Framework for the Treatment And Reporting of Missing data in Observational Studies: The TARMOS framework

James Carpenter and Katherine Lee

on behalf of STRATOS TG1: Missing Data

ISCB 2020

Acknowledgements

- Kate Tilling
- Rosie Cornish
- Rod Little
- Melanie Bell
- Els Goetghebeur
- Joe Hogan

JC is funded by the UK MRC; KL by the Australian NHMRC; KT and RC by the UK MRC, Wellcome Trust and University of Bristol.

Outline

- Background – why do we need a framework?
- Case study – exploring a causal effect of teen smoking on educational achievement
- The **T**reatment **A**nd **R**eporting of **M**issing data in **O**bservational **S**tudies (TARMOS) framework
- Application
- Discussion

Background

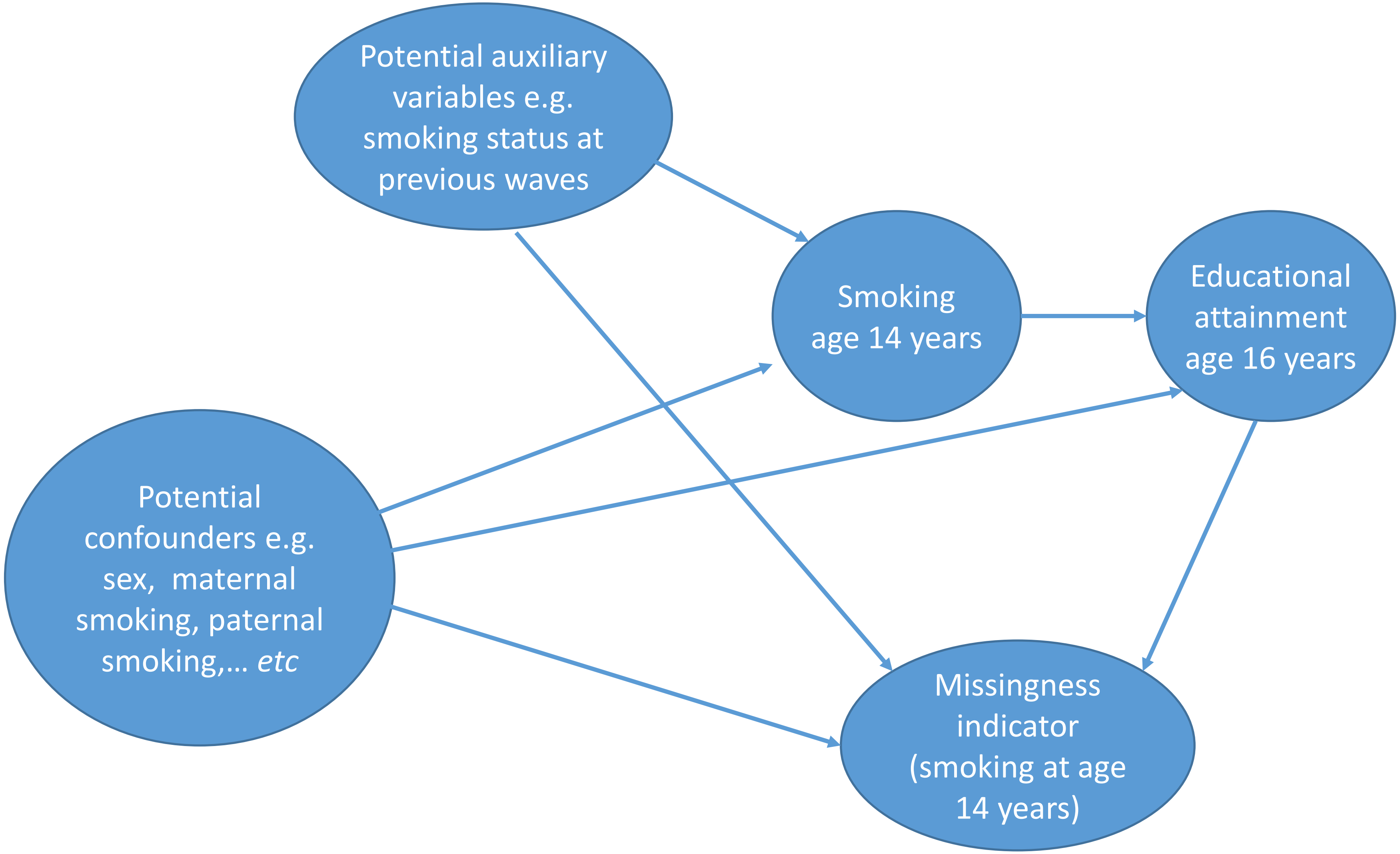
- Missing data are common in medical research
- Guidance is available, but appears not to have connected with many analysts: missing data are still often not handled appropriately
- Particularly problematic in observational research
- Therefore, we propose a practical framework for the **T**reatment **A**nd **R**eporting of **M**issing data in **O**bservational **S**tudies (TARMOS)
- Focus on multiple imputation (MI) because of its flexibility and practicality
- Focus on the estimation of an exposure-outcome association

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

- Is there a causal relationship between smoking at 14 years and educational attainment at 16 years?
  - 14,684 adolescents
  - Outcome: Educational attainment score obtained via linkage to the National Pupil Database
  - Exposure: current or non-smokers obtained via a computerised questionnaire during a clinic assessment and a postal questionnaire

Case Study: ALSPAC



Potential auxiliary variables e.g. smoking status at previous waves

Smoking age 14 years

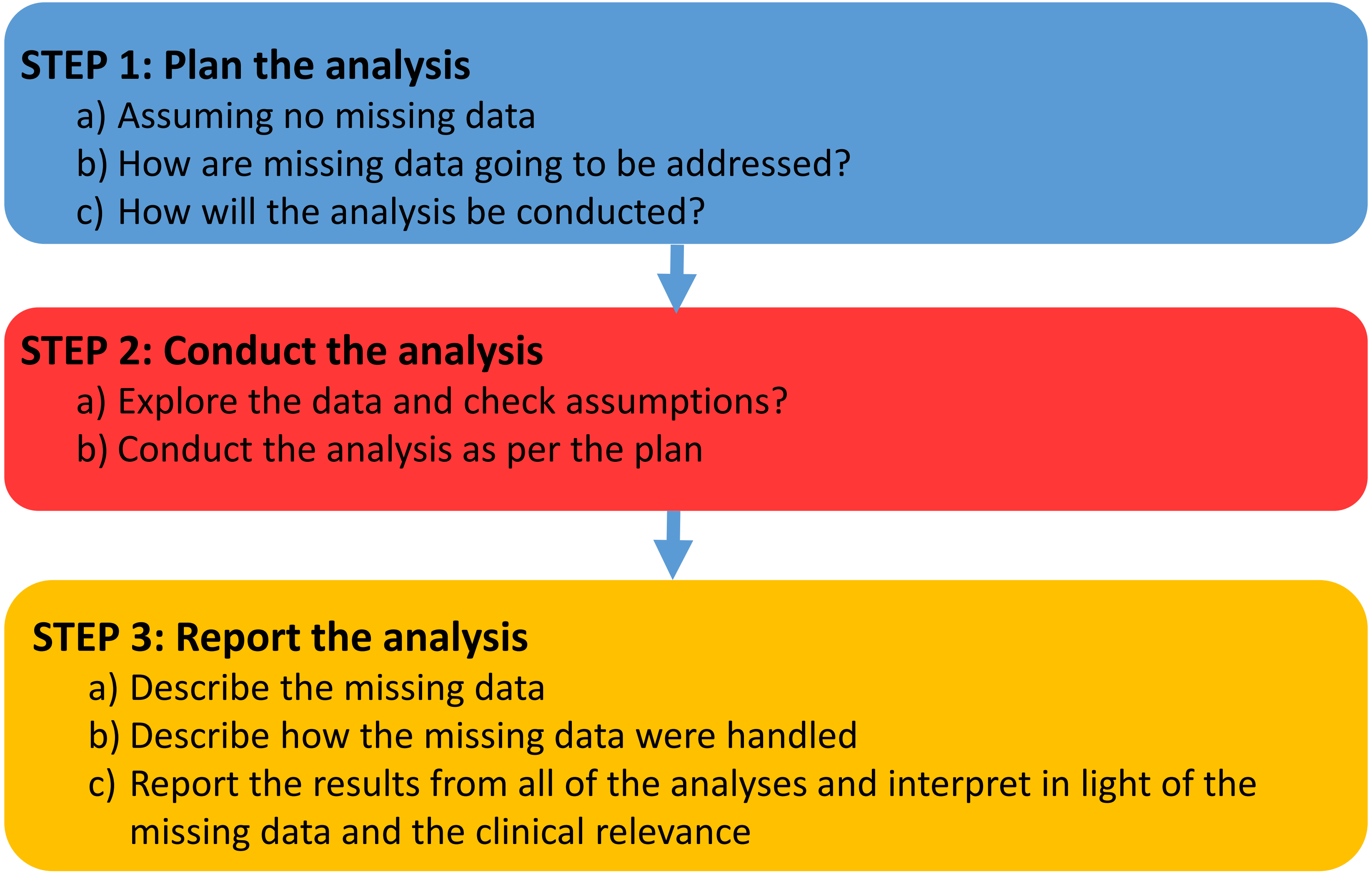
Educational attainment age 16 years

Potential confounders e.g. sex, maternal smoking, paternal smoking,… *etc*

Missingness indicator (smoking at age 14 years)

Missing data in all variables required for analysis (except sex)

The Framework



STEP 1: Plan the analysis

a) Assuming no missing data
b) How are missing data going to be addressed?
c) How will the analysis be conducted?

STEP 2: Conduct the analysis

a) Explore the data and check assumptions?
b) Conduct the analysis as per the plan

STEP 3: Report the analysis

a) Describe the missing data
b) Describe how the missing data were handled
c) Report the results from all of the analyses and interpret in light of the missing data and the clinical relevance

STEP 1a: Plan the analysis *if there were no missing data*

Pre-specify an analysis plan stating the primary and any secondary analyses

ALSPAC: Consistent with the causal graph, fit a linear regression of educational attainment score at 16 years on smoking at 14 years, adjusting for confounders

- sex, parity, maternal smoking, paternal smoking, maternal education, paternal education, behaviour at 81 months, educational attainment age 11 years

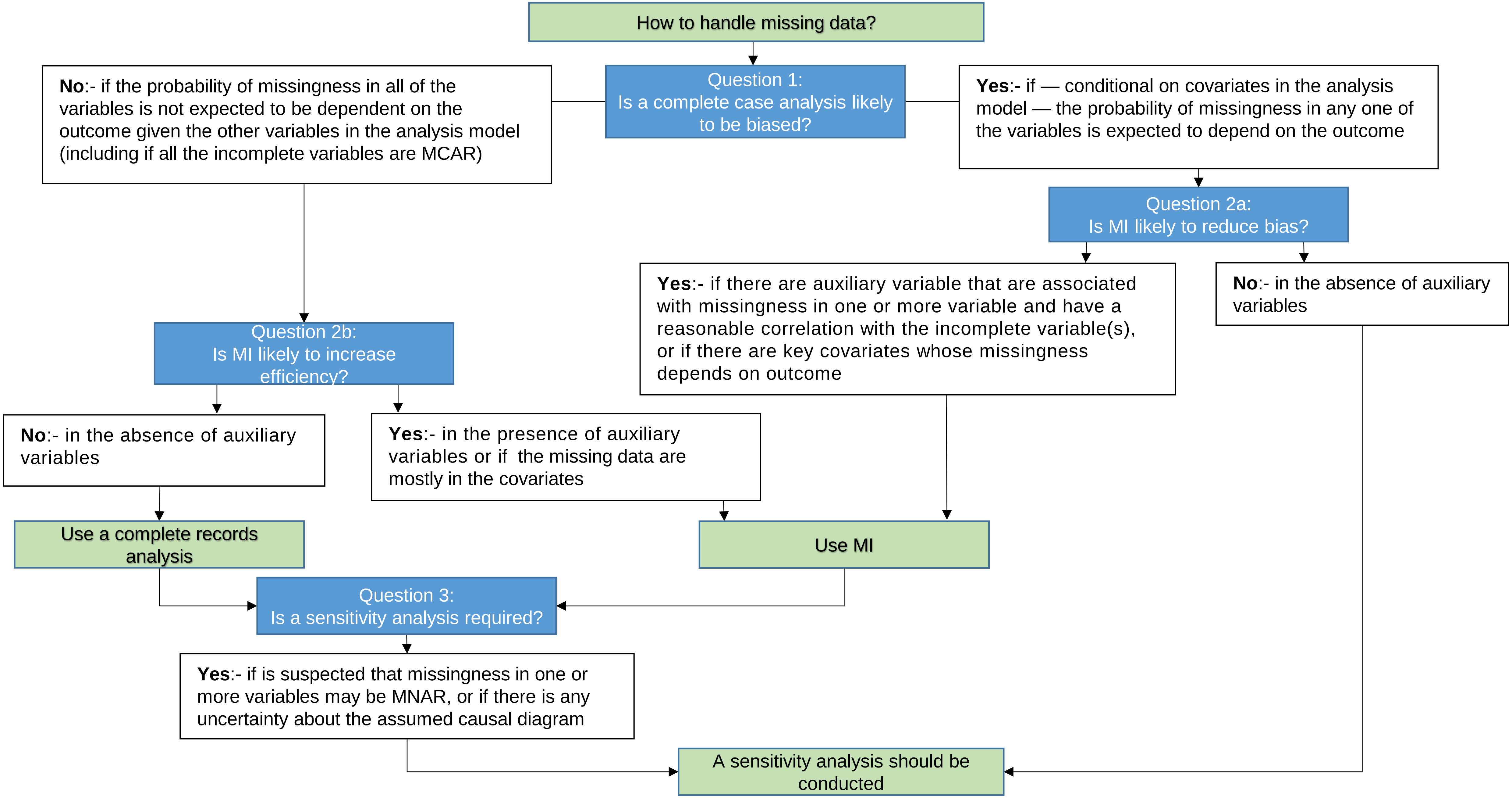
Step 1b: How are missing data to be addressed?

1. Is complete case analysis likely to be biased?
   - ➢ Yes, if the chance of missing values is related to outcome

2. Is MI likely to reduce the bias?
   - ➢ Yes, if either (a) incomplete data plausibly MAR given variables in model and (b) have good auxiliary variables

3. Is MI likely to increase efficiency?
   - ➢ Yes, if have good auxiliary variables and missing data mostly in the covariates

4. Is sensitivity analysis required?
   - ➢ Yes, if suspect data are MNAR or there is uncertainty about the missingness mechanism



How to handle missing data?

Question 1:
Is a complete case analysis likely to be biased?

No:- if the probability of missingness in all of the variables is not expected to be dependent on the outcome given the other variables in the analysis model (including if all the incomplete variables are MCAR)

Yes:- if — conditional on covariates in the analysis model — the probability of missingness in any one of the variables is expected to depend on the outcome

Question 2a:
Is MI likely to reduce bias?

Yes:- if there are auxiliary variable that are associated with missingness in one or more variable and have a reasonable correlation with the incomplete variable(s), or if there are key covariates whose missingness depends on outcome

No:- in the absence of auxiliary variables

Question 2b:
Is MI likely to increase efficiency?

No:- in the absence of auxiliary variables

Yes:- in the presence of auxiliary variables or if the missing data are mostly in the covariates

Use a complete records analysis

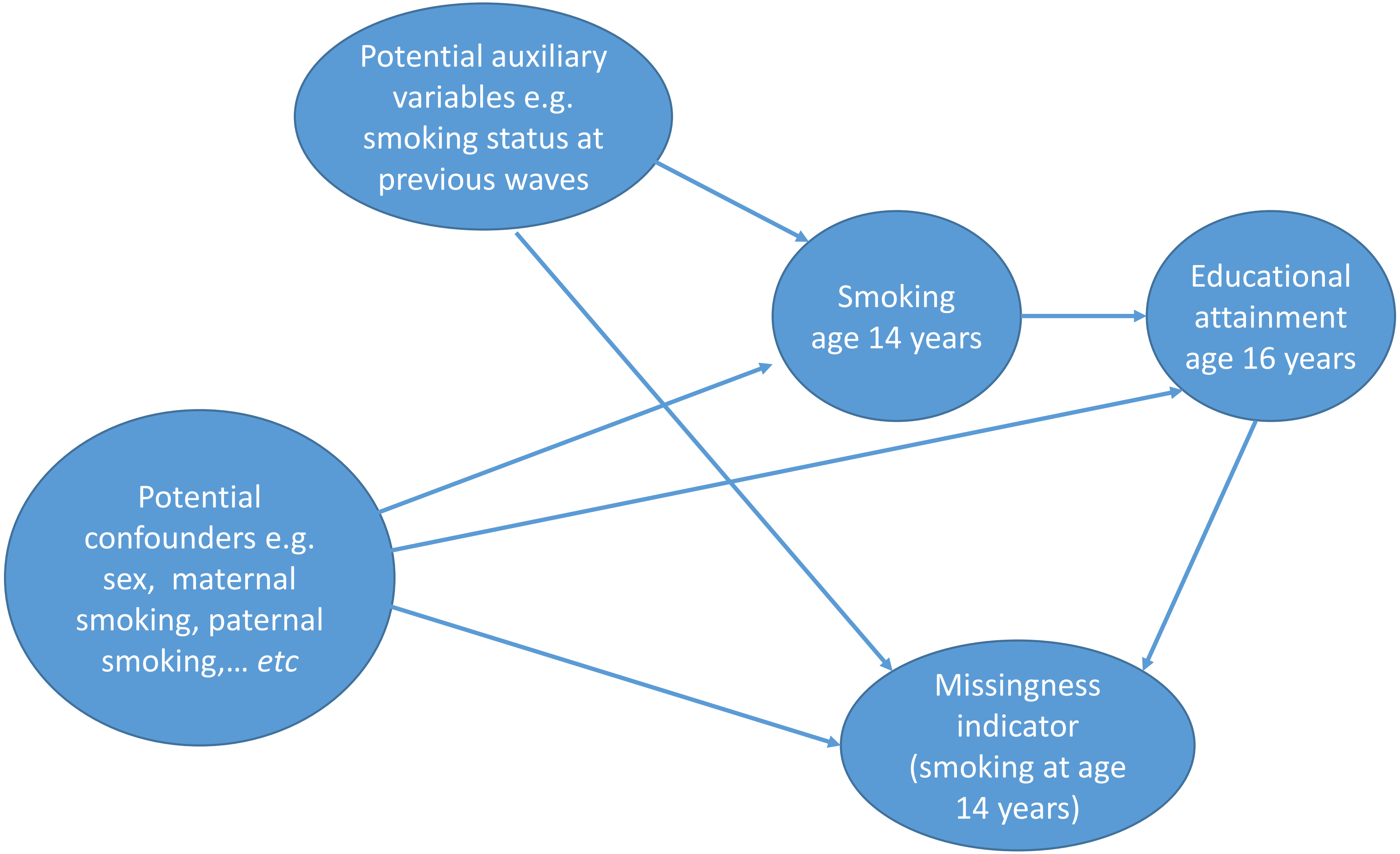
Use MI

Question 3:
Is a sensitivity analysis required?

Yes:- if is suspected that missingness in one or more variables may be MNAR, or if there is any uncertainty about the assumed causal diagram

A sensitivity analysis should be conducted

ALSPAC: Analysis planning


Potential auxiliary variables e.g. smoking status at previous waves
Smoking age 14 years
Educational attainment age 16 years
Potential confounders e.g. sex, maternal smoking, paternal smoking,… *etc*
Missingness indicator (smoking at age 14 years)

ALSPAC: Analysis planning

1. Is complete case analysis likely to be biased?
   - ➢ Yes, if the chance of missing values is related to outcome – this is true here

2. Is MI likely to reduce the bias?
   - ➢ Yes, if either (a) incomplete data plausibly MAR given variables in model and (b) have good auxiliary variables – both true in this example

3. Is MI likely to increase efficiency?
   - ➢ Yes, if have good auxiliary variables and missing data mostly in the covariates – both true in this example

4. Is sensitivity analysis required?
   - ➢ Yes, if suspect data are MNAR or there is uncertainty about the missingness mechanism – suspect educational attainment may be MNAR

Step 1c: How will the analysis be conducted?

Also need to plan the details of how the analysis will be conducted (including the justification)

- e.g. for MI
  - Method of imputation
  - Variables to be included
  - Form of variables
  - Nature of the relationships between variables
  - Method of imputation
  - Number of imputations
  - Software to be used
- Also details of how sensitivity analyses will be conducted
  - E.g. using pattern mixture approach
  - How the sensitivity parameter will be selected

ALSPAC: The planned analyses

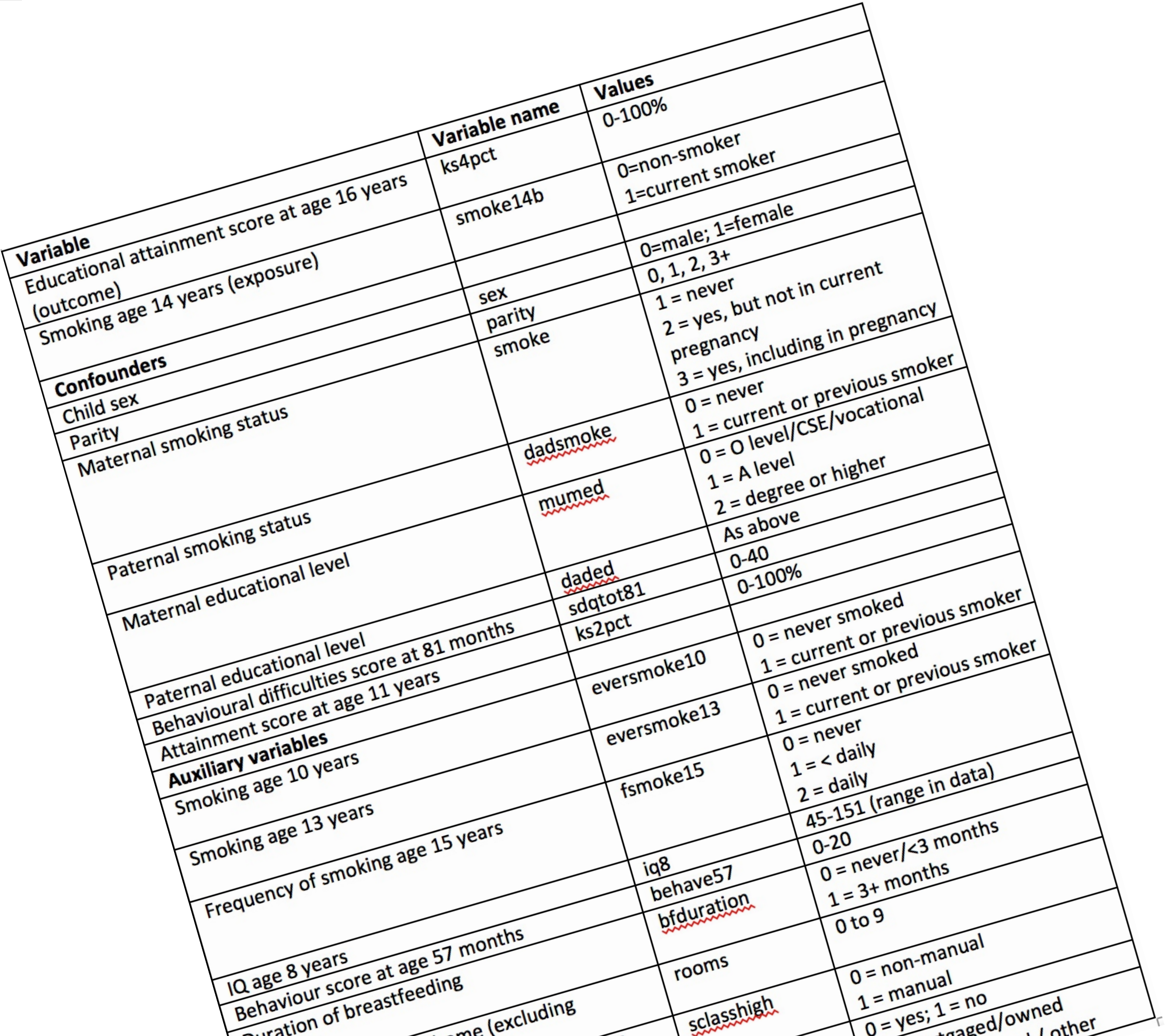
- MI (with auxiliary variables)
- Complete records (for comparison)
- Sensitivity analysis - MNAR
  - Pattern-mixture approach
  - Add the fixed log-odds of 0.1, 0.25, 0.5, 1 and 10 (extreme MNAR mechanism) within the logistic regression model used to impute smoking status
  - Conducted using the “offset” option within Stata’s *mi impute chained* command

Step 2a: Explore the data

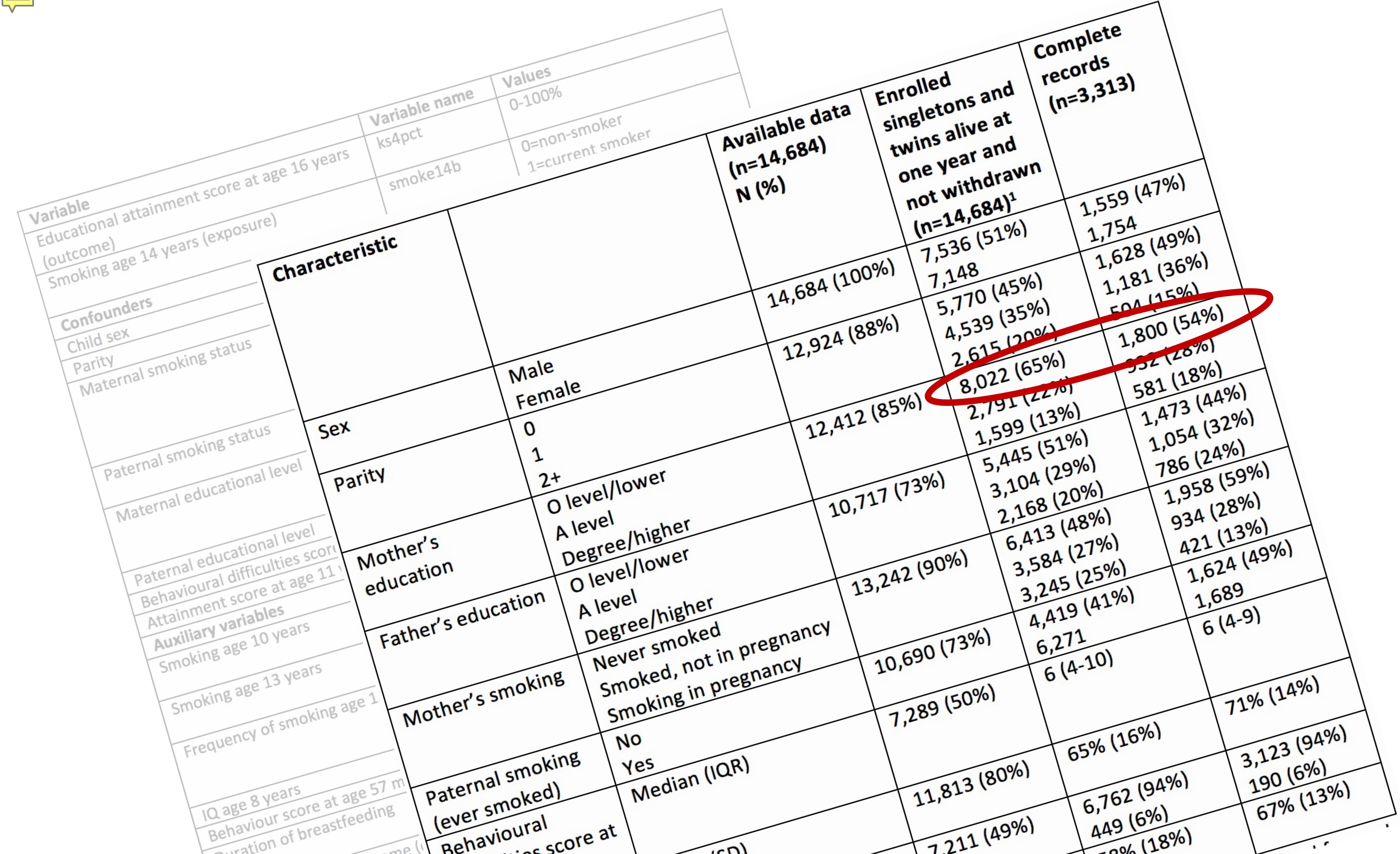
Provide:

- A table showing the proportion of missing data for all variables individually, and for the analysis model
- A table of the observed characteristics for the complete versus the incomplete participants
- An assessment of the predictors of missing (e.g. using logistic regression)

Use it to judge whether the methods outlined in the analysis plan are appropriate

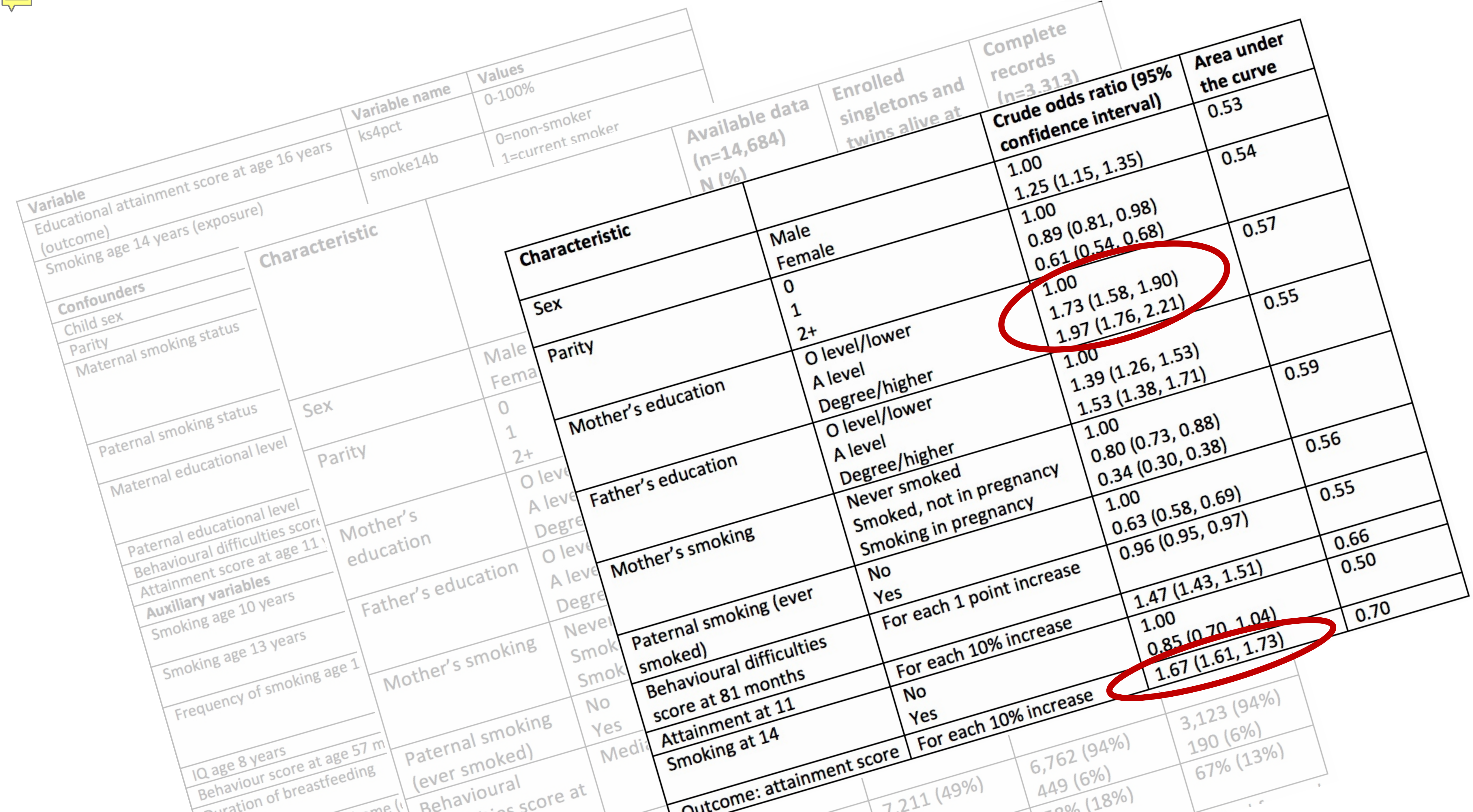


Variable	Variable name	Values
Educational attainment score at age 16 years (outcome)	ks4pct	0-100%
Smoking age 14 years (exposure)	smoke14b	0=non-smoker 1=current smoker
Confounders		
Child sex	sex	0=male; 1=female
Parity	parity	0, 1, 2, 3+
Maternal smoking status	smoke	1 = never 2 = yes, but not in current pregnancy 3 = yes, including in pregnancy
Paternal smoking status	dadsmoke	0 = never 1 = current or previous smoker
Maternal educational level	mumed	0 = O level/CSE/vocational 1 = A level 2 = degree or higher
Paternal educational level	daded	As above
Behavioural difficulties score at 81 months	sdqtot81	0-40
Attainment score at age 11 years	ks2pct	0-100%
Auxiliary variables		
Smoking age 10 years	eversmoke10	0 = never smoked 1 = current or previous smoker
Smoking age 13 years	eversmoke13	0 = never smoked 1 = current or previous smoker
Frequency of smoking age 15 years	fsmoke15	0 = never 1 = < daily 2 = daily
IQ age 8 years	iq8	45-151 (range in data)
Behaviour score at age 57 months	behave57	0-20
Duration of breastfeeding	bfduration	0 = never/<3 months 1 = 3+ months
me (excluding	rooms	0 to 9
	sclasshigh	0 = non-manual 1 = manual
		0 = yes; 1 = no
		tgaged/owned / other



Variable	Variable name	Values
Educational attainment score at age 16 years (outcome)	ks4pct	0-100%
Smoking age 14 years (exposure)	smoke14b	0=non-smoker 1=current smoker
Confounders		
Child sex		
Parity		
Maternal smoking status		
Paternal smoking status		
Maternal educational level		
Paternal educational level		
Behavioural difficulties score		
Attainment score at age 11 y		
Auxiliary variables		
Smoking age 10 years		
Smoking age 13 years		
Frequency of smoking age 1		
IQ age 8 years		
Behaviour score at age 57 m		
Duration of breastfeeding		
me (		

Characteristic		Available data (n=14,684) N (%)	Enrolled singletons and twins alive at one year and not withdrawn (n=14,684)[1]	Complete records (n=3,313)
Sex	Male Female	14,684 (100%)	7,536 (51%) 7,148	1,559 (47%) 1,754
Parity	0 1 2+	12,924 (88%)	5,770 (45%) 4,539 (35%) 2,615 (20%)	1,628 (49%) 1,181 (36%) 504 (15%)
Mother's education	O level/lower A level Degree/higher	12,412 (85%)	8,022 (65%) 2,791 (22%) 1,599 (13%)	1,800 (54%) 932 (28%) 581 (18%)
Father's education	O level/lower A level Degree/higher	10,717 (73%)	5,445 (51%) 3,104 (29%) 2,168 (20%)	1,473 (44%) 1,054 (32%) 786 (24%)
Mother's smoking	Never smoked Smoked, not in pregnancy Smoking in pregnancy	13,242 (90%)	6,413 (48%) 3,584 (27%) 3,245 (25%)	1,958 (59%) 934 (28%) 421 (13%)
Paternal smoking (ever smoked)	No Yes	10,690 (73%)	4,419 (41%) 6,271	1,624 (49%) 1,689
Behavioural difficulties score at	Median (IQR)	7,289 (50%)	6 (4-10)	6 (4-9)
	(SD)	11,813 (80%)	65% (16%)	71% (14%)
		7,211 (49%)	6,762 (94%) 449 (6%)	3,123 (94%) 190 (6%)
			78% (18%)	67% (13%)



Variable	Variable name	Values
Educational attainment score at age 16 years (outcome)	ks4pct	0-100%
Smoking age 14 years (exposure)	smoke14b	0=non-smoker 1=current smoker
Confounders		
Child sex		
Parity		
Maternal smoking status		
Paternal smoking status		
Maternal educational level		
Paternal educational level		
Behavioural difficulties score		
Attainment score at age 11		
Auxiliary variables		
Smoking age 10 years		
Smoking age 13 years		
Frequency of smoking age 1		
IQ age 8 years		
Behaviour score at age 57 m		
Duration of breastfeeding		

Characteristic		Crude odds ratio (95% confidence interval)	Area under the curve
Sex	Male Female	1.00 1.25 (1.15, 1.35)	0.53
Parity	0 1 2+	1.00 0.89 (0.81, 0.98) 0.61 (0.54, 0.68)	0.54
Mother's education	O level/lower A level Degree/higher	1.00 1.73 (1.58, 1.90) 1.97 (1.76, 2.21)	0.57
Father's education	O level/lower A level Degree/higher	1.00 1.39 (1.26, 1.53) 1.53 (1.38, 1.71)	0.55
Mother's smoking	Never smoked Smoked, not in pregnancy Smoking in pregnancy	1.00 0.80 (0.73, 0.88) 0.34 (0.30, 0.38)	0.59
Paternal smoking (ever smoked)	No Yes	1.00 0.63 (0.58, 0.69)	0.56
Behavioural difficulties score at 81 months	For each 1 point increase	0.96 (0.95, 0.97)	0.55
Attainment at 11	For each 10% increase	1.47 (1.43, 1.51)	0.66
Smoking at 14	No Yes	1.00 0.85 (0.70, 1.04)	0.50
Outcome: attainment score	For each 10% increase	1.67 (1.61, 1.73)	0.70

Step 2b: Conduct the planned analysis

- Proceed once satisfied the assumptions made in the analysis plan are acceptable
- If the analysis plan needs to be revised, any changes should be acknowledged and justified
- In ALSPAC, data exploration confirmed the assumptions in the analysis plan, hence we proceed with the pre-planned MI and sensitivity analysis

Step 3: Report the analysis

- Describe the extent of missing data and reasons for missing values if possible
- State how the missing data were addressed in the analyses and whether this was pre-specified
- Report the inference from the various analyses
- Interpret results in light of the missing data and the clinical relevance

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

ALSPAC: Results

Method of Analysis	Regression coefficient (95% CI)	p	% of missing smoking values imputed as “smokers”
Primary analysis: Multiple imputation	-10.8 (-12.2, -9.4)	<0.001	13.3
Complete records analysis	-7.9 (-9.1, -6.7)	<0.001	N/A
Sensitivity Analysis – sensitivity parameter = 0.1	-10.9 (-12.4, -9.4)	<0.001	14.2
Sensitivity Analysis – sensitivity parameter = 0.25	-11.0 (-12.3, -9.6)	<0.001	15.5
Sensitivity Analysis – sensitivity parameter = 0.5	-11.0 (-12.3, -9.6)	<0.001	18.1
Sensitivity Analysis – sensitivity parameter = 1	-10.7 (-11.8, -9.6)	<0.001	24.2
Sensitivity Analysis – sensitivity parameter = 10	-4.3 (-4.7, -3.8)	<0.001	99.8

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

Discussion

- The TARMOS framework gives practical, non-technical guidance with the aim of facilitating
  - **Planning**: informed discussion of the key issues among the research team, whether complete records is likely to be biased and the extent that MI may help
  - **Conduct**: choice of an appropriate imputation strategy, including use of auxiliary variables
  - **Reporting:** accurate reporting, including (i) the pattern and extent of missing data; (ii) comparison of complete records and MI analysis, and (iii) results of sensitivity analysis
- The framework adopts MI as the most general, practical method for the majority of researchers; however the principles apply whatever statistical method is used to handle the missing data.

STRATOS TG1: future plans

- Forthcoming manuscripts on
  - ➢Level 1: comparison of complete cases, weighting and multiple imputation with a social science application
  - ➢Level 2: Illustrated comparison of direct likelihood, EM algorithm, MI, IPW and AIPW (doubly robust) approaches
  - ➢Level 2/3: guidance for handling missing data in longitudinal causal models

Reference

- 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. http://arxiv.org/abs/2004.14066