Missing data:

an overview of the problem and the solutions

James R. Carpenter

James.Carpenter@lshtm.ac.uk

London School of Hygiene and Tropical Medicine &

MRC Clinical Trials Unit at UCL

www.missingdata.org.uk

Support from ESRC and MRC

1st EDEN Forum, Madrid, 31 March 2017

Acknowledgements

John Carlin, Lyle Gurrin, Helena Romaniuk, Kate Lee (Melbourne)
Mike Kenward, Harvey Goldstein (LSHTM)
Geert Molenberghs (Limburgs University, Belgium)
James Roger (GlaxoSmithKline Research)
Sara Schroter (BMJ, London)
Jonathan Sterne, Michael Spratt, Rachael Hughes (Bristol)
Stijn Vansteelandt (Ghent University, Belgium)
Ian White (MRC Biostatistics Unit, Cambridge)

Much of the material for this lecture is taken from *Multiple Imputation and its Application*, by Carpenter & Kenward, published by Wiley (2013).

Outline

Aim: to explain when a complete records analysis is insufficient, and how multiple imputation (MI) can help

- When is a complete records analysis OK?
  - Understanding the bias with complete records analysis
  - Examples: Youth Cohort Study; Bed Sharing Study (missing covariates)
- Beyond complete records analysis: the missing data mechanism
- A principled approach to missing data
- Multiple imputation: an overview
- Applications
  - Youth Cohort Study
  - Bed Sharing Study
- Sensitivity analysis with MI
  - Example: Improving the quality of peer review (missing outcomes)
- Discussion

Outline

Issues raised by missing data

Missing data generally raise three direct concerns:

- difficulty with performing the intended analysis
- loss of power
- bias

But when should we be worried about them?

Multiple imputation is being increasingly used, due to its practicality and versatility [1, Sterne *et al* (2009)], [2, Klebanoff & Cole (2008)].

However this raises the following issues:

- are 'off the shelf' imputation models suitably consistent with the substantive model?
- are inferences robust to assumptions (which may be implicit)?

Complete records analysis

It is often said that a complete records analysis is generally biased unless data are Missing Completely At Random (MCAR)

This is true, but we can say more, especially if our substantive model is a regression.

Suppose we are interested in the regression of Y on multivariate X.

> If — **given the covariates** — the probability of a complete record does not depend on Y_i, then the complete records analysis is valid — though it may be quite inefficient.

Note: this statement does not depend on *where* the missing data occur.

Additional observation

The substantive model coefficients of the variables predicting the probability of a complete record (in conjunction with the response) will typically be most biased.

In other words it is the variables that are involved in the missingness mechanism that are important for determining the bias of the complete records analysis: these may not be the variables with missing data!

For logistic regression, only the substantive model coefficients of covariates predicting complete records will be biased [3, p. 32].

Consider a regression of Y on covariates X and Z. The following table summarises the bias:

Logistic regression: bias in complete records analysis

logistic regression of Y on X and Z.

Pr(compelte record) depends on:	Biased estimation of parameters using CR		
	constant	coef. of X	coef. of Z
X	No	No	No
Z	No	No	No
X, Z	No	No	No
Y	Yes	No	No
Y, X	Yes	Yes	No
Y, Z	Yes	No	Yes
Y, X, Z	Yes	Yes	Yes

For details, see [4].

Implications for exploratory analysis

We should look for:

1. the pattern of missingness;
2. key predictors of the probability of a complete record / missing values, and
3. key predictors of the underlying (missing) values.

> Note: the observed data cannot tell us definitively if complete records is valid.

In multiple imputation, variables of type (3) recover information. Variables of type (2) & (3) recover information and correct for bias.

Example: Youth Cohort Study (YCS)

The YCS is an ongoing UK government funded representative survey of pupils in England and Wales at school leaving age.

We consider data from pupils attending comprehensive schools from five cohorts, who reached the end of Year 11 (age 16+) in 1990, 1993, 1995, 1997, 1999.

Our analysis is illustrative, and relates GCSE score (range 0–84) to parental occupation, ethnicity, cohort and sex.

Variable name	Description
cohort	year of data collection: 1990, 93, 95, 97, 99
boy	indicator variable for boys
occupation	parental occupation, categorised as managerial, intermediate or working
ethnicity	categorised as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White

Missing data pattern in YCS

Pattern	GCSE score	occupation	ethnicity	No.	% of total
1	✔	✔	✔	55145	87%
2	✔	.	✔	6821	11%
3	.	✔	✔	697	1%
4	✔	.	.	592	1%

Predictors of complete records

Variable	Models				
	1	2	3	4	5
cohort '93	-0.085				-0.168
	(0.036)				(0.039)
cohort '95	0.044				-0.212
	(0.038)				(0.042)
cohort '97	0.178				-0.032
	(0.040)				(0.043)
cohort '99	0.135				-0.165
	(0.040)				(0.046)
boy		-0.053			0.079
		(0.024)			(0.026)
GCSE score			0.037		0.038
			(0.001)		(0.001)
Non-white				-1.723	-1.698
				(0.0288)	(0.031)
ROC	0.53	0.51	0.68	0.62	0.74

Implications

We see that parental occupation (reported by the pupil) is the variable with the greatest proportion of missing data.

We see that the response in the substantive model, GCSE score, is a key predictor of missing parental occupation, so a complete records analysis is likely to be biased.

Next we observe that ethnicity is the key covariate in the substantive model which predicts missing parental occupation. The greatest bias is likely to occur in coefficients for ethnicity.

Results

Variable	Complete Records $n = 54872$
Cohort90	reference
Cohort93	5.66 (0.20)
Cohort95	9.42 (0.22)
Cohort97	8.09 (0.21)
Cohort99	12.70 (0.22)
Boys	−3.44 (0.13)
Managerial	reference
Intermediate	−7.42 (0.15)
Working	−13.74 (0.17)
White	reference
Black	−5.61 (0.57)
Indian	3.58 (0.44)
Pakistani	−2.03 (0.58)
Bangladeshi	0.27 (1.04)
Other asian	5.52 (0.68)
Other	−0.25 (0.70)
Constant	39.66 (0.19)

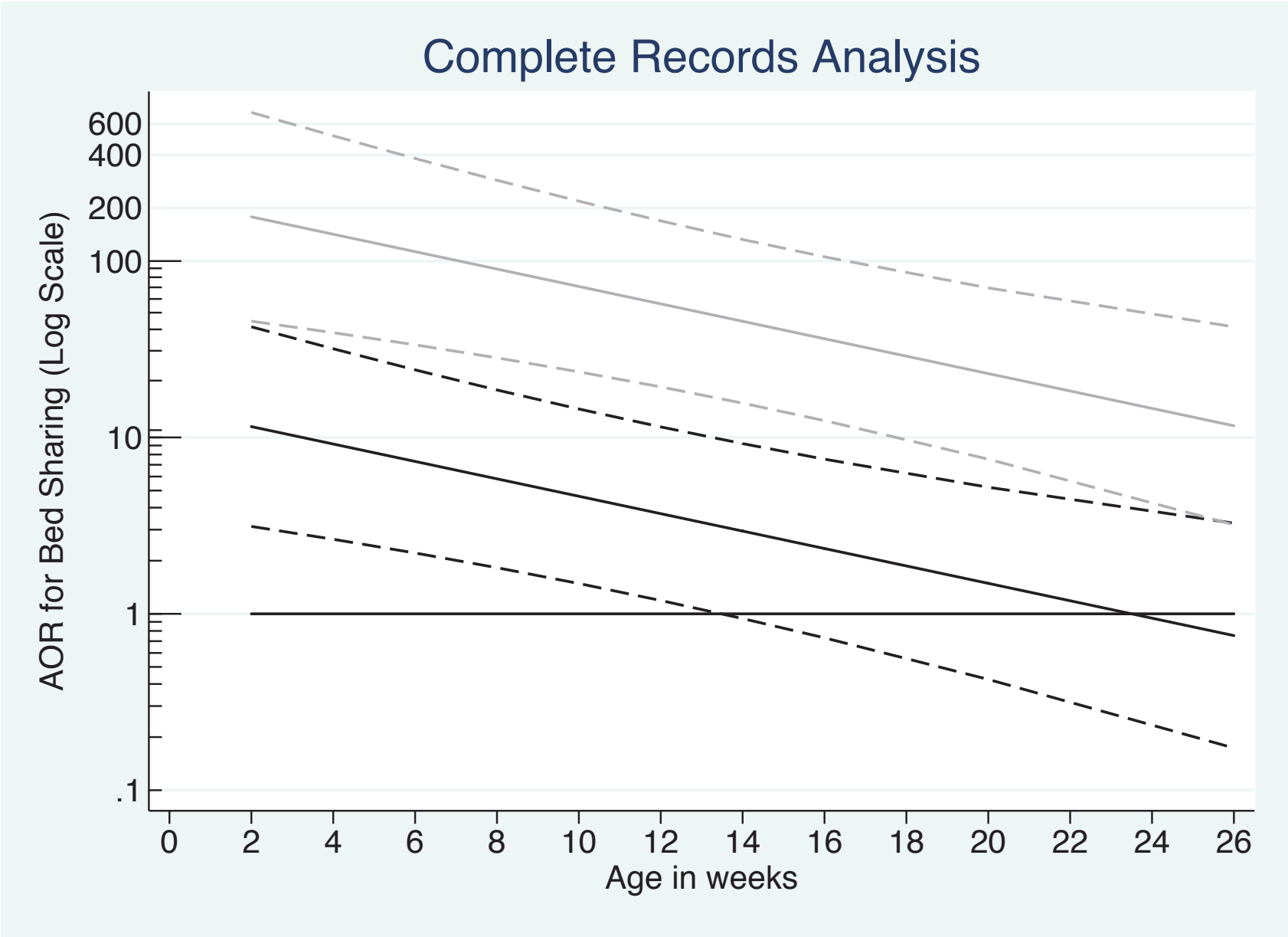
Case-control study of Sudden Infant Death Syndrome

Carpenter R. G. *et al* (2013) [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!

Results of Complete Records Analysis



Outline

Beyond Complete Records: key role of assumptions

CR will often be inefficient and sometimes biased. When we wish to go beyond this, we need to think about where the missing data occur, and why they are missing (i.e. the missingness mechanism)

The taxonomy of missingness mechanisms is:

Missing Completely At Random (MCAR)

The reason for the missing data is unrelated to our inferential question

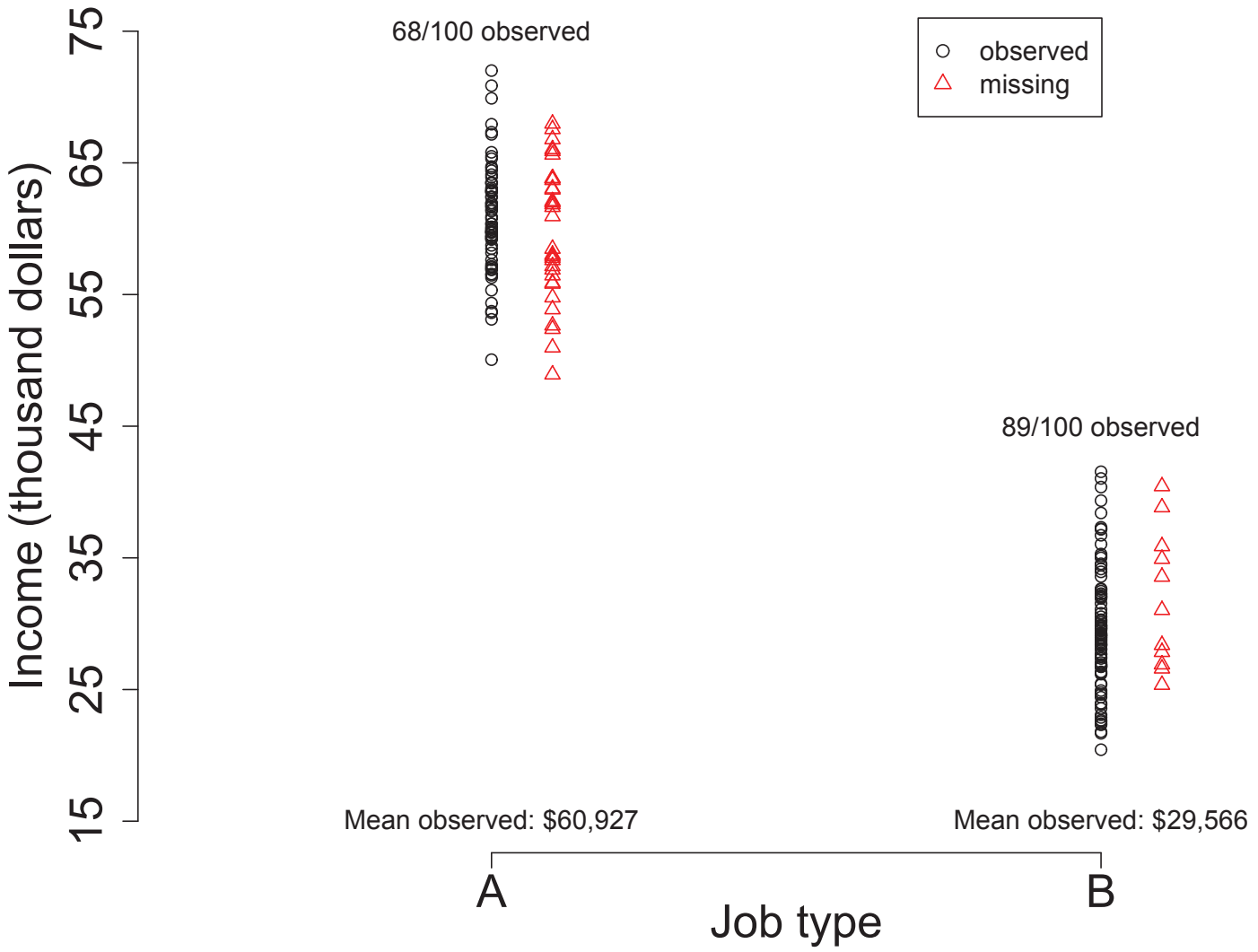
Missing At Random (MAR)

Suppose variables W are partially observed, and variables V fully observed. The distribution of $W|V$ is the same for observed and unobserved W.

Missing Not At Random (MNAR)

Consider the previous scenario. Now, the distribution of $W|V$ is *different* for observed and unobserved W.

Illustration: true mean income $45,000



Observed income: $\$43,149$.

MAR estimate: $\dfrac{100 \times 60,927 + 100 \times 29,566}{200} = \$45,246$

Outline

A principled approach to analysis when data are missing

Investigators discuss possible missingness mechanisms, say A–E, possibly informed by a (blind) review of the data, and consider their plausibility. Then

1. Under most plausible mechanism A, perform valid analysis, draw conclusions
2. Under similar mechanisms, B–C, perform valid analysis, draw conclusions
3. Under least plausible mechanisms, D–E, perform valid analysis, draw conclusions.

Investigators discuss the implications, and arrive at a valid interpretation of the study in the light of the possible mechanisms causing the missing data.

Why Multiple Imputation (MI)?

Typically, the primary assumption will be MAR. MI is a flexible way for analyzing data assuming MAR.

Assumes the analyst has selected the substantive model, then:

- Missing data are imputed K times from the predictive distribution of the missing values given the observed values, taking full account of the uncertainty.
- The substantive model is fitted to each imputed dataset, and the parameter estimate and its standard error recorded.
- The results are combined for final inference using Rubin's rules.

MI is attractive because

1. we can use our substantive model;
2. imputation, and Rubin's rules, are very general, and
3. we can include additional variables (not in our substantive model) to improve the imputations.

Outline

Consider regression of Y on X, Y partially observed, and assume Y MAR given X.

The regression of Y on X is validly estimated from the complete records.

Now consider the same regression, but suppose Y is fully observed but X is MAR given Y.

The complete records regression of Y on X will now be biased.

But the complete records of X on Y is not biased!

This suggests the following approach:

Intuition for multiple imputation

We continue to consider the regression of Y on X, with X values MAR given Y.

We can:

1. use the complete records to get a valid estimate of the regression of X on Y.
2. use the model in (1) to impute the missing values of X
3. Fit our substantive regression model of Y on X to the 'complete' data (where the missing X values are imputed)

The problem with this is that in (3) we do not distinguish between the observed and imputed data.

Intuition for multiple imputation (continued)

To address this, and reflect the fact that we can only ever know the *distribution* of the missing data, we repeat step (2), creating say K imputed data sets.

We then fit the model to each dataset, getting K sets of point estimates and standard errors.

These are then combined for final inference using Rubin's combination rules.

There is now a range of well established software packages for MI, so users do not typically need to engage in the technical details.

Some key considerations are as follows:

Considerations for multiple imputation

- Ensure all the variables in the substantive model are in the imputation model.
- Use at least 10 imputations; considerably more (~ 100) if you want your results to be reproducible at the quoted precision.
- Include auxilary (i.e. additional) variables that are predictive of missing values—these help recover lost information.
  If they are also predictive of data being missing, they can correct for bias.
  Variables that only predict whether data are missing are not useful.
- Be careful our imputation model is consistent with our model of interest:
  - multilevel structure is maintained in the imputation model: [3, ch. 9], [6],
  - Non-linear structure is reflected in the imputation model: [3, ch. 6–7], [7] [8]

Software

For cross-sectional data, most imputation is done using the 'Full Conditional Specification' algorithm.

This is available in `R` using the `mice` or `mi` software, in Stata using `mi impute chained` command, and in SAS PROC MI.

To preserve non-linear relationships, we have SMCFCS software in `R` and `Stata`. Details from `http://www.missingdata.org.uk` and [7].

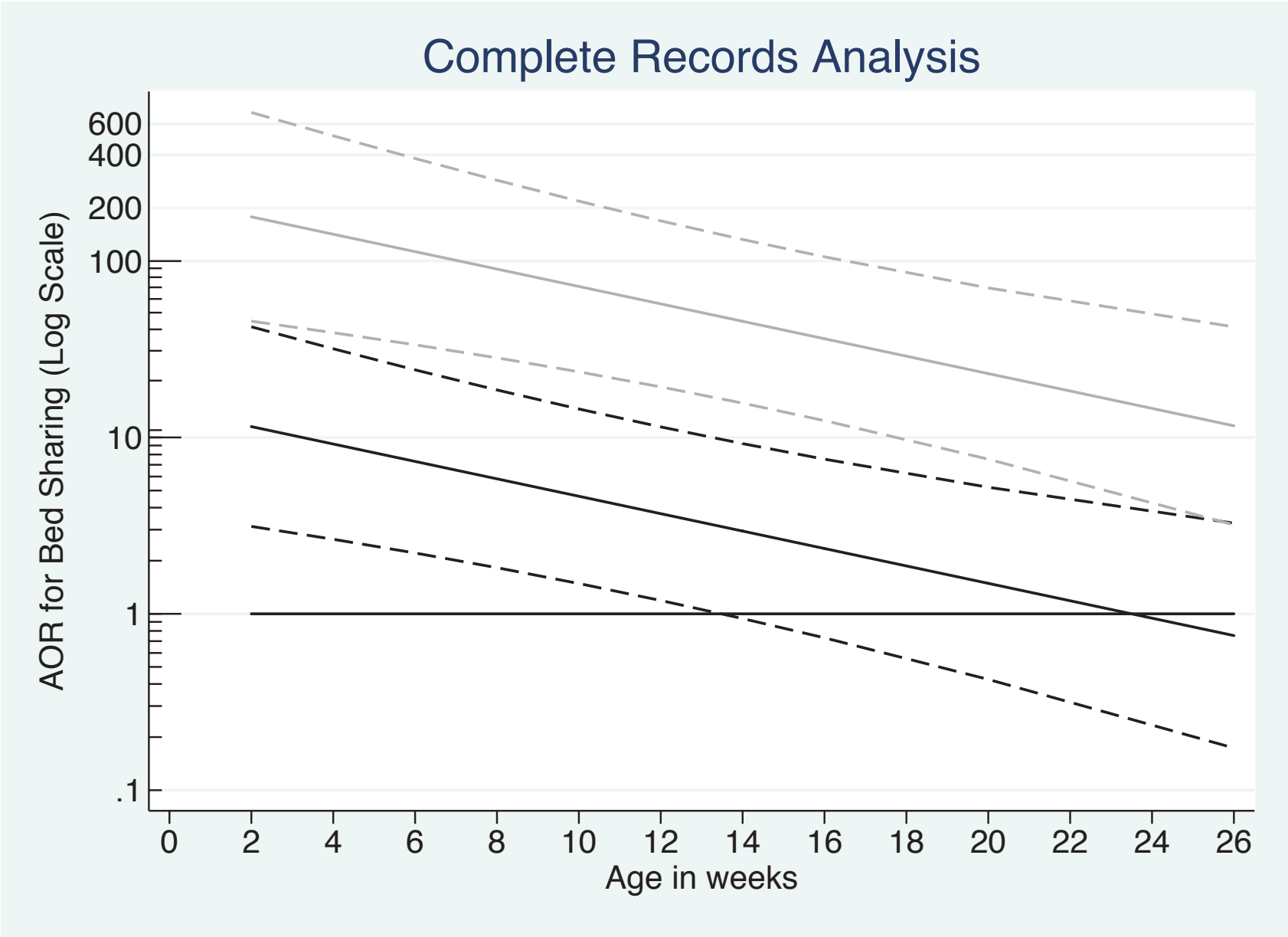
For hierarchical data, we have recently developed the `R` package `jomo`, available from CRAN; see [9].

Outline

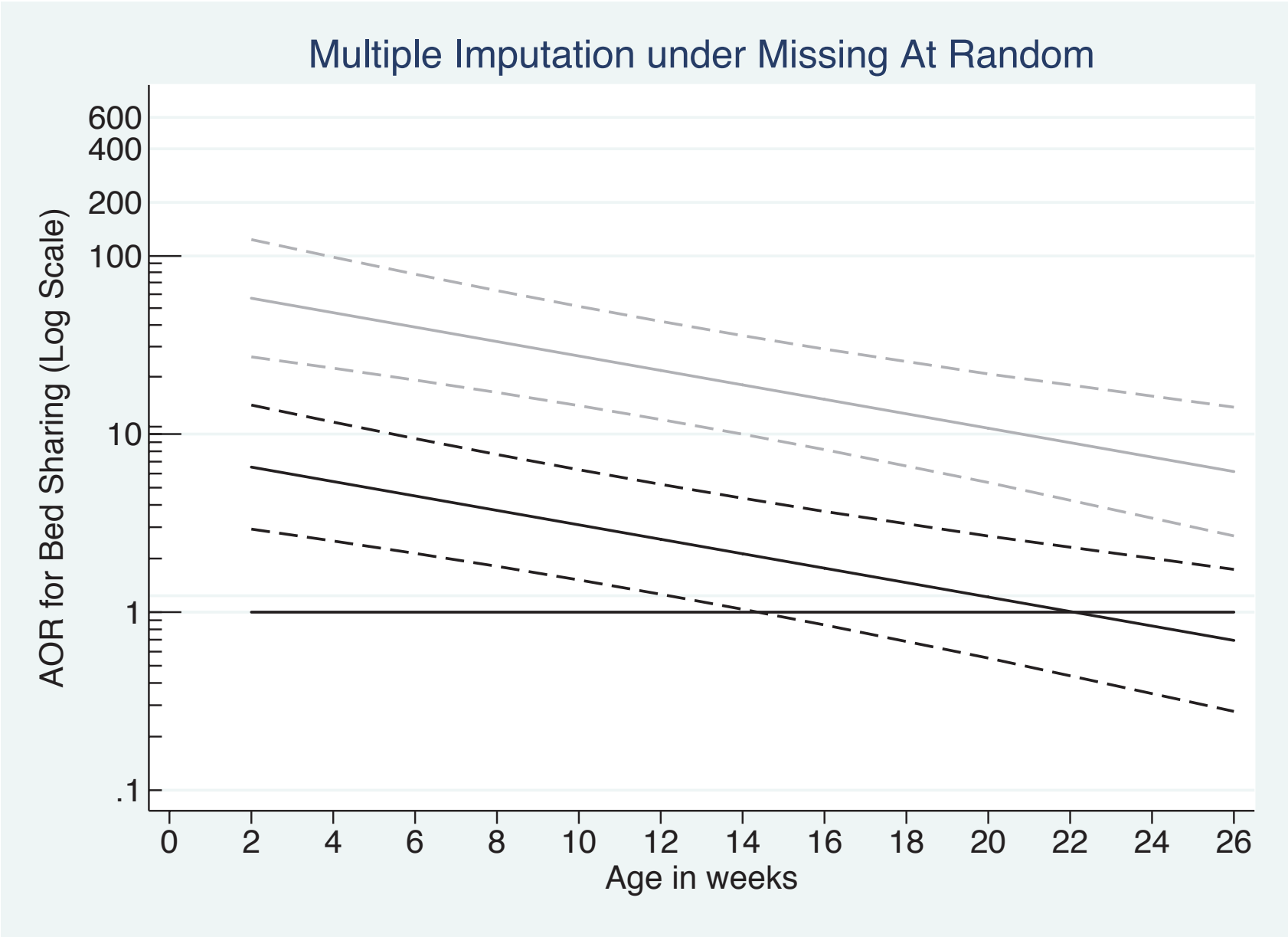
Youth Cohort Study: results of MI under MAR

Variable	Complete Records $n = 54872$	Multiple imputation $n = 62578$
Cohort90	reference	
Cohort93	5.66 (0.20)	5.44 (0.20)
Cohort95	9.42 (0.22)	9.21 (0.20)
Cohort97	8.09 (0.21)	8.03 (0.20)
Cohort99	12.70 (0.22)	12.91 (0.21)
Boys	−3.44 (0.13)	−3.35 (0.13)
Managerial	reference	
Intermediate	−7.42 (0.15)	-7.75 (0.16)
Working	−13.74 (0.17)	-14.32 (0.17)
White	reference	
Black	−5.61 (0.57)	−7.16 (0.51)
Indian	3.58 (0.44)	2.97 (0.42)
Pakistani	−2.03 (0.58)	−3.63 (0.47)
Bangladeshi	0.27 (1.04)	−3.20 (0.74)
Other asian	5.52 (0.68)	4.49 (0.63)
Other	−0.25 (0.70)	−1.32 (0.66)
Constant	39.66 (0.19)	39.09 (0.18)

Bed sharing study: results of MI assuming MAR



Bed sharing study: results of MI assuming MAR



Outline

Sensitivity analysis

Thus far our analyses have assumed data are MAR.

We now briefly explore how to use MI to do sensitivity analysis to this unverifiable assumption.

Recall from the income/job-type example that:

- MAR means the conditional distribution of income data given job is the same, whether or not income data are observed.
- MNAR means the distribution of income given job type is differs between observed and unobserved incomes

We have to specify this difference in order to proceed.

Example: peer review trial

Schroter *et al* (2004) [10] report a single blind randomised controlled trial among reviewers for a general medical journal. The aim was to investigate whether training improved the quality of peer review.

The study compared two different types of training (face to face training, or a self-taught package) with no training.

We restrict ourselves to the comparison between those randomised to the self-training package and no-training. Each participating reviewer was pre-randomised into their intervention group. Prior to any training, each was sent a baseline article to review (termed paper 1). If this was returned, then according to their randomised group, the reviewer was either (i) mailed a self-training package or (ii) received no further intervention.

...ctd

Two to three months later, participants who had completed their first review were sent a further article to review (paper 2).

The analysis excluded all participants who did not complete their first review: this was not expected to cause bias since these participants were unaware of their randomised allocation.

Analysis assuming MAR

Assume: outcome (score of three month review) MAR given baseline and intervention group.

Since the outcome is MAR given the covariates in the model, MI and full data analysis will agree[1]

Analysis	Est	SE	MI df	p-value	95% CI
Complete records, MAR	0.237	0.070	N/A	<0.001	(0.099, 0.376)
MAR, $K = 20$	0.245	0.073	302	<0.001	(0.102, 0.389)
MAR, $K = 10,000$	0.237	0.070	$\approx \infty$	<0.001	(0.099, 0.375)

[1] In general, for partially observed longitudinal outcome data, MI and the full data analysis will agree. [11, Ch. 4].

Is MAR plausible here?

Review Quality Index of paper 1 by whether or not paper 2 was reviewed.

		Group	
		Control	Postal
Returned review of paper 2	n	162	120
	mean	2.65	2.80
	SD	0.81	0.62
Did not return review of paper 2	n	11	46
	mean	3.02	2.55
	SD	0.50	0.75

Pattern mixture model: basic idea

Our analysis under MAR assumed — given baseline and intervention group — that the distribution of review quality index at 3 months was the same *whether or not it was observed*.

This suggests the following approach for sensitivity analysis *within a specific trial arm*:

- Ask experts what they think the *average* difference between observed and missing review quality is. Summarize the expert opinion into a distribution, say $\Delta \sim N(\mu, \sigma^2)$.
- Impute the data K times under MAR.
- For each imputed dataset, k, draw Δ_k and add Δ_k to each imputed value.
- Fit the substantive model to each imputed dataset, and use Rubin's Rules to combine the results for final inference.

Peer review study

Focusing on the baseline adjusted comparison of the self-taught training package with no training, the model of interest is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i, \quad e_i \overset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

where i indexes participant, $Y_i, X_{1,i}$ are the mean review quality index for paper 2 and paper 1 respectively and $X_{2,i}$ is an indicator for the self-training group.

Ideal for expert opinion:

Ideally, we want experts to tell us the likely departure from MAR *in each treatment arm*, and *how they are correlated*.

In other words, we want expert information on plausible values of the parameters of the (assumed normal) distribution

$$\begin{pmatrix} \Delta_N \\ \Delta_S \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_N \\ \mu_S \end{pmatrix}, \begin{pmatrix} \sigma_N^2 & \rho\sigma_S\sigma_S \\ \rho\sigma_S\sigma_S & \sigma_S^2 \end{pmatrix} \right],$$

where N, S index the no-intervention and self-training arms.

Specifying the departure from MAR

Carpenter and White [12] devised a questionnaire which was completed by 2 investigators and 20 editors and other staff at the *British Medical Journal*. The questionnaire was designed to elicit the experts' prior belief about the de facto difference between the average missing and average observed review quality index.

They showed that it was reasonable to pool information from the experts.

The resulting distribution is negatively skewed, with mean -0.21 and SD 0.46 (on the review quality index scale).

Unfortunately, it was not possible to collect information on how this differed by arm, or the correlation across arms, ρ.

Simplifying assumption

Suppose we denote by (Δ_N, Δ_S) draws from the distribution of the mean difference in review quality between observed and unobserved reviews, in respectively the no-intervention and self-training trial arms. We adopt a bivariate normal model approximation to the prior:

$$\begin{pmatrix} \Delta_N \\ \Delta_S \end{pmatrix} \sim N\left[\begin{pmatrix} -0.21 \\ -0.21 \end{pmatrix}, 0.46^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

As it was not possible to elicit a prior on ρ from the experts, we therefore analyse the data with $\rho = 0, 0.5, 1$ below.

Algorithm

Given a draw (Δ_N, Δ_S) from this distribution the model is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i \quad \text{if } Y_i \text{ observed,}$$
$$Y_i = (\beta_0 + \Delta_N) + \beta_1 X_{1,i} + (\beta_2 + \Delta_S - \Delta_N) X_{2,i} + e_i \quad \text{if } Y_i \text{ unobserved,}$$

Thus the mean review quality, relative to that in the observed data, is changed by Δ_N in the no-intervention arm and Δ_S in the self-taught arm.

...ctd

Following the general approach for estimating pattern mixture models [3, Ch 10], we proceed as followsS:

1. Fit the imputation model to the observed data and draw from the posterior distribution of the parameters $\eta = (\beta_0, \beta_1, \beta_2, \sigma^2)$.
2. Draw (Δ_N, Δ_S).
3. Using the draws obtained in steps 1 and 2, impute the missing Y_i.

Steps 1–3 are repeated to create K imputed data sets. Then we fit the model of interest to each imputed data set and apply Rubin's rules for inference.

Results

Analysis	Est	SE	MI df	p-value	95% CI
Complete records, MAR	0.237	0.070	N/A	<0.001	(0.099, 0.376)
MAR, $K = 20$	0.245	0.073	302	<0.001	(0.102, 0.389)
MAR, $K = 10,000$	0.237	0.070	$\approx \infty$	<0.001	(0.099, 0.375)
MNAR, $\rho = 0$, $K = 20$	0.209	0.178	27	0.25	$(-0.158, 0.575)$
MNAR, $\rho = 0.5$, $K = 20$	0.205	0.167	27	0.23	$(-0.141, 0.234)$
MNAR, $\rho = 1$, $K = 20$	0.213	0.134	34	0.12	$(-0.059, 0.486)$

Comments

- The approach is accessible to experts, and we can therefore frame questionnaires to elicit information about relevant departures from MAR;
- As usual, we should be careful to create proper imputations;
- We can summarise the imputed variables to check the process is working as we expect;
- We fit the substantive model to each imputed data set, and use Rubin's rules to summarise the results for final inference;
- We can use the imputed data to explore the implied selection mechanism if desired.
- Sensitivity analysis does not have to be local.

Challenges

- The main challenge with sensitivity analyses is not technical, but rather lies in eliciting meaningful information from experts.
- We have developed an intuitive web-based elicitation tool, which is described in a forthcoming article in *Clinical Trials* [13]
- Nevertheless, in longitudinal settings elicitation is very hard. Further, it can result in a substantial loss of information compared to the MAR analysis.
- We have developed an alternative, *reference based imputation*, which 'borrows' the information that is needed for the MNAR distribution from other arms [14].
- In recent work [15] we have shown this approach is *information anchored*.

Outline

Summary and discussion

- Design to minimize missing data.
- Overall strategy: we need to look carefully at the patterns of missingness, and the predictors of missing values *alongside our substantive model*, to understand whether a complete records analysis is adequate, and how it might differ from an analysis under MAR.
- MI gives a very flexible, practical, way of carrying out analysis under MAR.
- The most common pitfall with MI is that the imputation model does not reflect the variables and structure in the substantive model [1, 16]
- As the MAR assumption is untestable, we should consider sensitivity analysis.
- A pattern mixture approach, using expert opinion, can be readily implemented using MI; other approaches are reference based imputation, and best/worst scenarios.

References I

[1] 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.

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

[3] J R Carpenter and M G Kenward.
Multiple Imputation and its Application.
Chichester, Wiley, 2013.

[4] 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.

[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] H Goldstein, J Carpenter, M Kenward, and K Levin.
Multilevel models with multivariate mixed response types.
Statistical Modelling, 9:173–197, 2009.

[7] 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.

[8] H Goldstein, J R Carpenter, and W Browne.
Fitting multilevel multivariate models with missing data in responses and covariates, which may include interactions and non-linear terms.
Journal of the Royal Statistical Society, Series A, 177:553–564, 2014.

References II

[9] M. Quartagno and J. R. Carpenter.
Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates.
Statistics in Medicine, 35:2938–2954, 2015.

[10] S Schroter, N Black, S Evans, J Carpenter, F Godlee, and R Smith.
Effects of training on quality of peer review: randomised controlled trial.
British Medical Journal, 328:673–675, 2004.

[11] James R Carpenter and Michael G Kenward.
Missing data in clinical trials — a practical guide.
Birmingham: National Health Service Co-ordinating Centre for Research Methodology, 2008.

[12] I White, J Carpenter, Stephen Evans, and Sara Schroter.
Eliciting and using expert opinions about non-response bias in randomised controlled trials.
Clinical Trials, 4:125–139, 2007.

[13] Alexina J Mason, Manuel Gomes, Richard Grieve, Pinar Ulug, Janet T Powell, and James R Carpenter.
Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the improve trial.
Clinical Trials (in press), XX:XXX–XXX, 2017.

[14] James R Carpenter, James H Roger, and Michael G Kenward.
Analysis of longitudinal trials with protocol deviation:— a framework for relevant accessible assumptions and inference via multiple imputation.
Journal of Biopharmaceutical Statistics, 23:1352–1371, 2013.

[15] S Cro, J R Carpenter, and M G Kenward.
Information anchored sensitivity analysis, 2016.

[16] 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.