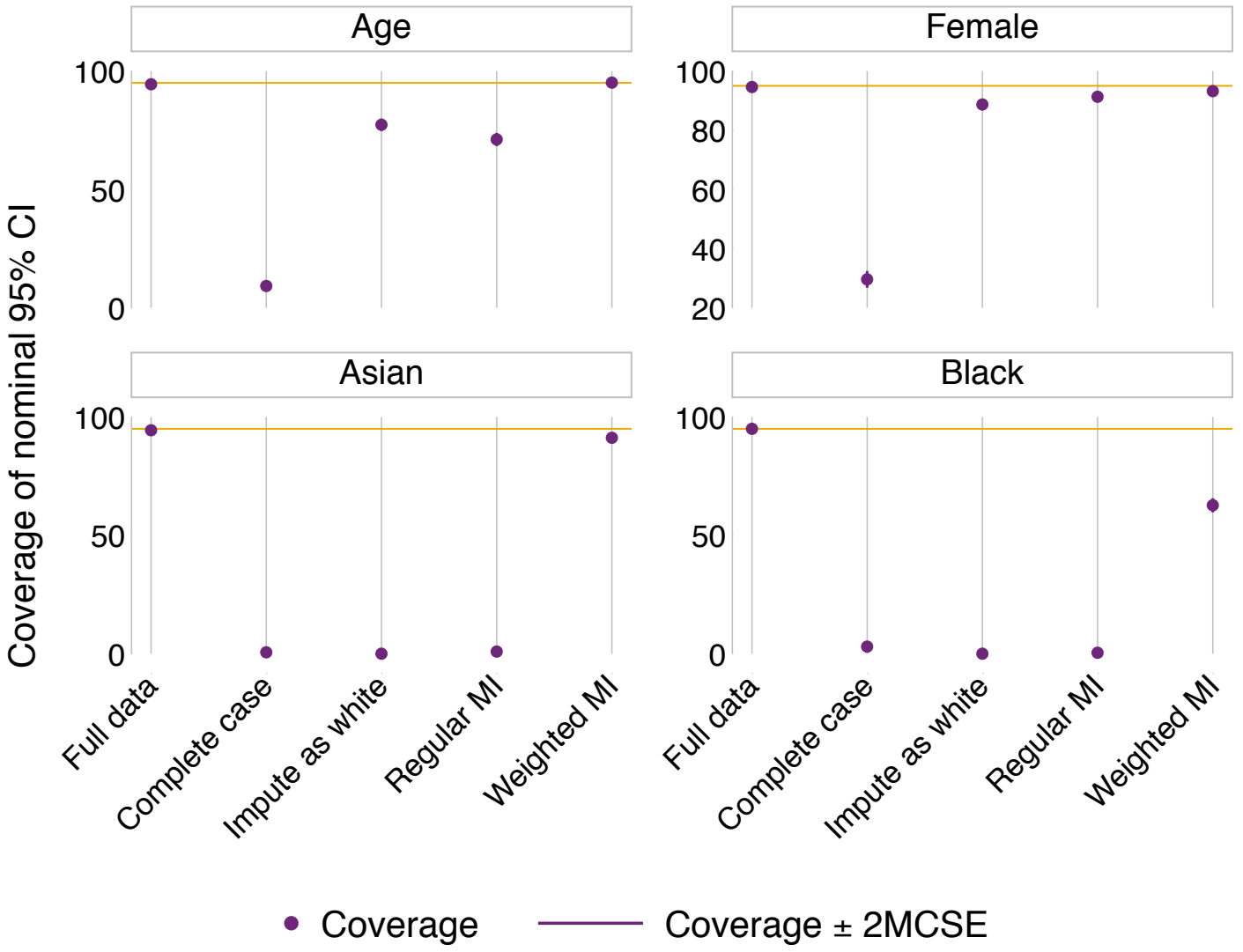
RESULTS - EMPIRICAL STANDARD ERRORS

- Empirical standard error $= \sqrt{\frac{1}{K-1}\sum_k(\hat{\beta}_k - \bar{\beta})^2}$



RESULTS - COVERAGE

- Coverage $= \frac{1}{K}\sum_{k} 1(|\ \hat{\beta}_k - \beta\ | < z_{\alpha/2} s_k)$



SUMMARY OF RESULTS

1. Complete case analysis leads to biased results and loss in efficiency
2. Impute missing data as white biases the association between non-white groups and outcome
3. Regular MI does not correct for bias when ethnicity is MNAR
4. Weighted MI gives closest estimates to true values and substantial coverages

WEIGHTED MI BY CHAINED EQUATIONS

- Weighted MI only implemented in univariate missing data of ethnicity so far
- Multiple imputation of chained equations (MICE) for multivariate missing data problems
  - To get one set of imputed values, iterate over $t = 0, 1, \ldots, T$ and impute:

$$
\begin{aligned}
&Y_1^{t+1} \text{ using } Y_2^t,\ Y_3^1, \ldots,\ Y_k^t \\
&Y_2^{t+1} \text{ using } Y_1^{t+1},\ Y_3^t, \ldots,\ Y_k^t \\
&\vdots \\
&Y_k^{t+1} \text{ using } Y_1^{t+1},\ Y_2^{t+1},\ \ldots, Y_{k-1}^{t+1}
\end{aligned}
$$

- `mi impute chained` only allows specification of global weights, which are applied to all equations

WEIGHTED MULTIPLE IMPUTATION BY CHAINED EQUATIONS IN STATA

- Weighted imputation by chained equations - `wice`
- Allows weights specification for one conditional models while no weights for others:

$$
\begin{aligned}
&Y_1^{t+1} \text{ using } Y_2^t,\ Y_3^1, \ldots,\ Y_k^t \longleftarrow pw_1 \\
&Y_2^{t+1} \text{ using } Y_1^{t+1},\ Y_3^t, \ldots,\ Y_k^t \\
&\vdots \\
&Y_k^{t+1} \text{ using } Y_1^{t+1},\ Y_2^{t+1},\ \ldots, Y_{k-1}^{t+1}
\end{aligned}
$$

WEIGHTED MULTIPLE IMPUTATION BY CHAINED EQUATIONS IN STATA

- Stata syntax:

`wice (method`$_1$ `impvar`$_1$ `[pweight`$_1$`]) ... (method`$_K$ `impvar`$_K$ `[pweight`$_K$`]) : indvars [, options]`

- Current features:
  1. Univariate imputation methods: `regress`, `logit`, `ologit`, `mlogit`
  2. Global options: `add`, `cycles`, `seed`, `noisily`, `dryrun`

CONCLUSIONS

- Ad-hoc methods can lead to biased results and misleading conclusions
- Regular MI fails to produce valid inferences when data are MNAR
- Weighted MI is a simple solution and performs better than regular MI
- `wice` offers flexibility for weighted MI by chained equations