Handling missing data in observational studies:

challenges for training and research

James Carpenter

London School of Hygiene and Tropical Medicine, London, UK, &
MRC Clinical Trials Unit, London, UK

james.carpenter@lshtm.ac.uk

Acknowledgements

TG1 Topic Group:

Melanie Bell, Els Goetghebeur, Joe Hogan, Kate Lee, Rod Little, Andrea Rotnitzky, Kate Tilling, Ian White

Elizabeth Williamson & Clemence Leyrat (LSHTM)

Tim Morris (LSTHM and MRC-CTU)

Mike Kenward (LSHTM)

Outline

- Context
- Proposed framework & examples:
    - Careful analysis of the complete records (complete cases)
    - Keep in mind the scientific context
    - Perform an analysis under MAR & use auxiliary variables
    - Perform simple sensitivity analyses
    - Know when a standard approach will be inadequate
    - Report the results clearly
- Discussion

Context

Missing data are ubiquitous, and the and the problem is not going away, partly because of

- the increasing use of routinely collected data (collected for clinical, not research needs), alongside
- the increasing reluctance of people to participate in studies.

Despite a large number of review papers (e.g. [11], [7],[8]), missing data are often poorly handled ([9],[16]) and things are only changing slowly [4].

Why is this?

There are a variety of reasons, for example innate conservatism; lack of a single ‘solution’; the variety of methodolgies and software.

But given the focus of the STRATOS initiative, I am drawn to two points:

- Level 1 Analysts
  Many analyses are undertaken by research staff with limited formal statistical training, who often find it tough to cope with missing data.
- Level 2 Analysts
  Many of the issues featured on our Banff programme are relatively peripheral to many applied statisticians’ education. For example, students on the MSc Medical Statistics at the LSHTM have only one day on missing data.

Possible way forward

We need to accelerate the adoption of the methodological developments in this area.

I propose addressing this by arguing for a framework (which all researchers can relate to) and then showing how this applies to analysts with different levels of statistical training.

A related approach in the clinical trials arena has been making progress (e.g. [13], and related publications).

Proposed framework

- Careful analysis of the complete records (complete cases)
- Keep in mind the scientific context
- Perform an analysis under MAR and use auxiliary variables
- Perform simple sensitivity analyses
- Know when a standard approach will be inadequate
- Report the results clearly

Complete Records/Cases Analysis

The data should be carefully explored with complete records analysis before more sophisticated techniques are used.

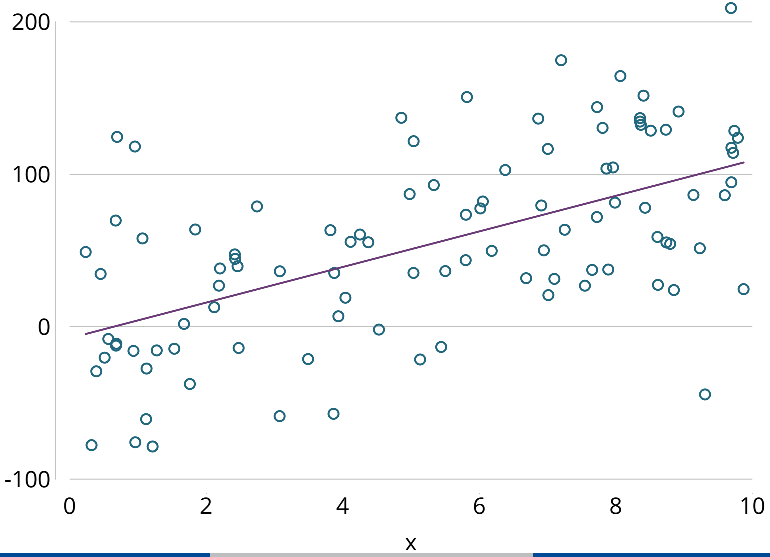
The conditions under which a complete records analysis gives valid inference are not difficult to grasp, but are not widely appreciated.

These are that, if the probability of a record being complete does not depend on the dependent ('Y') variable in the regression, the complete records analysis gives valid inference.

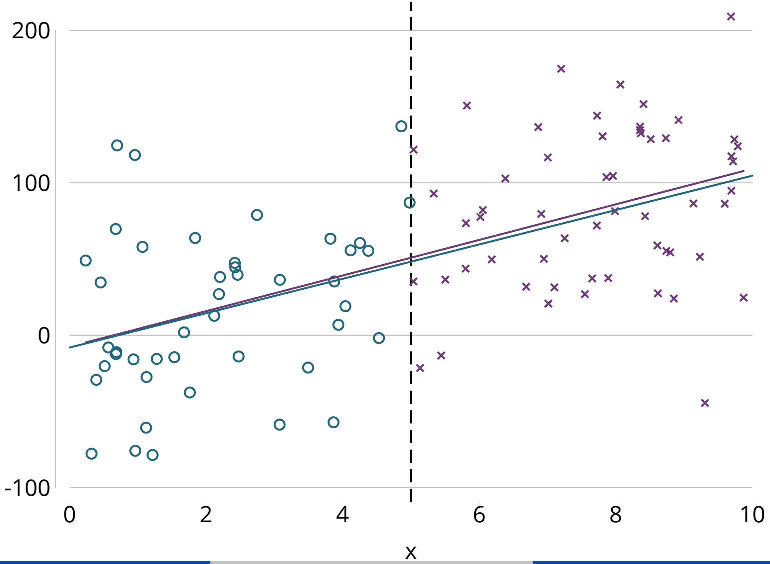
Thus the complete records analysis will be valid (if inefficient) in many situations, and should always be compared to the results of more sophisticated analyses.

How can this be explained to 'Level-1' analysts? A simple graphical approach works well:

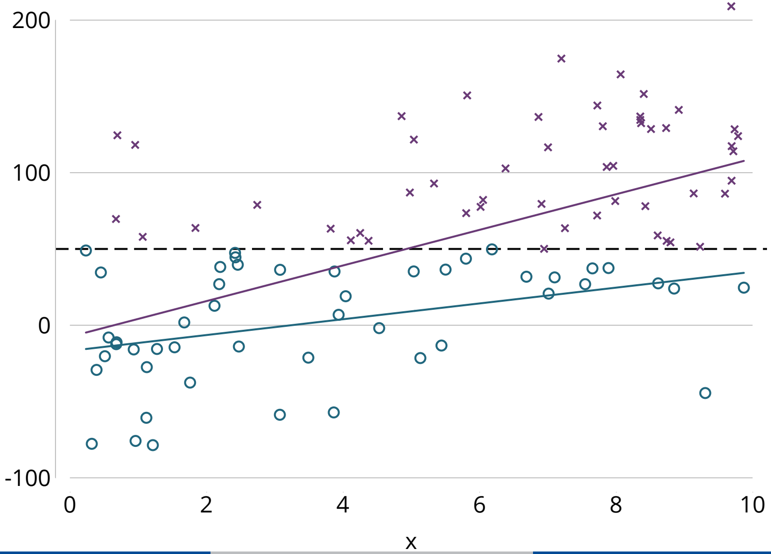
Intuition – full data analysis



Intuition – extreme MNAR



Intuition – extreme MAR



Example

- Bartlett *et al* [1] report results of an illustrative analysis based on cross-sectional data from the US NHANES 2003–04 study.
- They fit a regression model for systolic blood pressure (SBP) with no. of alcoholic drinks, BMI, and age as covariates.
- No. of alcoholic drinks was missing for 34% of individuals.
- Missingness in this variable may well be related to level of alcohol consumption (*i.e.* MNAR), age, and (maybe) BMI, but given these is probably unrelated to SBP.
- If this assumption is true, the CCA is valid, even though the covariate is (assumed to be) MNAR.

Other situations in which Complete Records are valid

A. If the substantive model is a logistic regression, e.g.

$$\mathrm{logit}\{\Pr(Y=1)\} = \beta_0 + \beta_X X + \beta_Z Z,$$

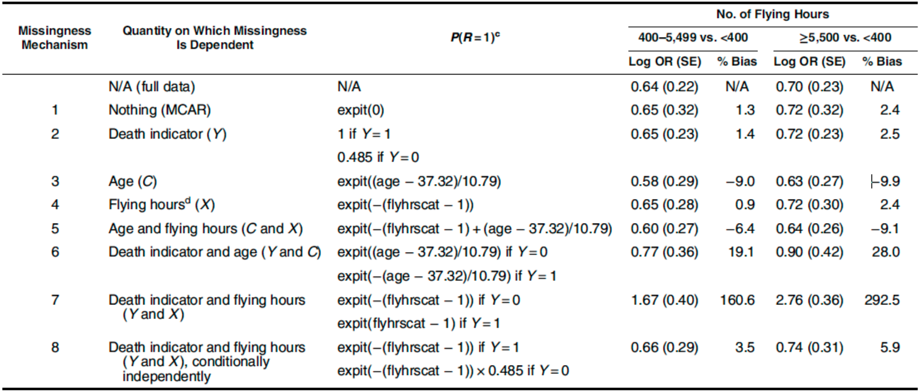
then, assuming the probability of a complete record involves Y,

- if X is not involved, β_X is unbiased;
- if Z is not involved, β_Z is unbiased [17, 2].

> If the probability of a complete record does not involve the exposure, then the exposure effect is unbiased.

B. If we set aside a part of the data where missingness causes bias, leaving a partial likelihood where the missingness is ignorable, then we may be able to use complete records analysis; or at least an analysis assuming missing at random [10]

Example: Flight Crew Mortality and Flying Hours, 1969–99[2]



Missingness Mechanism	Quantity on Which Missingness Is Dependent	$P(R=1)^c$	No. of Flying Hours			
			400–5,499 vs. <400		≥5,500 vs. <400	
			Log OR (SE)	% Bias	Log OR (SE)	% Bias
	N/A (full data)	N/A	0.64 (0.22)	N/A	0.70 (0.23)	N/A
1	Nothing (MCAR)	expit(0)	0.65 (0.32)	1.3	0.72 (0.32)	2.4
2	Death indicator (*Y*)	1 if $Y=1$ 0.485 if $Y=0$	0.65 (0.23)	1.4	0.72 (0.23)	2.5
3	Age (*C*)	expit((age − 37.32)/10.79)	0.58 (0.29)	−9.0	0.63 (0.27)	−9.9
4	Flying hours[d] (*X*)	expit(−(flyhrscat − 1))	0.65 (0.28)	0.9	0.72 (0.30)	2.4
5	Age and flying hours (*C* and *X*)	expit(−(flyhrscat − 1) + (age − 37.32)/10.79)	0.60 (0.27)	−6.4	0.64 (0.26)	−9.1
6	Death indicator and age (*Y* and *C*)	expit((age − 37.32)/10.79) if $Y=0$ expit(−(age − 37.32)/10.79) if $Y=1$	0.77 (0.36)	19.1	0.90 (0.42)	28.0
7	Death indicator and flying hours (*Y* and *X*)	expit(−(flyhrscat − 1)) if $Y=0$ expit(flyhrscat − 1) if $Y=1$	1.67 (0.40)	160.6	2.76 (0.36)	292.5
8	Death indicator and flying hours (*Y* and *X*), conditionally independently	expit(−(flyhrscat − 1)) if $Y=1$ expit(−(flyhrscat − 1)) × 0.485 if $Y=0$	0.66 (0.29)	3.5	0.74 (0.31)	5.9

Scientific Context

Ignoring the scientific context, and rushing to publish a more complex analysis, is unlikely to end well.

Example

- The QRISK[6] study aimed to derive a new cardiovascular disease (CVD) risk score for the UK, based on routinely collected data from general practice.
- The score was derived using data from 1.28 million patients registered at UK GP practices between 1995 and 2007, who were free from CVD at registration
- The outcome of interest was time to first recorded diagnosis of CVD
- Cox proportional hazards models were used to model time to CVD, as a function of risk factors measured at registration

Missing data in QRISK

- Inevitably there was substantial missingness in ‘baseline’ risk factor data
- In particular, 70% of subjects had HDL cholesterol missing
- The investigators used MI to deal with missing baseline data, using the `ice` (the forerunner to `mi impute chained`) command in Stata

Cholesterol and CVD

- In the final model, the adjusted hazard ratio for the ratio of total to HDL cholesterol was 1.001 (95% 0.999 to 1.002)
- This suggested that, after adjusting for other baseline risk factors, cholesterol had no effect on CVD risk
- Given that cholesterol has been shown to have an independent effect on CVD risk in many previous studies, this result was unexpected

Selected comments on the paper

> *the hazard ratio for total cholesterol/high density lipoprotein cholesterol is completely inconsistent with numerous previous studies (Wild)*

> *Until more details of the materials and methods of this new QRISK study are made available, the reliable conclusion should be retained from previous studies that the ratio of total to HDL cholesterol (undistorted by lipid-lowering drugs) is strongly predictive of the primary incidence of coronary disease (Peto)*

> *It is no surprise therefore that cholesterol did not appear to contribute to risk in the entire QRISK population (Simpson)*

All comments at:
`http://www.bmj.com/content/335/7611/136/rapid-responses`

What had gone wrong (Tim Morris)[12]

The authors stated:

> *We imputed total serum cholesterol and HDL separately in the original [imputation] model. We then calculated the ratio by dividing total serum cholesterol by HDL*

Such passive imputation is best avoided but here was disastrous. Imputed HDL values had relatively high variance, so that many were close to zero, massively inflating the total/HDL ratio and removing the association with heart disease.

Perform an analysis under MAR and use auxiliary variables

Use of appropriate auxiliary variables can markedly improve the analysis. However, the following points need to be remembered:

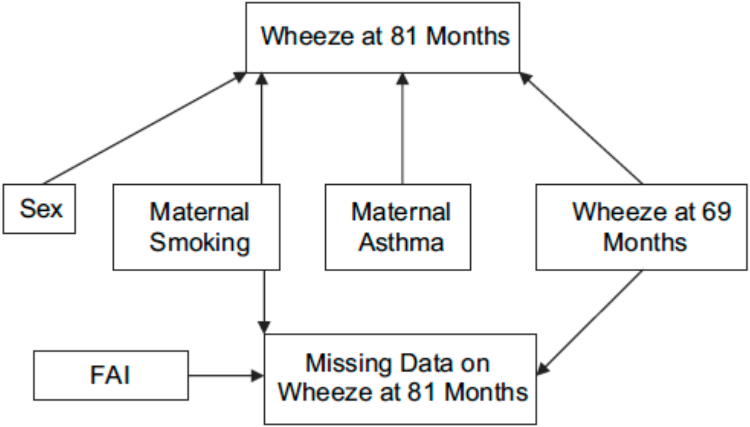
Auxiliary variables that:

1. only predict missingness are not useful;
2. only predict the underlying (missing) values improve precision;
3. do both (1) and (2) remove bias and increase precision.

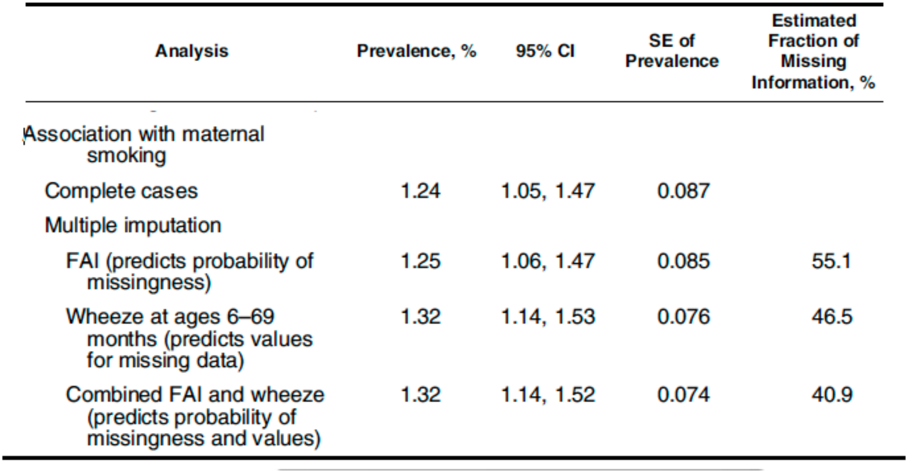
Analysts should be taught how to identify and use such variables.

Depending on the statistical method used, this is relatively straightforward.

Example: Wheeze at 81 months in the ALSPAC data[15]



Results



Analysis	Prevalence, %	95% CI	SE of Prevalence	Estimated Fraction of Missing Information, %
Association with maternal smoking				
Complete cases	1.24	1.05, 1.47	0.087	
Multiple imputation				
FAI (predicts probability of missingness)	1.25	1.06, 1.47	0.085	55.1
Wheeze at ages 6–69 months (predicts values for missing data)	1.32	1.14, 1.53	0.076	46.5
Combined FAI and wheeze (predicts probability of missingness and values)	1.32	1.14, 1.52	0.074	40.9

Appropriate use of auxiliary variables recovers substantial information and corrects bias.

Perform simple sensitivity analyses

Case-control study of Sudden Infant Death Syndrome: [5] report a case control study to investigate whether bed sharing is a risk factor for Sudden Infant Death Syndrome (SIDS). This is an IPD meta-analysis of data from five case-control studies, with in total 1472 cases and 4679 controls.

Unfortunately, data on alcohol and drug use were unavailable in three of the five studies (about 60% of the data).

The reason was the study did not collect them: i.e. study is the predictor of missing data!

The substantive model adjusted for study, so missingness was MAR dependent on covariates.

We expect MI under MAR to gain information (especially on coefficients of variables with data excluded from the CR analysis), but not change the associations substantially.

SIDS study

Sometimes, a simple best/worst case scenario analysis in a key subgroup will be sufficient, particularly if the partially observed variable is binary.

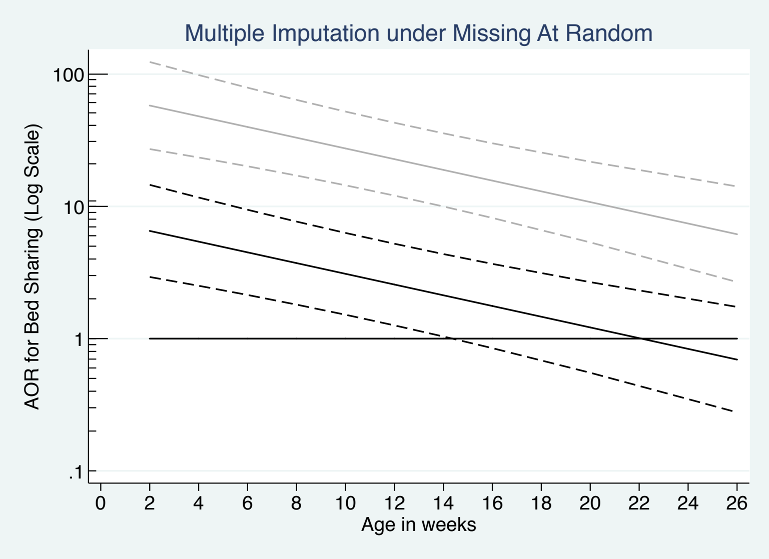
For example, in the SIDS study, critics have argued that the risk of bed-sharing is confounded with alcohol use.

To explore the robustness of our inference to this assumption, the simplest approach is to impute 'alcohol use' to the mothers of all the cases who were bed-sharing, and 'no alcohol use' to all controls who were bed-sharing.

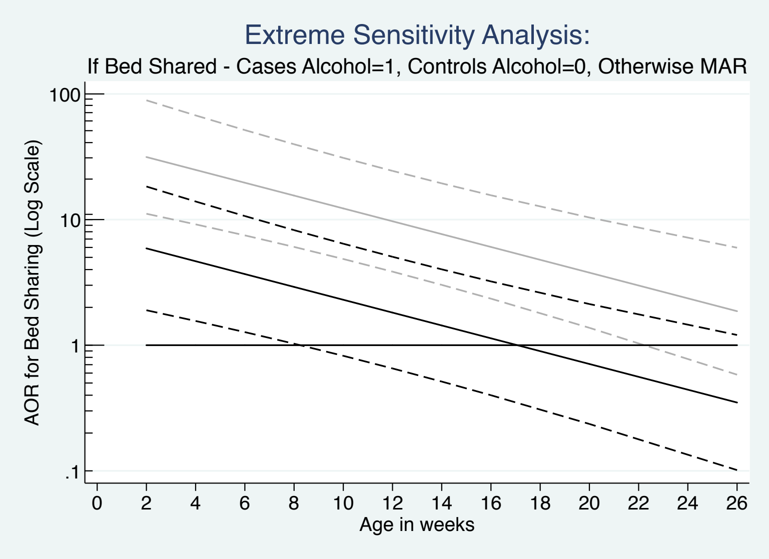
Remaining missing values (in each imputed data set) are left at the values imputed by MI assuming MAR.

We do a similar analysis for drug use.

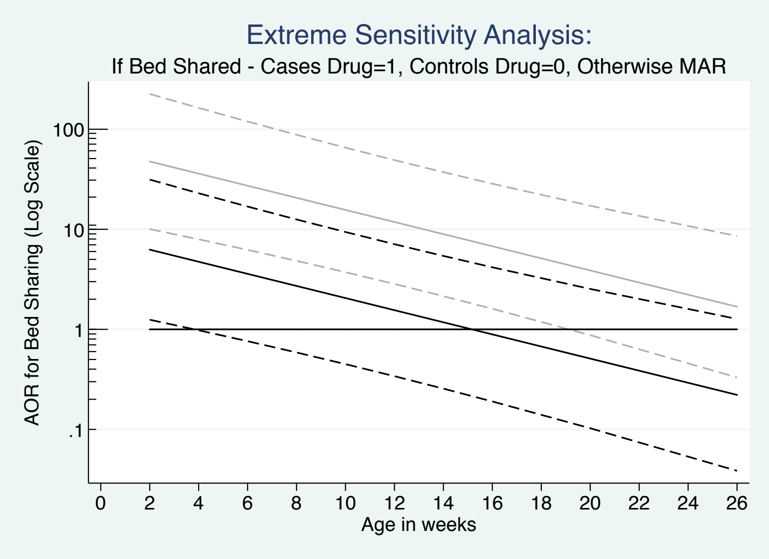
Results


Multiple Imputation under Missing At Random
AOR for Bed Sharing (Log Scale)
100
10
1
.1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
Age in weeks

Results


Extreme Sensitivity Analysis:
If Bed Shared - Cases Alcohol=1, Controls Alcohol=0, Otherwise MAR
AOR for Bed Sharing (Log Scale)
100
10
1
.1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
Age in weeks

Results


Extreme Sensitivity Analysis:
If Bed Shared - Cases Drug=1, Controls Drug=0, Otherwise MAR
AOR for Bed Sharing (Log Scale)
100
10
1
.1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
Age in weeks

Know when a standard approach is inadequate

There are now a number of well established software packages for MI, and other methods. These can be expected to work well, for all levels of users, provided the substantive model has a linear structure, and $p << n$.

However analysts often wish to fit more complex models, e.g.

- splines & interactions [3]
- hierarchical structure [14]
- survival; competing risks [18, 3]
- causal models
- combinations of the above
- ...

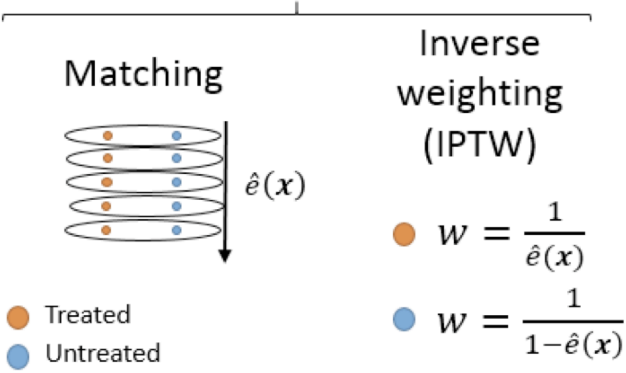
and handle different missing data issues (e.g. linkage)

In general these require more sophisticated methods, which are increasingly available, but require greater understanding (e.g. level 2?) to use appropriately.

Example: propensity scores (Williamson & Leyrat)

Propensity scores (PS) proposed in 1983 to **balance groups** in observational studies:

$$\hat{e}(\boldsymbol{x}) = P(T = 1|\boldsymbol{x})$$



● $w = \frac{1}{\hat{e}(\boldsymbol{x})}$

● $w = \frac{1}{1-\hat{e}(\boldsymbol{x})}$

● Treated

● Untreated

Two key questions:

- Should the outcome be included in the imputation model ?

Omitting the outcome gives biased results.

- How to apply Rubin's rules?

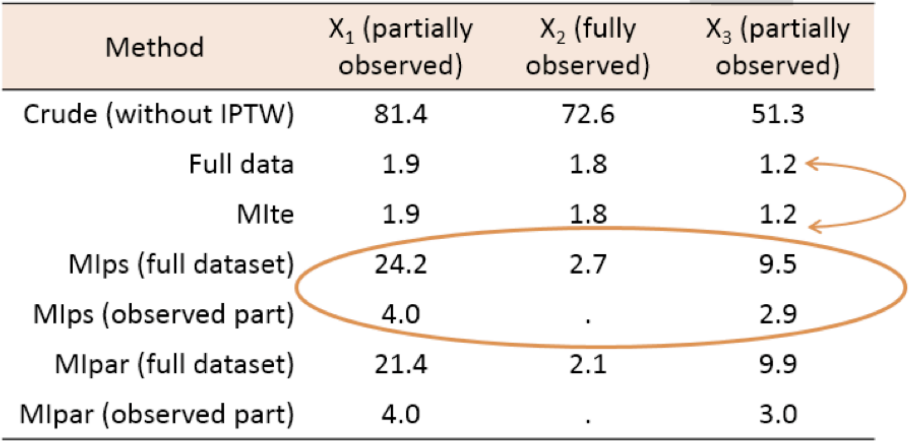
$\Longrightarrow$ pooled treatment effect or pooled PS?

Simulation study to confirm theory:

- Complete case: exclusion of participants with partial data
- Missingness pattern: 4 different PS models
- MIte: the K IPTW estimates of the treatment effect are pooled according Rubin's rules
- MIps: 1 IPTW estimate obtained from the average PS
- MIpar: 1 IPTW estimate obtained from the PS of the average covariates

Results: Balancing properties

Standardized differences (in%) between groups: SD $= \frac{100\times|\bar{X}_1-\bar{X}_0|}{\sqrt{\frac{s_0^2+s_1^2}{2}}}$



Method	X_1 (partially observed)	X_2 (fully observed)	X_3 (partially observed)
Crude (without IPTW)	81.4	72.6	51.3
Full data	1.9	1.8	1.2
MIte	1.9	1.8	1.2
MIps (full dataset)	24.2	2.7	9.5
MIps (observed part)	4.0	.	2.9
MIpar (full dataset)	21.4	2.1	9.9
MIpar (observed part)	4.0	.	3.0

PS obtained from MP, MIps and MIpar **do not balance the missing part** of the covariates

Report the results clearly

For any analysis potentially affected by missing data[16]:

1. Report the number of missing values for each variable of interest. Give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when reporting the flow of participants through the study. If possible, describe reasons for missing data in terms of other variables.

2. Clarify whether there are important differences between individuals with complete and incomplete data.

3. For analyses that account for missing data, describe the nature of the analysis (e.g. multiple imputation), and the assumptions that were made (e.g. missing at random).

For analyses based on multiple imputation:

1. Provide details of the imputation modelling: software, number of imputations, variables in imputation model, use of interactions, transformations.
2. If a large fraction of the data is imputed, give a comparison of observed and imputed values. Marked differences need a careful explanation.
3. Where possible, provide results from analyses restricted to complete cases, for comparison with results based on multiple imputation. If there are important differences between the results, suggest explanations, bearing in mind that analyses of complete cases may suffer more chance variation, and that under the MAR assumption multiple imputation should correct biases that may arise in complete-cases analyses.
4. Discuss whether the variables included in the imputation model make the missing at random assumption plausible.
5. Include discussion of the robustness of key inferences to possible departures from the MAR assumption

How to choose between the methods

The assumptions are key, but given this the following considerations are important for widespread use:

- established software for ‘standard’ settings (essentially where relationships are linear);
- ability to include auxiliary variables;
- relatively simple to do sensitivity analyses;
- can be used for a range of linked issues (e.g. linkage, measurement error, disclosure)
- can handle large datasets

These issues suggest that MI may be the tool to focus on, although other methods (direct likelihood, EM, IPW, AIPW) may be preferable in specific situations.

Level 3 question: How far should we develop MI algorithms?

- MI is a Bayesian method with good frequentist properties.
- In tackling more complex problems (causal modelling, high dimensional data, modelling non-linear relationships), when does directly fitting the Bayesian model become preferable to MI?
- In other words, how much work should we put into developing MI algorithms that accommodate specific substantive models?
    - For non-linear relationships [3] and hierarchical models [14] the effort has been worthwhile.
    - For network meta-analysis? For high-dimensional models?

Discussion

What should STRATOS contribute to the missing data field?

- There are now a large number of tutorials and reviews of the missing data literature.
- However the literature is somewhat disconnected.

The proposal is to accelerate the adoption of good practice among ‘Level 1’ and ‘Level 2’ analysts by

- proposing a framework for considering and addressing missing data issues which is
  - accessible to all analysts;
  - links — at the appropriate level — to existing tutorials and software
  - helps highlight priority areas for future work: e.g. with increasingly sophisticated analyses, is a generic approach appropriate
- A similar approach has been effective in handling missing data in clinical trials (see [13] and subsequent papers).

References I

[1] J W Bartlett, J R Carpenter, K Tilling, and S Vansteelandt. Improving upon the efficiency of complete case analysis when covariates are mnar. *Biostatistics*, 15:719–730, 2014.

[2] J W Bartlett, O Harel, and J R Carpenter. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology*, 182:730–736, 2015.

[3] J W Bartlett, S Seaman, I R White, and J R Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24:462–487, 2015.

[4] Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in rcts; a review of the top medical journals. *BMC Medical Research Methodology*, 14:118, 2014.

[5] R G Carpenter, C McGarvey, E A Mitchell, D M Tappin, M M Vennemann, M Smuk, and J R Carpenter. Bed sharing when parents to not smoke: is there a risk of sids? an individual level analysis of five major case-control studies. *BMJ Open*, 3:e002299, 2013.

[6] J Hippisley-Cox, C Coupland, Y Vinogradova, J Robson, M May, and P Brindle. Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. *British Medical Journal*, 335:7611–7623, 2007.

[7] J W Hogan, J Roy, and C Krokontzelou. Tutorial in biostatistics: handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455–1497, 2004.

References II

[8] N J Horton and K P Kleinman. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, 61(1):79–90, 2007.

[9] M A Klebanoff and S R Cole. Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168:355–357, 2008.

[10] R J Little and N Zhang. Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 60:591–605, 2011.

[11] R J A Little. Regression With Missing X's: A Review. *Journal of the American Statistical Association*, 87:1227–1237, 1992.

[12] T P Morris, I R White, P Royston, S R Seaman, and A M Wood. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine*, 33:88–104, 2014.

[13] National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials.* Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2010.

[14] M. Quartagno and J. R. Carpenter. Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, pages n/a–n/a, 2015.

[15] M Spratt, J R Carpenter, J A C Sterne, B Carlin, J Heron, J Henderson, and K Tilling. Strategies for Multiple Imputation in Longitudinal Studies. *American Journal of Epidemiology*, 172:478–487, 2010.

References III

[16] J A C Sterne, I R White, J B Carlin, M Spratt, P Royston, M G Kenward, A M Wood, and J R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 339:157–160, 2009.

[17] W Vach and M Blettner. Logistic regression with incompletely observed categorical covariates — investigating sensitivity against violation of the Missing at Random assumption. *Statistics in Medicine*, 54:1315–1329, 1999.

[18] I R White and P Royston. Imputing missing covariate values for the cox model. *Statistics in Medicine*, 28:1982–1998, 2009.