Level 1 guidance on conducting and reporting sensitivity analyses for missing data

Katherine Lee, Rheanna Mainzer, Kate Tilling and James Carpenter

on behalf of STRATOS TG1: Missing Data

ISCB 2023

Acknowledgements

- **Rheanna Mainzer**
- **James Carpenter**
- **Kate Tilling**
- Rosie Cornish
- Rod Little
- Melanie Bell
- Els Goetghebeur
- Joe Hogan

Outline

- The TARMOS Framework
- Aim of this work
- The case study
- Sensitivity analyses
  - Step 1: Planning
  - Step 2: Conducting
  - Step 3: Reporting
- Analysis of the case study
- Discussion

The TARMOS Framework

- Missing data are common in medical research
- Guidance is available, but missing data are still often not handled appropriately
- Particularly problematic in observational research
- Proposed a practical framework for the Treatment And Reporting of Missing data in Observational Studies (TARMOS)
- Focus on multiple imputation (MI) because of its flexibility and practicality

The TARMOS Framework

1. Plan the analysis

a) What is the analysis model if no missing data?

b) How are missing data going to be handled?

- Is a complete records analysis likely to be valid?
- Is MI likely to offer benefits over a complete records analysis?
- Is a sensitivity analysis required?

2. Conduct the analysis

a) Examine the data – consistent with analysis plan?

b) Conduct the analysis as per the plan – justifying any amendments

3. Report the analysis

a) Describe missing data

b) Describe and justify how missing data were handled

c) Report all analyses

The TARMOS Framework

1. Plan the analysis

a) What is the analysis model if no missing data?

b) How are missing data going to be handled?

- Is a complete records analysis likely to be valid?
- Is MI likely to offer benefits over a complete records analysis?
- Is a sensitivity analysis required?

2. Conduct the analysis

a) Examine the data – consistent with analysis plan?

b) Conduct the analysis as per the plan – justifying any amendments

3. Report the analysis

a) Describe missing data

b) Describe and justify how missing data were handled

c) Report all analyses

Sensitivity analysis

- In some scenarios, it may not be possible to estimate an estimand of interest consistently using the observed data alone, i.e. is not "recoverable"
- For example, if missingness in a variable depends on the variable itself
  - E.g. smokers may be less likely than non-smokers to answer a survey about smoking habits
- Requires external information about the missing values
  - "Missing not at random analysis"
  - "Delta-adjusted analysis"
  - "Bias analysis"
- External information may be used to inform the missing values, or may be expressed in the form of a sensitivity parameter
- Important step that is often overlooked and poorly reported

Aim

Provide accessible practical guidance on the planning, conduct and reporting of sensitivity analyses which incorporate external information about the missing values

- ➢ Start from the point where it has been decided that a sensitivity analysis is required…
- ➢ Accessible for level 1 researcher
- ➢ Include code for different approaches
- ➢ Provide example text for reporting

CAVEAT: This is a work in progress….

Case Study: ALSPAC

- The **A**von **L**ongitudinal **S**tudy of **P**arents and **C**hildren
  - Transgenerational prospective observational study
  - 14,541 women recruited initially (14,062 live births) with additional children enrolled subsequently

- Three estimands of interest:
  1. Proportion who are current at age 14 years (obtained via a computerised questionnaire during a clinic assessment and a postal questionnaire)
  2. Mean educational attainment at age 16 years (obtained via linkage to the National Pupil Database)
  3. Causal relationship between smoking at 14 years and educational attainment at 16 years

Step 1: Planning the sensitivity analysis

Should be pre-specified….

1. Start with a plausible m-DAG

2. What variables are we going to consider a sensitivity analysis for?

3. What analytic method are we going to use?

4. How are we going to chose the sensitivity parameter?

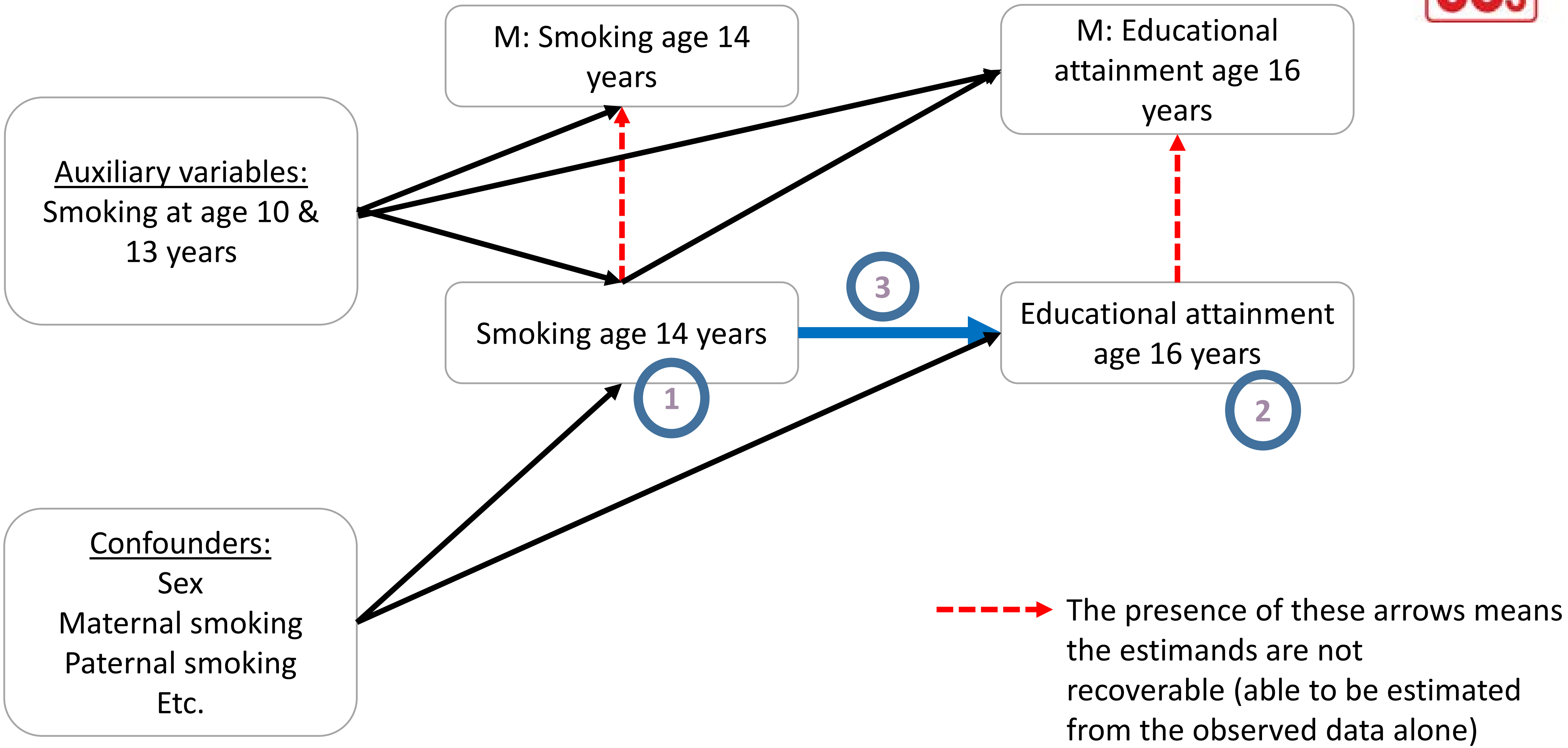
Step 1: Planning the sensitivity analysis

Should be pre-specified….

1. Start with a plausible m-DAG

Step 1: Planning the sensitivity analysis

Step 1: Planning the sensitivity analysis



Step 1: Planning the sensitivity analysis

Should be pre-specified….

1. Start with a plausible m-DAG

2. What variables are we going to consider a sensitivity analysis for?
    - Often lots of incomplete variables
    - Best to focus on a single key variable

→ Estimand 1: smoking age 14 years
Estimand 2: educational attainment age 16 years
Estimand 3: educational attainment age 16 years

Step 1: Planning the sensitivity analysis

Should be pre-specified….

1. Start with a plausible m-DAG

2. What variables are we going to consider a sensitivity analysis for?

3. What analytic method are we going to use?

Step 1: Planning the sensitivity analysis

Best case/worst case: missing values replaced with extreme values representing the best and worst case scenario

E.g. (1) missing data on smoking at age 14 years = non-smoker and (2) missing data on smoking at age 14 years = smoker

- Simple to understand and conduct
- Provides bounds on the estimate that we would have obtained if we had complete data

- Choice of extremes arbitrary for continuous variables
- Estimates unlikely to be the estimate that we would have observed had all data been complete

Step 1: Planning the sensitivity analysis

Start with a joint model for the data and its missingness

e.g. single (binary) outcome (Y) and missingness indicator (R)

$$P(Y, R)$$

Pattern mixture models

Factorise into:

$$P(R)P(Y|R)$$

i.e. a marginal model for missingness, and a model for the outcome conditional on missingness status

Sensitivity parameter = difference in log odds of the outcome between those with observed vs missing data

Selection models

Factorise into:

$$P(Y)P(R|\mathrm{Y})$$

i.e. a marginal model for the outcome, and a model for missingness given the outcome

Sensitivity parameter = difference in log-odds of missingness between those with and without the outcome

Step 1: Planning the sensitivity analysis

Pattern mixture models $P(R)P(Y|R)$

- Can be implemented using multiple imputation – delta-adjusted MI
  - Fit the imputation model using the complete cases
  - Modify imputed values to reflect expected differences between observed and missing values (delta)
  - Fit analysis to each (modified) imputed dataset and combine using Rubin's rules
- Available via NARFCS (not at random fully conditional specification)

Straight-forward to communicate

Sensitivity parameter straightforward to understand

Can be conducted using R/Stata

Is used in practice

Need to fit the analysis model as a separate step

Step 1: Planning the sensitivity analysis

Selection models $P(Y)P(R|Y)$

- Can also be fitted using MI using a stacking approach
  - Conduct standard MI and stack the imputed datasets
  - Each observation assigned a weight proportional to the odds of Y being observed conditional on the imputed value and other variables in the dataset (Y missing), or 1/M (Y observed)
  - Analysis conducted using a weighted version of the target analysis
- Available via R package StackImpute.

👍 Directly fits the analysis model P(Y)

👎 P(R|Y) not very intuitive to understand

👎 Sensitivity parameter hard to interpret

👎 Standard errors not straightforward to calculate: a jack-knife approach proposed

👎 Not commonly used in practice

Step 1: Planning the sensitivity analysis

Other approaches

- Reference based imputation
- Trimmed mean
- Shared parameter model
- Inverse probability weighting
- Full Bayes
- …

Step 1: Planning the sensitivity analysis

As part of the plan, should also specify:

- Assumption being made about the missing data in the primary analysis
  - Ideally would be the most realistic assumption
  - In practice would typically be no arrow

- Assumption being made about the missing data in secondary analyses

- What we will do if we encounter difficulties e.g. with model convergence when using MI

Step 1: Planning the sensitivity analysis

Should be pre-specified….

1. Start with a plausible m-DAG

2. What variables are we going to consider a sensitivity analysis for?

3. What analytic method are we going to use?

4. How are we going to chose the sensitivity parameter?

Step 1: Planning the sensitivity analysis

Choosing the sensitivity parameter

- Elicitation – ask/survey content experts
- Literature – review of the literature
- Tipping point analysis – consider a range of values to assess whether there is a point at which qualitative conclusions changes
  - focuses on binary conclusions about rejecting or accepting null hypotheses

Step 2: Conduct the planned analysis

As per TARMOS:

- Check the assumptions made in the analysis plan are acceptable
- Follow the pre-specified analysis plan
- If the analysis plan needs to be revised, any changes should be acknowledged and justified

Step 3: Report the results

- Describe & justify **assumptions** made about the missing data, e.g. via an m-DAG, and arrows of interest
- Describe **method** for conducting the sensitivity analysis, and why (reproducibility)
  - State the values used for the sensitivity parameter(s) and how they were chosen
  - Which computer package and tuning parameters
- Present results of the **primary analysis**
- Report results for each of the **alternative analyses**/values of the sensitivity parameter
  - Figures can be helpful to summarise multiple estimates & highlight trends
- **Interpret** the results
  - How/if main conclusions change for different values of the sensitivity parameter
  - What the most likely value/result is

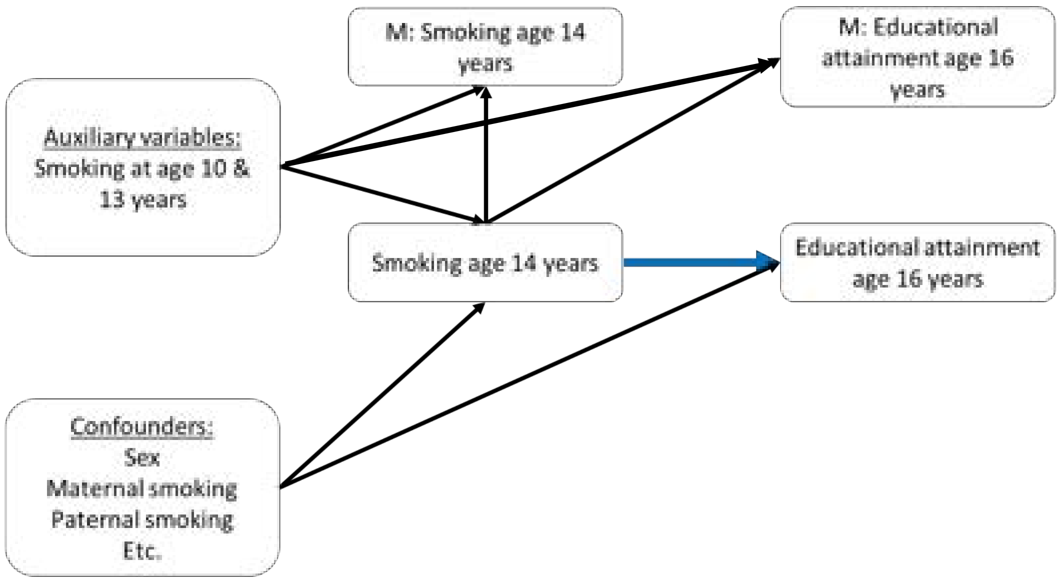
[Some of this may be included in the supplementary material]

Paper will include example text based on the case study….

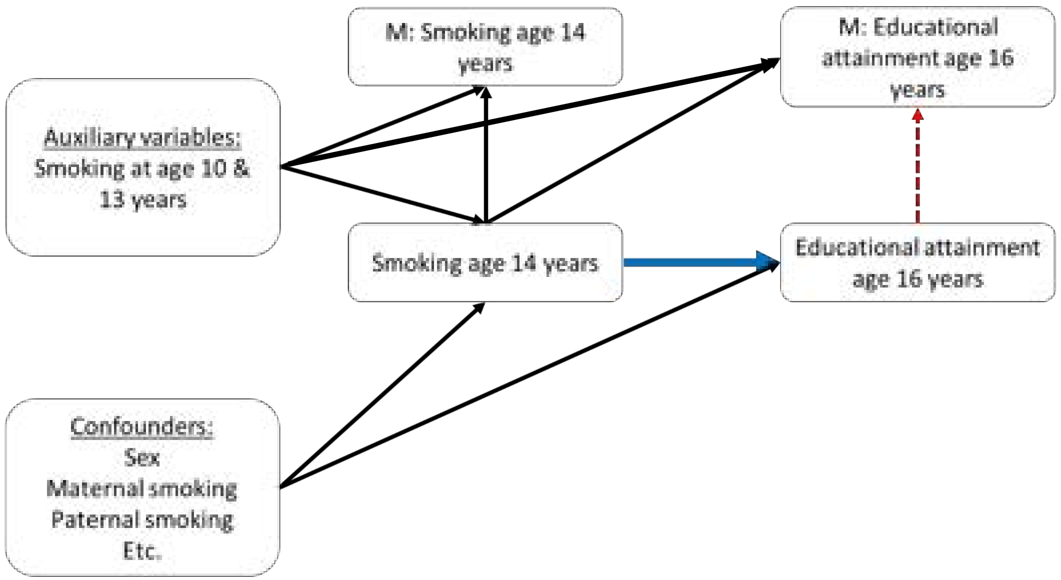
Case Study: ALSPAC Estimand 3

Step 1

Primary



Sensitivity



Case Study: ALSPAC Estimand 3

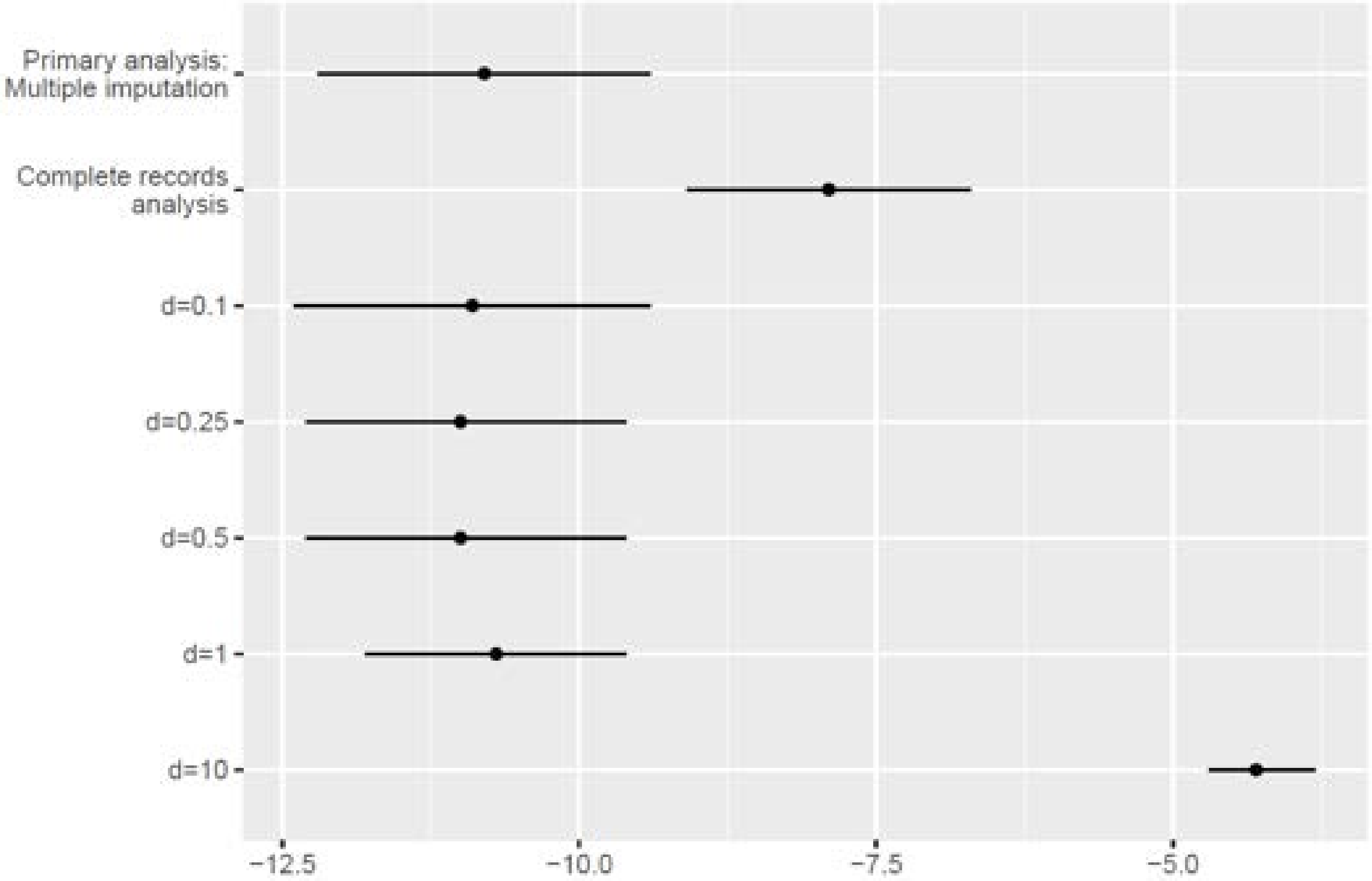
Step 1 (& 2)

- Primary analysis: MI
- Sensitivity analysis conducted using pattern mixture approach fitted using delta-adjusted MI in Stata
  - Intuitive
  - Straight-forward to fit
- Values of sensitivity parameter (d) from discussion with content experts
  - Plausible values: 0.1, 0.25, 0.5 and 1.
  - Extreme value: 10
- Complete records analysis for comparison

d = difference in log odds of the outcome between those with observed vs missing data

Case study: ALSPAC Estimand 3

Step 3



Mean difference in educational attainment in smokers compared with non-smokers with 95% confidence interval

All analysis suggest a causal relationship between smoking age 14 years and educational attainment age 16 years

Discussion

- Hope this tutorial will provide much needed guidance to make this form of sensitivity analyses more accessible
  - Increase uptake
- Strong focus on pre-planning and transparent reporting (with example code & text)
  - Encourage the reliability and reproducibility of research

Reference

- Beesley LJ and Taylor JMG. Accounting for not-at-random missingness through imputation stacking . Stats in Med, 2021. 40(27): 6118-6132. https://doi.org/10.1002/sim.9174
- Cro S, Morris TP, Kenward MG and Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. Stats in Med, 2020. 39 (21); 2815-2842.
- Héraud-Bousquet, V., Larsen, C., Carpenter, J. et al. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. BMC Med Res Methodol 12, 73 (2012). https://doi.org/10.1186/1471-2288-12-73
- Lee, K. J, Tilling, K, Cornish, R. P, Little, R. J. A, Bell, M. L, Goetghebeur, E., Hogan J.W. and Carpenter, J. R., on behalf of the STRATOS initiative (2020). Framework for the treatment and reporting of missing data in observational studies: the TARMOS framework. JCE 2021. 134: P79-88. 2021 Jun;134:79-88. doi: 10.1016/j.jclinepi.2021.01.008.
- Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Stat Med. 2018 Jul 10;37(15):2338-2353. doi: 10.1002/sim.7643. Epub 2018 Apr 2. PMID: 29611205; PMCID: PMC6001532.