Sensitivity analysis within multiple imputation framework using delta-adjustment: Application to Longitudinal Study of Australian Children

Panteha Hayati Rezvan — The University of California, Los Angeles, USA and The University of Melbourne, Australia
Katherine J. Lee — Murdoch Children's Research Institute, Australia and The University of Melbourne, Australia
Julie A. Simpson — The University of Melbourne, Australia
julieas@unimelb.edu.au

(Received February 2018 Revised June 2018) http://dx.doi.org/10.14301/llcs.v9i3.503

Abstract

Multiple imputation (MI) is a powerful statistical method for handling missing data. Standard implementations of MI are valid under the unverifiable assumption of missing at random (MAR), which is often implausible in practice. The delta-adjustment method, implemented within the MI framework, can be used to perform sensitivity analyses that assess the impact of departures from the MAR assumption on the final inference. This method requires specification of unknown sensitivity parameter(s) (termed as delta(s)).

We illustrate the application of the delta-adjustment method using data from the Longitudinal Study of Australian Children, where the epidemiological question is to estimate the association between exposure to maternal emotional distress at age 4–5 years and total (social, emotional, and behavioural) difficulties at age 8–9 years. We elicited the sensitivity parameters for the outcome (Y) and exposure (X) variables from a panel of experts. The elicited quantile judgements from each expert were converted into a suitable parametric probability distribution and combined using the linear pooling method. We then applied MI under MAR followed by sensitivity analyses under missing not at random (MNAR) using the delta-adjustment method. We present results from sensitivity analyses that used different percentile values of the pooled distributions for the delta parameters for Y and X, and demonstrate that twofold increases in the magnitude of the association between maternal distress and total difficulties are only observed for large departures from MAR.

Keywords

Missing data; multiple imputation; missing not at random; sensitivity analysis; delta-adjustment method; elicitation

Background

Missing data commonly occur in longitudinal studies with multiple waves of data collection over long periods of follow-up (Burton & Altman, 2004; Karahalios, Baglietto, Carlin, English, & Simpson, 2012; Sterne et al., 2009). The simplest and widely used approach for handling missing data in these studies is to omit participants with any missing observations from the statistical analysis (known as a complete case analysis), which can greatly reduce the sample size, resulting in loss of precision and statistical power (i.e. inefficiency). More importantly, complete case analyses may give rise to bias if participants with complete records do not represent the entire study sample (and if the statistical analyses do not adjust for predictors of missingness).

An alternative statistical approach for dealing with missing data is multiple imputation (MI) (Rubin, 1987), a flexible and sophisticated method, which has grown in popularity among researchers (Hayati Rezvan, Lee, & Simpson, 2015; Mackinnon, 2010; Manly & Wells, 2015). MI involves two stages. First, missing data are imputed by sampling multiple times (denoted *m*) from an imputation model based on the observed data to create multiple completed (observed and imputed values) datasets. Second, the completed datasets are analysed separately using the statistical method required for the target analysis, resulting in *m* sets of parameter estimates and associated variances. The estimates obtained from each completed dataset are then combined using special formulae, known as Rubin's rules (Rubin, 1987), to obtain one overall MI estimate and corresponding variance, which accounts for the within- and between- imputation variability. Unlike a complete case analysis, MI enables all participants to be included in the target analysis by replacing the missing data with plausible values, thereby, potentially improving efficiency and reducing the bias obtained from a complete case analysis (Little & Rubin, 2002; White & Carlin, 2010). Of note, this bias correction achieved with MI is only obtained if all of the variables associated with non-response are included in the imputation model.

The standard implementation of MI is typically valid under the assumption that data are missing at random (i.e. the probability of a value being missing in any variable depends on the observed data and is conditionally independent of any unobserved values (MAR)). However, in many practical settings, the plausibility of the MAR assumption is questionable, and it is more likely that the probability of data being missing depends on the unobserved values (i.e. data are missing not at random (MNAR)). In such cases, performing the standard MI procedure may not capture the true underlying missing data mechanism, and may lead to biased results (White & Carlin, 2010). Since it is not possible to verify whether the missingness depends on the missing data, it is desirable to assess the robustness of the MI results for the target analysis by conducting sensitivity analyses that explore the effects of plausible departures from the MAR assumption.

The necessity of performing such sensitivity analyses within the MI framework has been emphasised in a number of guidelines (Burzykowski et al., 2010; Little et al., 2012; Sterne et al., 2009; White, Horton, Carpenter, & Pocock, 2011) and reviews (Hayati Rezvan, Lee & Simpson, 2015; Mackinnon, 2010), and was also recommended in the 2010 report produced by the National Research Council (NRC) expert panel on handling missing data in clinical trials (National Research Council, 2010). In general, there are two frameworks for modelling data that are MNAR, both of which are based on different factorisations of the joint distribution of the data and the mechanism leading to missing data. These two broad classes of models are the selection model and pattern-mixture model (Diggle & Kenward, 1994; Hogan & Laird, 1997; Kenward & Molenberghs, 1999; Little, 1993; Little, 1995).

In the context of MI, there are a number of approaches that have been proposed in the statistical literature for performing sensitivity analyses to the MAR assumption under these two frameworks (Carpenter, Kenward, & White, 2007; Siddique, Harel, & Crespi, 2012; Siddique, Harel, Crespi, & Hedeker, 2014; van Buuren, Boshuizen, & Boshuizen, 1999). A selection-model-based weighting approach, which was proposed by Carpenter, Kenward and White (2007), involves re-weighting the parameter estimates obtained from analyses of the imputed values under the assumption of MAR in such a way that reflect the MNAR mechanism (i.e. a weighted version of Rubin's rules). This approach was proposed as an approximate method for performing sensitivity analyses following MI, for a single variable with missingness that is weakly MNAR (Carpenter et al., 2007; Carpenter, Rücker, & Schwarzer, 2011;

Carpenter & Kenward, 2013). The performance of the weighting approach was evaluated through a series of simulation experiments, where it was shown that the method does not recover unbiased estimates even when the number of imputations used is large (Hayati Rezvan, White, Lee, Carlin, & Simpson, 2015). Possible reasons for the failure of the method were explained in detail, and the method was not recommended for performing sensitivity analyses to assess the effects of departure from the MAR assumption.

An alternative approach to sensitivity analysis in the context of MI is the delta-adjustment method, which was proposed by Rubin (1987) within the pattern-mixture modelling framework, where the MAR imputed values are modified to reflect an assumed MNAR mechanism. Practical examples of this method in different settings are provided by a number of authors (Carpenter & Kenward, 2007; Carpenter & Kenward, 2013; Leacy, Floyd, Yates, & White, 2017; Liublinska & Rubin, 2014; Moreno-Betancur & Chavance, 2013; Ratitch, O'Kelly, & Tosiello, 2013; van Buuren et al., 1999; van Buuren, 2012). For continuous incomplete variables, the method often proceeds by adding some fixed constant to the imputed values obtained under MAR from a standard MI procedure. For categorical incomplete variables, the missing values are drawn from an imputation model assuming MNAR, which proceeds by adding offsets to the linear predictors of the variables with missing observations. The modified imputations are then analysed and the resulting estimates and variances for each completed dataset combined in the usual way (i.e. Rubin's rules). "Multiple-model multiple imputation", developed by Siddique et al. (Siddique et al., 2012; Siddique et al., 2014) within the pattern-mixture modelling framework, is an alternative approach which takes into account the uncertainty in the missingness mechanism by specifying a distribution for the value of the offset, so that imputations are generated from multiple imputation models with different offsets. The resulting estimates are combined using nested imputation combining rules (Shen, 2000) to obtain final estimates.

Despite recent emphasis on the importance of performing sensitivity analyses following MI to assess the influence of plausible departures from MAR, the approaches outlined earlier have not been widely adopted in practice, due to lack of explicit guidance for conducting such sensitivity analyses. The majority of pattern-mixture model-based methods have been implemented in the context of clinical trials, generally with missing data in the outcome variable only. Recently, a number of statistical packages (SAS 9.4, Proc MI (Yuan, 2014) and R, SensMice (Resseguier, 2010; Resseguier, Giorgi, & Paoletti, 2011; Resseguier, Verdoux, Giorgi, Clavel-Chapelon, & Paoletti, 2013) have been developed to impute missing values in multiple variables under MNAR using the delta-adjustment method (with the strong assumption that the MNAR mechanism is independent for multiple incomplete variables), however, they have not been widely used in practice.

One of the most important aspects of conducting sensitivity analyses to the MAR assumption using any of the above methods is selecting one or more sensitivity parameters that represent plausible departures from MAR. These parameters are generally unidentifiable values and must be specified by one or more experts who have relevant knowledge of the subject matter. For applications of the methods described above, in general, extreme values or a range of plausible values for the sensitivity parameters have been selected by the analyst instead of carefully elicited from content experts. Although there is a vast amount of literature on the different approaches for eliciting uncertain values from experts' knowledge (Kadane & Wolfson, 1998; O'Hagan, 2006; White, 2015; White, Carpenter, Evans, & Schroter, 2007), there is limited research on eliciting unknown sensitivity parameters in practice for sensitivity analyses within the MI framework.

The outline for the rest of this paper is as follows. We begin with an overview of pattern-mixture models and explain the delta-adjustment method for performing sensitivity analyses within the MI framework. We also describe the process of obtaining prior information regarding the sensitivity parameters from subject-matter experts (i.e. elicitation). We then describe the Longitudinal Study of Australian Children (LSAC) and the case study that motivated this work. We address how we elicited the required information regarding the unknown values of sensitivity parameters from a panel of three experts, and then present results from the LSAC case study where the delta-adjustment method was implemented for imputing missing data assuming MNAR for the outcome and

exposure of interest using the elicited sensitivity parameters. Finally, we conclude with a discussion of the delta-adjustment method.

Methods

Pattern-mixture models

In brief, within the pattern-mixture modelling framework, the incomplete data are modelled conditional on the missingness mechanism (i.e. the response pattern). Let Y be a partially observed variable, with Y^{obs} and Y^{mis} representing its observed and missing components, respectively. Suppose also that X is a fully observed variable and R_y is the usual indicator of missingness, taking value of 0 or 1, depending on whether Y is missing or observed. Under the pattern-mixture framework, the joint distribution of the complete data and the missing data mechanism, $f(Y, R_y|\, X)$, is factorised as:

$$f\left(Y^{obs}, Y^{mis}, R_y|X\right) = f\left(Y^{obs}, Y^{mis}|R_y, X\right) f(R_y|X) \quad (1)$$

i.e. the distribution of complete data (Y^{obs}, Y^{mis}) conditional on the missingness mechanism (R_y), and the marginal distribution of the missing data mechanism. Under this factorisation, there is a different joint distribution of the observed and missing data for each missing data pattern. Equation (1) can be further decomposed to:

$$f\left(Y^{obs}, Y^{mis}, R_y|X\right) = f\left(\, Y^{mis}|Y^{obs}, R_y, X\right) f\left(Y^{obs}\middle|R_y, X\right) f\left(R_y\middle|X\right) \quad (2)$$

which enables researchers to distinguish the conditional distribution of the missing data given the observed data from the distribution of the observed data. Under MAR, $f\left(\, Y^{mis}|Y^{obs}, R_y, X\right) = f\left(\, Y^{mis}|Y^{obs}, X\right)$, that is, $f\left(\, Y^{mis}|Y^{obs}, X, R_y = 1\right) = f\left(\, Y^{mis}|Y^{obs}, X, R_y = 0\right)$, implying that the imputed values are drawn from the posterior distribution of the observed data, $f\left(\, Y^{mis}|Y^{obs}, X, R_y = 1\right)$. However, under MNAR, $f\left(\, Y^{mis}|Y^{obs}, X, R_y = 1\right) \neq f\left(\, Y^{mis}|Y^{obs}, X, R_y = 0\right)$, indicating that the conditional distribution of the missing data given the observed data will differ with missingness patterns represented by R_y.

In order to fit these models, participants are initially divided into different groups based on their missing data pattern, and then the grouping variable is used to model the effect of missing data patterns on outcome(s). Using these models, the overall estimate is then obtained by averaging the outcome over the missing data patterns. A simple pattern-mixture model could be in the form of a linear function as below:

$$E\left(Y|R_y, X\right) = \varphi_0 + \varphi_1 X + \delta(1 - R_y) \quad (3)$$

where Y is a continuous variable with missing data, X is a fully observed covariate, and R_y is a missingness indicator of Y as described above. δ in equation (3) is known as the sensitivity parameter and quantifies the degree of departure from the MAR assumption, where $\delta = 0$ represents the MAR mechanism as there is no dependency between the missingness mechanism and the missing values in Y. This parameter shows the mean difference (shift) of the partially observed variable Y between the missing and observed data. If $\delta > 0$ (or $\delta < 0$), the mean of Y among non-respondents is δ units higher (or lower) than respondents, given a fixed value of X. If Y is a partially observed binary variable (0/1), then the pattern-mixture model could be in the form of the following logit function:

$$logit[Pr\left(Y = 1|R_y, X\right)] = \gamma_0 + \gamma_1 X + \delta(1 - R_y) \quad (4)$$

where δ represents the difference in the $\log_e$ odds of Y=1 between non-respondents and respondents. Of note, equations (3) and (4) represent models that allow for shifts of the intercepts, φ_0 and γ_0, respectively. In realistic scenarios, where the covariate X has different impacts on the outcome Y among different missing data patterns, it is more relevant that the pattern-mixture models allow for the association between X and Y to differ between non-respondents and respondents (i.e. additional sensitivity parameters are required to allow for shifts of the regression coefficients φ_1 and γ_1).

Implementation of the delta-adjustment method

For continuous variables with MNAR missingness, the imputation procedure using the delta-adjustment method proceeds as follows:

(i) Missing values in a partially observed variable are imputed using a standard MI procedure under MAR. Returning to the example explained earlier for an incomplete continuous variable, first the point estimates of φ_0, φ_1, and σ^2 are obtained by fitting the imputation model $Y = \varphi_0 + \varphi_1 X + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, to the observed data (i.e. $R_y = 1$) to characterise the joint posterior distribution of φ_0, φ_1 and σ^2. Then, imputed values are drawn from the posterior distribution of the parameters.

(ii) The imputed values are shifted by adding some fixed value, δ, which is obtained from content experts, to reflect the MNAR mechanism.

(iii) The completed (observed plus imputed values) datasets are analysed separately using standard statistical methods, and the resulting point estimates and standard errors are combined using Rubin's rules to give a single MNAR estimate.

The procedure for imputing MNAR missing data in categorical variables proceeds similarly but in step (i) an offset of δ is included in the univariate imputation model so that the missing values are drawn from an imputation model assuming MNAR, and thus step (ii) is omitted. In the example described previously for an incomplete binary variable, missing values in Y variable are imputed using the MNAR imputation model, $logit\ [Pr(Y = 1|X)] = \gamma_0 + \gamma_1 X + \delta(1 - R_y)$, instead of the MAR imputation model, $logit\ [Pr(Y = 1|X)] = \gamma_0 + \gamma_1 X$.

Specification of the sensitivity parameters, δ

In practice the sensitivity parameters are unidentifiable values and cannot be estimated from the observed data since, by definition, they depend on the missing data. The only principled approach to determine plausible values of the unknown sensitivity parameters is to extract them from experts who have relevant knowledge about the subject matter. However, elicitation of the sensitivity parameters is often a challenging task in practice, and alternatively some investigators prefer to conduct a tipping-point analysis (Yan, Lee, & Li, 2009) in which sensitivity parameters are varied across a large range of values.

This section briefly describes the process of elicitation (O'Hagan, 2006), which, by definition, aims to formulate the expert's beliefs into a probability distribution for the parameter of interest, in this case the sensitivity parameter of the pattern-mixture model, δ. Briefly, as described by Garthwaite, Kadane and O'Hagan (2005), a good elicitation depends on the quality of the four important stages listed below.

Stage 1: Setting up the elicitation problem by identifying one or more unknown quantities for which expert judgement is required. It is common practice to elicit the opinion from a number of experts using a questionnaire and/or a face-to-face interview. Although, using a questionnaire has the advantage of the same structured questions being given to each expert, designing a questionnaire that will be clear to all experts is not a trivial task. Thus, for elicitation a face-to-face interview is considered the optimal approach (White, 2015).

Stage 2: Eliciting experts' opinion about the unknown quantities. Since the aim of the elicitation is to present experts' judgements in a form of a probability distribution, they are often asked to suggest values for suitable summary statistics of the associated parameters of that distribution. Usually these summaries are probabilities (e.g. single probabilities or quantiles), measures of location (e.g. mean, median or mode), measures of spread (e.g. usually variance or standard deviation), etc.

Stage 3: Fitting an appropriate probability distribution to those elicited summaries. In situations where elicitation is conducted for each expert individually, the resulting distributions must be combined into a single probability distribution (see O'Hagan (2006) and Garthwaite et al. (2005) for more details). Several methods have been proposed in the literature for combining distributions from multiple experts into a single distribution (O'Hagan, 2006). The simplest aggregation method is linear opinion pooling (Cooke, 1991; Genest & Zidek, 1986; Mcconway, 1981), which is a weighted average of each expert's distribution. Under this approach, the investigator must decide whether all experts should be weighted equally, or whether to assign larger weights to experts whose distributions are believed to be more accurate. There are number of software packages available to help elicitation in practice, which are mostly designed within the "Sheffield Elicitation Framework" (O'Hagan, 2013); examples include SHELF package (Oakley, 2017) in R software (R Development Core Team, 2005), Elicitator software (Kynn, 2005.), UncertWeb-The Elicitator (Bastin et al., 2013), and MATCH Uncertainty Elicitation Tool (Morris, Oakley, & Crowe, 2014).

Stage 4: Evaluating the adequacy of the elicitation process by providing feedback to the experts. Once the desirable summaries have been elicited from experts and an adequate probability distribution has been specified by the investigator, the experts should be informed about the implications of that fitted distribution (e.g. provided with visual feedback) and asked whether the estimated quantities adequately represent their opinions. In cases where experts believe that the fitted distribution does not express their opinions,

repeating the procedure until the probability distribution accurately reflects their beliefs is required. Further, sensitivity analyses considering a number of alternative probability distributions can be conducted to explore the impacts of experts' uncertainty.

In the 'Results' section, where the application of the delta-adjustment method is presented using the LSAC case study, we explain how and what quantities we elicited for obtaining a probability distribution for the sensitivity parameters.

Motivation: The LSAC case study

The LSAC is a national longitudinal study of childhood development in Australia. Data have been collected on a range of important aspects of childhood development such as wellbeing (physical and mental health), education and schooling, and social, emotional, and cognitive functioning across childhood. Details of the LSAC dataset have been published elsewhere (Australian Institute of Family Studies, 2015). In brief, LSAC consists of two cohorts of children: the birth cohort, which includes 5107 infants aged 0–1 years, and the kindergarten cohort, which includes 4983 children aged 4–5 years. The primary objective of the LSAC case study presented in this paper, which only uses data from the kindergarten cohort, is to estimate the association between exposure to maternal emotional distress at age 4–5 years (i.e. pre–school children) and total (social, emotional, and behavioural) difficulties at age 8–9 years, controlling for potential confounders. This case study was motivated by previously published research (see Bayer et al. (2011)). Since our aim was to use the present case study to evaluate the delta-adjustment method rather than make any substantive claims about the LSAC data, we modified the analysis from that used in the original article (Bayer et al., 2011) in order to keep the target analysis and imputation models simple.

The outcome variable of interest is total social, emotional and behavioural difficulties of children aged 8–9 years as assessed by the total score on Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997) at wave 3 of the data collection. This was also measured at wave 1 (age 4–5 years). The SDQ total score is summed over four subscales relating to conduct, hyperactivity, peer, and emotional problems. Each of these subscales is averaged over five items, where 1 is the minimum score for an answer 'Not true' and 3 is the maximum score for an answer 'Certainly true'. This score is then rescaled to be an integer between 0 and 10, giving a total score that can range from 0 to 40. A lower SDQ total score corresponds to a better overall behaviour status. The SDQ total score in the present study ranged from 0 to 35 (25th, 50th, and 75th percentiles were 4, 6, and 10, respectively) and was moderately right-skewed. While a number of approaches have been proposed in the MI literature for handling non-normally distributed variables, including transforming the skewed variables prior to imputation or predictive mean matching, we imputed missing values of the SDQ total score on the raw scale since it was modelled on the raw scale in the analysis model (Lee & Carlin, 2017; von Hippel, 2013).

The primary exposure of interest is maternal emotional distress at age 4–5 years as measured by the Kessler-6 (K-6) (Kessler et al., 2010) depression scale at wave 1. In LSAC, this measure is calculated as the mean of six items assessing mother's anxiety and depression symptoms in the most recent four weeks. For each item, the minimum score of 1 represents 'All of the time', and the maximum score of 5 indicates 'None of the time', with higher averaged scores representing better mother's mental health status. The distribution of maternal emotional distress was left-skewed, with the bulk of the observations at the higher end of the scale (i.e. 4–5). For the purpose of illustrating the missing data problem in the current case study, we dichotomised the variable such that 1 represents a category with 'Probable serious mental illness' (i.e. average K-6 score less than 4), and 0 represents a category with 'No probable serious mental illness' (i.e. average K-6 score greater or equal to 4).

The target analysis model is a multivariable linear regression of SDQ total score at 8–9 years on maternal emotional distress at 4–5 years, controlling for ten potential confounders measured at wave 1. The confounders selected *a priori* were: SDQ total score (possible range 0–40), child physical functioning score based on Paediatric Quality of Life inventory (PEDS QL) (possible range 0–100) (Varni, 2006), mother's age (years), consistent parenting score (possible range 1–5), family financial hardship score (possible range 0–6), sex of child (male/female), whether child has a sibling in the home (yes/no), mother current cigarette smoker (yes/no), mother consumes >2 standard drinks of

alcohol daily (yes/no), and mother completed high school (yes/no).

Table 1 presents the details of all the variables used in the statistical analysis along with the frequency of missing data. The percentage of missing observations in the dataset for a given variable ranged from 0% to 23.8% and the missing data pattern was non-monotonic. Out of the twelve variables included in the target analysis model, only sex of child and whether child has a sibling in the home were completely observed. While the SDQ total score at 8–9 years (outcome at wave 3) was the variable with the highest amount of missing observations (23.8%), data on maternal emotional distress at 4–5 years (exposure at wave 1) were also not available for 16.4% of the participants. Sixty-five percent of the sample (3244 participants) had complete data on the outcome and all of the covariates included in the target analysis model are listed in table 1. Among participants with observed SDQ total score at 8–9 years (i.e. 3798 out of 4331 participants at wave 3), the average total score was 7.5. Further, among mothers with observed scores for maternal emotional distress at 4–5 years, 20.9% had probable serious mental illness.

It is highly likely that the reason for missingness of SDQ total score is related to the underlying child behavioural status (i.e. MNAR). In a similar fashion, missing data for maternal emotional distress is of particular concern because the reason for a mother not completing certain survey questions or being unwilling to participate in a face-to-face interview is probably related to her underlying mental health status.

Table 1. Description of variables used in the LSAC case study analysis (*n*=4983).

Variable description	Grouping /range	Number missing (%)
Binary variables		
Sex of child	Male / Female	0(0)
Child has a sibling in the home	Yes / No	0(0)
Mother current cigarette smoker	Yes / No	852(17.1)
Mother alcohol consumption†	Yes / No	966(19.2)
Mother completed high school	Yes / No	44(0.9)
Mother emotional distress‡*	(<4) / (≥4)‡	819(16.4)
Continuous variables		
Mother's age	Years	39(0.8)
Consistent parenting score	[1 – 5]	81(1.6)
Family financial hardship score	[0 – 6]	14(0.3)
Child physical functioning score	[0 – 100]	785(15.8)
Child SDQ total score*	[0 – 40]	15(0.3)
Child SDQ total score**	[0 – 40]	1185(23.8)

† Mother consumes > 2 standard drinks of alcohol daily.

‡ Probable serious mental illness / No probable serious mental illness.

* Measured at 4–5 years (wave 1).

** Measured at 8–9 years (wave 3).

Results

We initially carried out a complete case analysis to investigate the association between exposure to maternal emotional distress at 4–5 years and SDQ total score at 8–9 years, controlling for potential confounders. Performing a complete case analysis reduced the number of participants to 3244 from 4983 recruited individuals in the kindergarten cohort (i.e. 35% reduction). Further, the assessment of missing data suggests that a complete case analysis may lead to biased estimates as there are some characteristics that differed between children with incomplete (i.e. who had data available on at least one characteristic but missing observations on one or more other characteristics) and complete (i.e. who had data available on all twelve variables listed in table 1) observations (see table A1 in appendix A).

We implemented a standard MI procedure under the assumption of MAR, and then performed sensitivity analyses using the delta-adjustment method to assess the research question under plausible departures from MAR. We elicited the required values for the sensitivity parameters from content experts, and we obtained marginal distributions that reflect the uncertainty about the quantities of interest.

Eliciting expert opinion about two sensitivity parameters

As described earlier, for performing sensitivity analyses, we assumed that only the missing data in the SDQ total score (Y) and maternal emotional distress (X) follow a MNAR missingness mechanism. For simplicity we assumed that the two missingness mechanisms are independent (i.e. $R_x \perp\!\!\!\perp R_y$). This means that we elicited experts' opinions regarding the marginal distributions of the sensitivity parameters for the outcome and exposure separately (elicitation of a joint distribution for two sensitivity parameters is beyond the scope of this manuscript). In the context of the pattern-mixture modelling, the sensitivity parameters for the outcome and exposure variables correspond to the difference between LSAC non-respondents and respondents, which represent (1) the mean differences in the SDQ total score at 8–9 years (denoted by $\delta_{(y)} = \mu_{R_y(0)} - \mu_{R_y(1)}$, where the subscripts $R_y(0)$ and $R_y(1)$ correspond to R_y=0 and R_y=1, respectively), and (2) the shift in the $\log_e$ odds of mothers with probable serious mental illness at 4–5 years (denoted by $\delta_{(x)} = log(\frac{\pi_{R_x(0)}/1-\pi_{R_x(0)}}{\pi_{R_x(1)}/1-\pi_{R_x(1)}})$, where $\pi_{R_x(0)}$ and $\pi_{R_x(1)}$ correspond to proportions of mothers with probable serious mental illness when R_x=0 and R_x=1, respectively). A summary of the elicitation process for these parameters is provided in the following sections.

Choice of experts and format of elicitation

Three experts, who had relevant knowledge about child health and development, were invited to form a panel for the elicitation task. The elicitation was implemented through a written questionnaire (see appendix B) that was explained and discussed face-to-face with each expert individually.

Elicitation task

Since the aim of elicitation is to present an expert's judgement on how those with missing data differ to those with observed data (i.e. the sensitivity parameter) in a form of a probability distribution, they are often asked to determine desirable summary statistics (e.g. probabilities, measure of location and/or measures of spread) for this difference. Investigators may assist experts in this regard by providing the distributions of the incomplete variables in the observed data based on their elicited summaries to determine whether the summaries provided reflect their opinion.

Following a brief description of the primary outcome and exposure of interest, the questionnaire asked experts' expectation on how individuals with missing data differ from those with observed data (i.e. sensitivity parameter) in a form of a probability distribution. In particular, they were asked to determine desirable summary statistics for this difference including the mean differences in the SDQ total score at 8–9 years, and the proportional change in mothers who were emotionally distressed at 4–5 years between LSAC participants with missing and observed data. Note that for the latter sensitivity parameter (i.e. shift in the proportions, $\pi_{R_x(0)} - \pi_{R_x(1)}$) this was not exactly the same as the sensitivity parameter used in the delta-adjustment method as an offset in the imputation model (i.e. shift in the $\log_e$ odds- $\delta_{(x)} = log(\frac{\pi_{R_x(0)}/1-\pi_{R_x(0)}}{\pi_{R_x(1)}/1-\pi_{R_x(1)}})$). Elicitation of the shift in the proportions was used to ease interpretation and avoid any confusion in eliciting the shift in $\log_e$

odds. The elicited proportional shift was then transformed into the $\log_e$ odds scale using the proportion of mothers with probable serious mental illness in the observed data (i.e. $\pi_{R_x(1)}$ =21%), and the corresponding $\pi_{R_x(0)}$ calculated from $\pi_{R_x(0)} - \pi_{R_x(1)}$.

The experts were asked to provide a small number of quantile judgements for each of the sensitivity parameters including a median value, upper and lower quartiles and minimum and maximum values (see table 2) for $\mu_{R_y(0)} - \mu_{R_y(1)}$ and $\pi_{R_x(0)} - \pi_{R_x(1)}$. To help the experts to provide these quantities, investigators provided the distributions of the incomplete outcome and exposure in the observed data based on their elicited summaries to determine whether the summaries provided reflect their opinion. In addition, an example of a hypothetical expert was given in the questionnaire with the graphical illustration of the expert's probability distribution function for $\mu_{R_y(0)} - \mu_{R_y(1)}$ and $\pi_{R_x(0)} - \pi_{R_x(1)}$ (see appendix B).

Feedback

Feedback was provided after eliciting the required summaries from each expert by asking them whether the elicited quantities adequately express each expert's opinion. For example, based on the information provided by expert 1 in table 2 (bottom), the following feedback was given.

"We know that the proportion of maternal emotional distress in the observed data is 21%. Based on the summaries that you provided, you think among those who are not observed, this proportion is going to be somewhere between 22% and 30%, and your best guess is 25%. Does that sound about right?"

If the experts believed that the chosen summaries did not represent their opinion, they were kindly asked to provide the summaries again.

Fitting a distribution and pooling experts' opinion

The elicited quantile judgements from each expert, corresponding to the points on the cumulative distribution functions (CDF), were converted into a suitable parametric probability distribution (PDF) for $\mu_{R_y(0)} - \mu_{R_y(1)}$ and $\pi_{R_x(0)} - \pi_{R_x(1)}$ using SHELF package (O'Hagan, 2013; Oakley, 2017) in R (R Development Core Team, 2005) (see appendix C for more details). The linear pooling method, explained earlier, with equal weight given to each expert, was adopted to combine the experts' individual PDFs into a single PDF (see figure 1).

Table 2. Elicitation of prior information for the distribution of the mean difference in the SDQ total score at 8–9 years ($\delta_{(y)} = \mu_{Ry(0)} - \mu_{Ry(1)}$), and the average change in the percentage of mothers who were emotionally distressed at 4–5 years ($\pi_{Rx(0)} - \pi_{Rx(1)}$), between LSAC non-respondents and respondents.

	Minimum	Lower quartile	Median	Upper quartile	Maximum
$\mu_{R_y(0)} - \mu_{R_y(1)}$					
Hypothetical example	**-1**	**1**	**3**	**7**	**10**
Expert 1 response	0.5	0.75	1.3	2.25	2.5
Expert 2 response	-1	1	2.6	8	10.6
Expert 3 response	-1	1	3	6	9
$\pi_{R_x(0)} - \pi_{R_x(1)}$					
Hypothetical example	**-5%**	**3%**	**5%**	**9%**	**20%**
Expert 1 response	1%	2.5%	4%	7%	9%
Expert 2 response	-3%	5%	10%	15%	20%
Expert 3 response	0	2%	10%	18%	26%

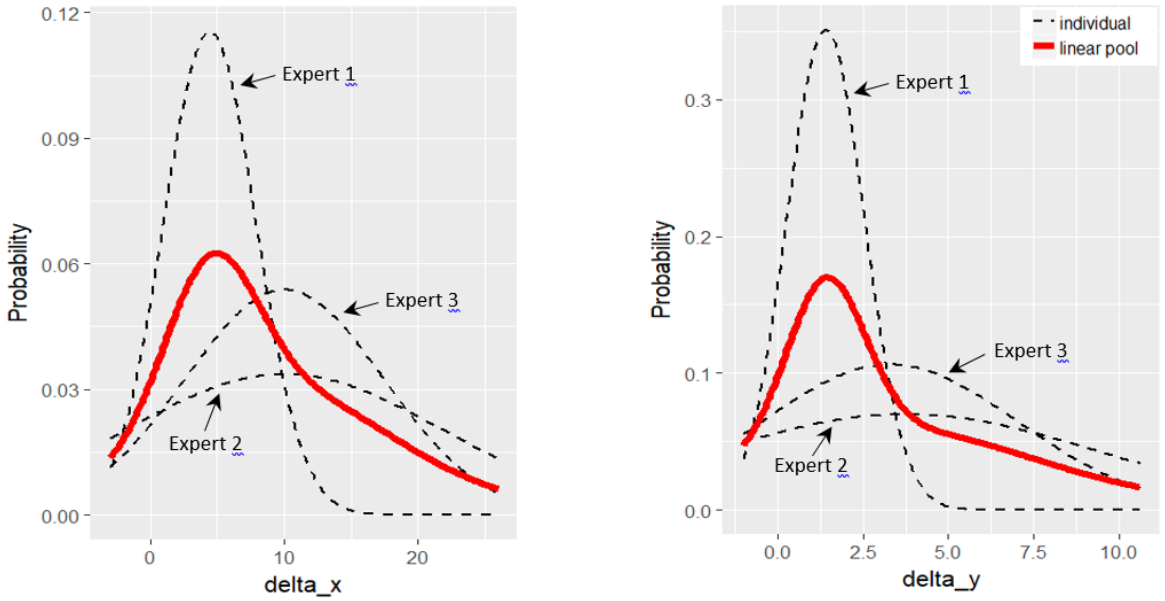

Figure 1. Graphical illustration of linear opinion pooling. Dashed lines represent experts' individual PDFs and solid lines correspond to the final pooled PDF. Left panel: elicited and pooled distributions for the average change in the proportion of maternal emotional distress at 4–5 years between non-respondents and respondents ($\pi_{Rx(0)}$- $\pi_{Rx(1)}$). Right panel: elicited and pooled distributions for the average differences of SDQ total score at 8–9 years between non-respondents and respondents ($\mu_{Ry(0)}$-$\mu_{Ry(1)}$).

Table 3. Percentiles of the pooled distributions for the sensitivity parameters of interest.

Sensitivity parameter	5th percentile	Lower quartile	Median	Upper quartile	95th percentile
$\mu_{R_y(0)} - \mu_{R_y(1)}$	-3.277	0.560	2.119	4.916	10.398
$\pi_{R_x(0)} - \pi_{R_x(1)}$	-0.042	0.028	0.069	0.130	0.239
$log(\frac{\pi_{R_x(0)}/1-\pi_{R_x(0)}}{\pi_{R_x(1)}/1-\pi_{R_x(1)}})^*$	-0.276	0.162	0.377	0.663	1.122

* $\pi_{R_x(1)}$ was set to 21%, the observed value in study sample.

It is apparent from the plots above that expert 1's distributions for both sensitivity parameters are quite different from the two other experts, and the pooled distributions are right-skewed. Consequently, instead of making random draws from the distributions for conducting sensitivity analyses, we used different percentile values of the pooled distributions (5th, 25th, 50th, 75th, and 95th percentiles) as an offset in separate pattern-mixture models. Table 3 presents the relevant percentiles of the pooled distributions for $\delta_{(y)} = \mu_{R_y(0)} - \mu_{R_y(1)}$, $\pi_{R_x(0)} - \pi_{R_x(1)}$, and $\delta_{pm(x)} = log(\frac{\pi_{R_x(0)}/1-\pi_{R_x(0)}}{\pi_{R_x(1)}/1-\pi_{R_x(1)}})$.

Multiple imputation under the assumption of MAR

MI by chained equations (MICE – also known as fully conditional specification (FCS) (van Buuren, 2007; van Buuren et al., 1999; van Buuren, 2015)), which is a widely used imputation approach for handling missing data in multiple incomplete variables with a general missingness pattern, was adopted using the user-written command –*ice*– (Royston & White, 2011) in Stata 15.0 (50 cycles and 100 imputations). The variables included in the MAR imputation models were the same as the variables in the target analysis model (see table 1) to ensure compatibility between the imputation and analysis models. In particular, the outcome

variable SDQ total score at 8–9 years was included in the imputation model to prevent the association between the outcome and covariates being falsely weakened (Moons, Donders, Stijnen, & Harrell, 2006; Sterne et al., 2009; White, Royston, & Wood, 2011). Further, two auxiliary variables measured at wave 1, which were predictors of missingness in the exposure and/or outcome, were also included in the imputation models to make the MAR assumption more plausible and to reduce bias (Collins, Schafer, & Kam, 2001; White et al., 2011). These were mother's primary language was not English (yes/no –2.9% missingness) and whether child has two parents in the home (yes/no – 0 % missingness).

All the variables included in the imputation model were imputed except the completely observed variables, sex of child, whether child has a sibling in the home, and whether child has two parents in the home. The estimates of the model parameters which were obtained from separate analyses of each of the 100 completed datasets were combined to provide an overall MI estimate (MI_{MAR}). The estimates of the regression coefficients obtained from MI under MAR as well as complete case analysis are shown in table 4 with their corresponding standard errors (SE) and 95% confidence intervals (95% CI).

As expected, MI increased the precision of the regression estimates slightly in comparison to the complete case analysis. The differences between the estimates obtained from the complete case analysis and MI under MAR were more apparent for the variables with higher proportion of missing data (e.g. maternal emotional distress, mother current cigarette smoker, and mother alcohol consumption; 16–19%) as well as mother high school completion, family financial hardship score, and whether child has a sibling in the home. Both complete case analysis and MI under MAR provide estimates, CIs and p-value that collectively provide evidence for positive relationship between total difficulties at 8–9 years and maternal emotional distress at 4–5 years.

Sensitivity analyses to the MAR assumption using the pattern-mixture method

As mentioned previously, different percentiles of the pooled distributions for the two sensitivity parameters were used to illustrate the application of the delta-adjustment method for performing sensitivity analyses following MI in the LSAC case study. The sensitivity analysis was implemented using –*uvis*– command (i.e. univariate imputation sampling, which imputes missing observations in a single variable given a set of predictors) in Stata 15.0 (Royston, 2004; Royston, 2005). We repeatedly called this procedure to iteratively draw missing values from a specified set of univariate conditional distributions for each incomplete variable, where we included the elicited sensitivity parameters as offsets in univariate imputation models for the outcome and exposure. Although the –*uvis*– command is repeatedly called by the standard routine of "*mi impute chained*" or the user written –*ice*– commands to perform multivariate imputation, incorporating offsets in each univariate imputation model for both continuous and binary incomplete variables simultaneously was not readily available in these procedures.

The variables included in the MNAR imputation models were the same as the MAR imputation models, that is, the target analysis variables as well as the two auxiliary variables. The missing data in the outcome and exposure were imputed such that imputations were drawn from imputation models assuming MNAR. We used different percentile values of the pooled distributions for $\delta_{(y)}$ and $\delta_{(x)}$ derived previously, and included these values as an offset to the linear predictors of the univariate imputation models for imputing the outcome and exposure, respectively. Estimates of the regression coefficients (95% CIs) obtained from the sensitivity analyses (MI under MNAR- MI_{MNAR}) are presented in figure 2, table 4, and table D1 in appendix D.

Table 4. Linear regression analysis results from complete case analysis, MI under MAR(MI_{MAR}), and MI under MNAR (MI_{MAR}) using the pattern-mixture method for the LSAC case study; outcome variable is total social, emotional and behavioural difficulties of children as assessed by the total score on Strength and Difficulties Questionnaire (SDQ total score) measured at 8–9 years (wave 3).

Variables	Complete analysis		MI_{MAR}		MI_{MNAR}*	
	Coefficient (SE)	95% CI	Coefficient (SE)	95% CI	Coefficient (SE)	95% CI
Maternal emotional distress†	0.59 (0.20)	(0.20, 0.98)	0.67 (0.20)	(0.27, 1.06)	0.82 (0.20)	(0.43, 1.21)
Child SDQ total score†	0.50 (0.02)	(0.47, 0.53)	0.50 (0.02)	(0.47, 0.53)	0.50 (0.02)	(0.47, 0.53)
Mother's age	-0.02 (0.02)	(-0.05, 0.01)	-0.03 (0.01)	(-0.05, 0.00)	-0.04 (0.01)	(-0.07, -0.01)
Sex of child	1.12 (0.15)	(0.83, 1.41)	1.09 (0.14)	(0.82, 1.36)	1.11 (0.14)	(0.84, 1.38)
Child has a sibling in the home	-0.78 (0.24)	(-1.25, -0.31)	-0.82 (0.22)	(-1.25, -0.40)	-0.85 (0.23)	(-1.29, -0.41)
Mother completed high school	-0.49 (0.16)	(-0.80, -0.18)	-0.58 (0.15)	(-0.87, -0.30)	-0.75 (0.16)	(-1.05, -0.44)
Mother current cigarette smoker	0.34 (0.20)	(-0.05, 0.72)	0.30 (0.19)	(-0.08, 0.68)	0.38 (0.21)	(-0.03, 0.80)
Mother alcohol consumption‡	-0.32 (0.37)	(-1.04, 0.41)	-0.31 (0.37)	(-1.03, 0.41)	-0.23 (0.41)	(-1.02, 0.57)
Consistent parenting score	-0.12 (0.12)	(-0.36, 0.12)	-0.12 (0.12)	(-0.35, 0.11)	-0.27 (0.12)	(-0.51, -0.04)
Child physical functioning score	-0.03 (0.01)	(-0.05, -0.02)	-0.03 (0.01)	(-0.04, -0.02)	-0.03 (0.01)	(-0.04, -0.02)
Family financial hardship score	0.51 (0.09)	(0.33, 0.69)	0.39 (0.09)	(0.22, 0.56)	0.46 (0.09)	(0.30, 0.63)

**50th* percentile value of the pooled distribution obtained from experts used as a sensitivity parameter in the MNAR analysis.
†Measured at 4–5 years (wave 1). ‡Mother consumes >2 standard drinks of alcohol daily

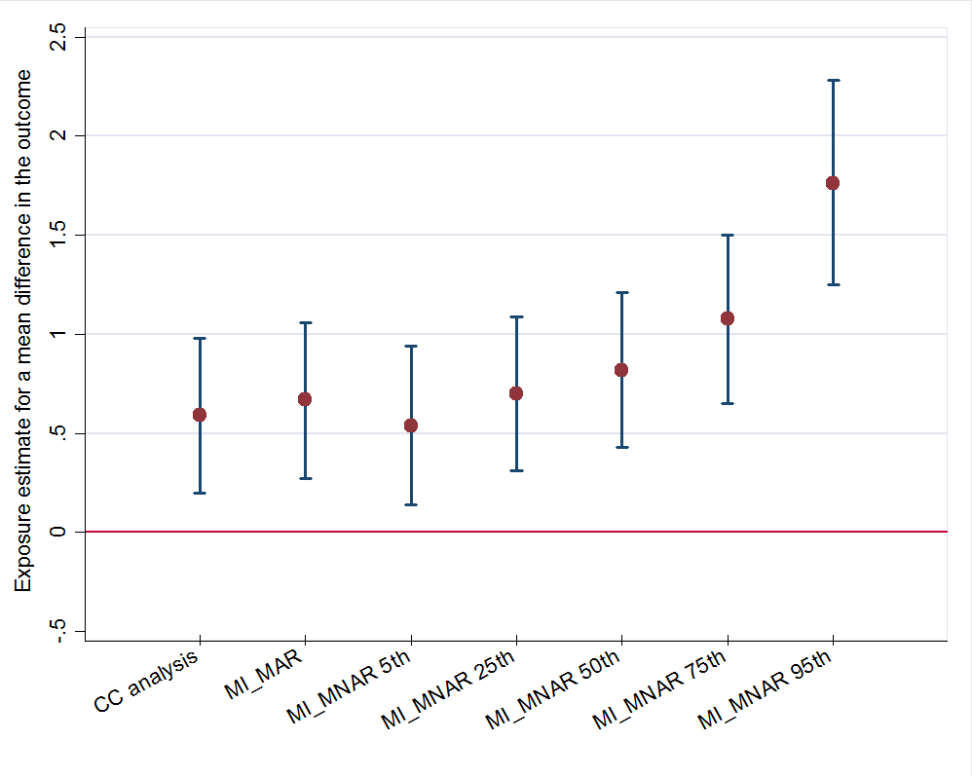

Figure 2. Estimates of the regression coefficient (95% CI) for the target analysis (i.e. the estimated adjusted mean difference in SDQ total score at 8–9 years associated with the exposure, maternal emotional distress at 4–5 years) obtained from a complete case analysis, MI under MAR, and MI under MNAR using different percentile values (5th, 25th, 50th, 75th, 95th percentile) of the pooled elicited distributions for the sensitivity parameters.

Table D1 in appendix D shows that the magnitude of regression estimates for the exposure and some of the confounder variables (e.g. mother current cigarette smoker and mother alcohol consumption) changes as the higher percentile values of the pooled distribution are used as a sensitivity parameter for both the exposure and outcome. This is not surprising since these variables have a high proportion of missing data and the missingness in these variables generally coincided with missingness in the exposure variable.

It is apparent from figure 2 that the MNAR estimates using the 5th and 25th percentiles of the pooled distributions for the sensitivity parameter are similar to the MAR estimate, since these percentile values represent minimal departures from MAR (see table 4).

Using the higher percentile values of the pooled distributions for the sensitivity parameter (i.e. 50th, 75th, and 95th) results in greater departures from MAR and therefore greater shifts in the MNAR compared with the MAR parameter estimates, as expected. Of note though, under all MNAR scenarios, the sensitivity analysis provides the same overall conclusion; that is, there is evidence that exposure to maternal emotional distress at 4–5 years is associated with the higher levels of SDQ total score at 8–9 years. Thus, we can conclude that the result obtained under the MAR analysis is robust (not sensitive) to plausible departures from the MAR assumption.

However, it is important to note that although the overall conclusion has not been changed under all MNAR scenarios, the magnitude of the

association between exposure to maternal emotional distress at 4–5 years and SDQ total score at 8–9 years has increased from 0.54 (95% CI: 0.14, 0.94) using 5th percentile values of sensitivity parameters for both the exposure and outcome to 1.76 (95% CI: 1.25, 2.28) for the 95th percentile values, potentially a clinically important difference.

Discussion

Drop-out and non-response are problematic in LSAC and many other large-scale longitudinal studies with multiple follow-up waves of data collection. Since the complete cases in the LSAC case study appeared to differ from the incomplete cases on a number of characteristics, we adopted MI for handling missing data to avoid producing biased results. While there were small differences between the magnitude of the regression estimates obtained from the complete case analysis and MI, both of these methods provided evidence for the positive relationship between maternal emotional distress at 4–5 years and SDQ total score at 8–9 years. Further, the findings did not differ dramatically in terms of precision of the parameter estimates, which may be related to the small number of auxiliary variables that were included in the imputation models and/or small proportion of missing observations in the variables imputed.

We conducted sensitivity analyses using the delta-adjustment method to assess the potential impact of departures from MAR on the final inference with elicitation from content experts for specification of the distributions of the sensitivity parameters. The elicited distributions suggested that children with higher SDQ total score and children with mothers who were emotionally distressed were more likely to be non-respondents, the elicited mean difference in SDQ total score (at 8–9 years) would most likely be 2 units, and the most probable average shift in the proportion of emotionally distressed mothers (at 4–5 years) would be 6.9%. Our results show that exposure to maternal emotional distress at 4–5 years was associated with higher levels of SDQ total score at 8–9 years, and that the magnitude of this association increases two-fold only for the extreme values of the pooled distributions of the sensitivity parameters.

Limitations and future work

In their paper, van Buuren et al. (1999) illustrated the δ-adjustment method with missing data in the highly correlated variables, systolic and diastolic blood pressure, where the effect of applying the δ-adjustment on systolic blood pressure (i.e. adding an offset δ to the univariate imputation model) was carried over to the diastolic blood pressure. This problem is due to the feedback of δ-adjustment, where the offset is amplified by the iterative FCS algorithm in the presence of highly correlated incomplete variables in the imputation model (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012). Such feedback could be problematic since the offset may not represent the actual departure from MAR, and may not correspond to the value of the sensitivity parameter that was elicited from the content expert. To avoid the issue of feedback, it was initially suggested to exclude the strongest predictors for the incomplete variables in the imputation model (van Buuren & Groothuis-Oudshoorn, 2011). A correction was also suggested to multiply δ by a damping factor (van Buuren, 2012); however, the performance of the δ-adjustment method has not been explored further, e.g. in scenarios with different types of incomplete variables using multiple offset values corresponding to each of the incomplete variables. Thus, further exploration of the performance of the delta-adjustment method under a range of realistic scenarios would be desirable to provide methodological insights and assist practical researchers to deal with incomplete data in complex settings, where the missing data mechanism is MNAR.

Our case study involved multiple incomplete variables, where it was suspected that the incomplete outcome and exposure of interest followed a MNAR mechanism, where the missingness in the outcome and exposure was independent (i.e. $R_x \perp\!\!\!\perp R_y$). For ease of exposition, the target analysis model was modified to include ten potential confounders with no interaction and non-linear terms. We also included two auxiliary variables, both of which contained a small proportion of missing data, in the imputation model to make the MAR assumption more plausible and reduce bias in implementing MI under MAR. Future research could be conducted to explore the performance of the delta-adjustment method through a number of complicated case studies using more complex imputation and analysis models. For example, exploration of missing data in more than two incomplete variables with MNAR missingness,

with different variable types and when the independence assumption of missingness indicators does not hold for multiple variables with MNAR missing data, which would require elicitation of the joint distribution between incomplete variables.

In general, for performing a MNAR sensitivity analysis, when the interest lies in the association between an outcome Y and a covariate X, considering a scenario that allows the association between X and Y to differ between non-respondents and respondents may seem more relevant and realistic. Although such elicitation in practice may be challenging, it ensures a more comprehensive investigation of departures from the MAR assumption. Further research would be to evaluate the performance of the delta-adjustment method for estimating the association between Y and X, when this association differs between non-respondents and respondents for the outcome Y by each category of X (i.e. when the assumption of independence between the missingness indicators is relaxed). For this evaluation, there would be three sensitivity parameters: $\delta_{(x)}$ to account for MNAR missing data in X (i.e. a shift in $\log_e$ odds of X =1 for R_x=0 and R_x=1), and $\delta_{(y|x=1)}$ and $\delta_{(y|x=0)}$ to account for MNAR missing data in Y for X =1 and X =0, respectively (i.e. a shift in mean of Y for $R_{y|x=1}$=0 vs. $R_{y|x=1}$=1 when X =1, and a shift in mean of Y for $R_{y|x=0}$=0 vs. $R_{y|x=0}$=1 when X=0, respectively). These three sensitivity parameters would need to be elicited from experts in order to include in the imputation models to frame the MNAR missingness in Y and X. For the LSAC case study, this would mean eliciting additional prior information to allow for the difference between missing and observed values in SDQ total score at 8–9 years to vary between those with and without maternal emotional distress at 4–5 years.

A work undertaken by Leacy and White (2015) has illustrated the importance of considering the dependence between the missingness indicators for incomplete variables when modelling MNAR missingness. They have extended the δ-adjustment method for imputing missing data under MNAR to allow for the dependence between the variables by including indicators for the missingness in the incomplete variables in each of the univariate imputation models. Further exploration is desirable to examine the proposed method in more realistic situations when the imputation model involves multiple continuous variables, as well as a mixture of binary and continuous variables with missing data, and potentially time-dependent variables.

Additional research exploring the delta-adjustment method when different MNAR imputation models for different groups of participants are required is also important. White, Kalaitzaki, and Thompson (2011) incorporated interaction terms to allow the sensitivity parameters to differ between the trial arms for a continuous outcome measured at a single time-point. Moreno-Betancur and Chavance (2013) proposed a method within the pattern-mixture modelling framework for imputing a continuous time-dependent outcome and performed a sensitivity analysis, in a clinical trial setting, to assess the robustness of the MAR results, allowing the sensitivity parameter for the incomplete repeated measure continuous outcome (baseline plus up to five follow-up visits) to differ between treatment and control groups at the last visit. Further development of the delta-adjustment method is required to incorporate sensitivity parameters for different groups of participants for variables of different data types with missing data.

As described in detail earlier, we carefully elicited the required quantile judgements from a panel of experts to specify distributions for the unknown sensitivity parameters of interest. We initially contacted six experts and invited them to be part of the elicitation task; however, the final elicitation panel was limited to three experts who were available for face-to-face interview. We gathered their opinions via separate interviews with the use of a written questionnaire and aggregated the results into a single probability distribution rather than carrying out a group elicitation, which may be subject to biases (e.g. quantile judgments provided by one expert may influence others' decisions). Of note, it is theoretically possible for experts to select a value of the sensitivity parameter that would strengthen the findings. In our study, we asked experts to consider the raw data i.e. to elicit on the raw scale, and they were not aware of the effect their elicited values/distribution had on the analysis until the end of the elicitation process. The elicited summaries from expert 1 were quite different from the other two experts who had similar responses to the hypothetical examples in the questionnaire. Providing more hypothetical examples, which consider multiple scenarios with different quantile

values, may minimise the potential to sway expert opinion.

We conducted sensitivity analyses over extreme values of the pooled distribution (i.e. 5th and 95th percentiles) to examine whether the study conclusion was sensitive to larger departures from MAR instead of randomly drawing the sensitivity parameters from their corresponding pooled distributions, since the incompatibility between the experts led to skewness in the pooled distribution. It may be of interest to examine how inferences might change across the range of values using the graphical tipping-point analysis (Liublinska & Rubin, 2014; Ratitch et al., 2013; Yan et al., 2009), where the impact of alternative assumptions regarding the missing data on the final study conclusions can be graphically assessed. Future research could also incorporate alternative approaches for summarising the results from sensitivity analyses such as a fully Bayesian methodology proposed by Scharfstein, Daniels, and Robins (2003), which allows researchers to draw a single conclusion by incorporating prior beliefs for sensitivity parameters, or uncertainty intervals developed by Vansteelandt, Goetghebeur, Kenward, and Molenberghs (2006), which accounts for both the uncertainty in the data and the uncertainty in the MNAR mechanism.

As noted earlier, the delta-adjustment method in the current research was implemented in Stata 15.0 using loops of the –*uvis*– commands; the standard routine of '*mi impute chained*' command did not offer the flexibility of including an offset in the univariate generalised linear models. For widespread uptake of the delta-adjustment method, development of a user-friendly package for conducting such sensitivity analyses in Stata and other commonly used statistical software is required. Furthermore, development of practical tools (e.g. visual representation), which allow for elicitation of multiple sensitivity parameters, and guidance on elicitation of sensitivity parameters from content experts within the MI framework is warranted.

Conclusions

Finally, the findings of this paper addressed some of the knowledge gaps in the literature by using a case study as a motivating example to illustrate the application of the delta-adjustment method in practice. Despite the limitations in this research, our investigation provides valuable insights into sensitivity analyses following MI using the delta-adjustment method when two incomplete variables follow the MNAR missingness mechanisms. Nevertheless, additional development of the method is desirable to assist the researcher for conducting sensitivity analyses in more complex studies in practice.

Acknowledgements

This work was supported by funding from the National Health and Medical Research Council: a Centre of Research Excellence grant, ID 1035261, awarded to the Victorian Centre of Biostatistics (ViCBiostat), Career Development Fellowship ID 1053609 (KJL), and Senior Research Fellowship ID 1104975 (JAS). PHR was funded by a University of Melbourne International Research Scholarship (MIRS). She is currently funded by National Institute of Mental Health (NIMH) [grant number: 1T32MH109205-01A1] and the UCLA Center for HIV Identification, Prevention, and Treatment Services (CHIPTS) grant (P30MH58107).

References

Australian Institute of Family Studies. (2015). *The Longitudinal Study of Australian Children data user guide*. Retrieved from http://www.growingupinaustralia.gov.au

Bastin, L., Cornford, D., Jones, R., Heuvelink, G. B. M., Pebesma, E., Stasch, C., . . . Williams, M. (2013). Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling and Software*, *39*, 116–134. https://doi.org/10.1016/j.envsoft.2012.02.008

Bayer, J. K., Ukoumunne, O. C., Lucas, N., Wake, M., Scalzo, K., & Nicholson, J. M. (2011). Risk factors for childhood mental health symptoms: National Longitudinal Study of Australian Children. *Pediatrics*, *128*(4), 865–879. https://doi.org/10.1542/peds.2011-0491

Burton, A., & Altman, D. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer*, *91*(1), 4–8. https://doi.org/10.1038/sj.bjc.6601907

Burzykowski, T., Carpenter, J. R., Coens, C., Evans, D., France, L., Kenward, M. G., . . . Yu, L. (2010). Missing data: Discussion points from the PSI missing data expert group. *Pharmaceutical Statistics*, *9*(4), 288–297. https://doi.org/10.1002/pst.391

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research*, *16*(3), 259–275.

Carpenter, J. R., & Kenward, M. G. (2007). *Missing data in randomised controlled trials: A practical guide*. Birmingham: Health Technology Assessment Methodology Programme. Accessed from http://researchonline.lshtm.ac.uk/4018500/

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application* (1st ed.). Chichester: Wiley. https://doi.org/10.1002/9781119942283

Carpenter, J. R., Rücker, G., & Schwarzer, G. (2011). Assessing the sensitivity of meta-analysis to selection bias: A multiple imputation approach. *Biometrics*, *67*(3), 1066–1072. https://doi.org/10.1111/j.1541-0420.2010.01498.x

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford: Oxford University Press.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 49–93. https://doi.org/10.2307/2986113

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–700. https://doi.org/10.1198/016214505000000105

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*(1), 114–135. https://doi.org/10.1214/ss/1177013825

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Hayati Rezvan, P., White, I. R., Lee, K. J., Carlin, J. B., & Simpson, J. A. (2015). Evaluation of a weighting approach for performing sensitivity analysis after multiple imputation. *BMC Medical Research Methodology*, *15*, 83. https://doi.org/10.1186/s12874-015-0074-2

Hayati Rezvan, P., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, *15*, 30. https://doi.org/10.1186/s12874-015-0022-1

Hogan, J. W., & Laird, N. M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, *16*(3), 259–272. https://doi.org/10.1002/(SICI)1097-0258(19970215)16:3<259::AID-SIM484>3.0.CO;2-S

Kadane, J. B., & Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society, Series D (the Statistician)*, *47*(1), 3–19. https://doi.org/10.1111/1467-9884.00113

Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., & Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology*, *12*(1), 96–105. https://doi.org/10.1186/1471-2288-12-96

Kenward, M., & Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, *8*(1), 51–83. https://doi.org/10.1177/096228029900800105

Kessler, R. C., Green, J. G., Gruber, M. J., Sampson, N. A., Bromet, E., Cuitan, M., . . . Zaslavsky, A. M. (2010). Screening for serious mental illness in the general population with the K6 screening scale: Results

from the WHO world mental health (WMH) survey initiative. *International Journal of Methods in Psychiatric Research*, *19 Suppl 1*, 4–22. https://doi.org/10.1002/mpr.310

Kynn, M. (2005.). *Eliciting expert knowledge for bayesian logistic regression in species habitat modelling*. (PhD thesis, Queensland University of Technology, Australia).

Leacy, F. P., Floyd, S., Yates, T. A., & White, I. R. (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *American Journal of Epidemiology*, *185*(4), 304–315. https://doi.org/10.1093/aje/kww107

Leacy, F. P., & White, I. R. (2015). *Multiple imputation for data that are missing not at random: Extending the fully conditional specification procedure* [Abstract]. (Joint Statistical Meeting) Seattle, United States. Accessed from https://ww2.amstat.org/meetings/jsm/2015/onlineprogram/AbstractDetails.cfm?abstractid=315410

Lee, K. J., & Carlin, J. B. (2017). Multiple imputation in the presence of non-normal data. *Statistics in Medicine*, *36*(4), 606–617. https://doi.org/10.1002/sim.7173

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112–1121. https://doi.org/10.1080/01621459.1995.10476615

Little, R. J. A., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355–1360. https://doi.org/10.1056/NEJMsr1203730

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, N.J.: Wiley.

Liublinska, V., & Rubin, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in Medicine*, *33*(24), 4170–4185. https://doi.org/10.1002/sim.6197

Mackinnon, A. (2010). The use and reporting of multiple imputation in medical research – a review. *Journal of Internal Medicine*, *268*(6), 586–593. https://doi.org/10.1111/j.1365-2796.2010.02274.x

Manly, C. A., & Wells, R. S. (2015). Reporting the use of multiple imputation for missing data in higher education research. *Research in Higher Education*, *56*(4), 397–409. https://doi.org/10.1007/s11162-014-9344-9

Mcconway, K. J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association*, *76*(374), 410–414. https://doi.org/10.1080/01621459.1981.10477661

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., & Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, *59*(10), 1092–1101. https://doi.org/10.1016/j.jclinepi.2006.01.009

Moreno-Betancur, M., & Chavance, M. (2013). Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research*, *25*(4), 1471–1489. https://doi.org/10.1177/0962280213490014

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling and Software*, *52*, 1–4. https://doi.org/10.1016/j.envsoft.2013.10.010

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*. Washington, D.C.: National Academies Press. Accessed from https://www.cytel.com/hs-fs/hub/1670/file-2411099288-pdf/Pdf/MissingDataNationalAcademyof_Medicine.2010.pdf

O'Hagan, A. (2006). *Uncertain judgements: Eliciting experts' probabilities* (1st ed.). London; Hoboken, NJ : John Wiley & Sons. https://doi.org/10.1002/0470033312

O'Hagan, T. (2013). *SHELF: The Sheffield Elicitation framework*. Accessed from http://www.tonyohagan.co.uk/shelf

Oakley, J. (2017). *Tools to support the Sheffield elicitation framework (SHELF) 1.3.0*. Accessed from http://www.tonyohagan.co.uk/shelf

R Development Core Team. (2005). *R: A language and environment for statistical computing, reference index version 2.2.1*. Vienna, Austria: R Foundation for Statistical Computing.

Ratitch, B., O'Kelly, M., & Tosiello, R. (2013). Missing data in clinical trials: From clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*, *12*(6), 337–347. https://doi.org/10.1002/pst.1549

Resseguier, N. (2010). Package 'SensMice'. Accessed from http://download.lww.com/wolterskluwer_vitalstream_com/PermaLink/EDE/A/EDE_2010_12_08_PAOLETTI_200963_SDC1.pdf

Resseguier, N., Giorgi, R., & Paoletti, X. (2011). Sensitivity analysis when data are missing not-at-random. *Epidemiology* (Cambridge, Mass.), *22*(2), 282–283. https://doi.org/10.1097/EDE.0b013e318209dec7

Resseguier, N., Verdoux, H., Giorgi, R., Clavel-Chapelon, F., & Paoletti, X. (2013). Dealing with missing data in the center for epidemiologic studies depression self-report scale: A study based on the French E3N cohort. *BMC Medical Research Methodology*, *13*(1), 1–11. https://doi.org/10.1186/1471-2288-13-28

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, *4*(3), 227–241.

Royston, P. (2005). Multiple imputation of missing values: Update. *Stata Journal*, *5*(2), 188–201.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software*, *45*(4). https://doi.org/10.18637/jss.v045.i04

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (1st ed.). New York, NY: Wiley. https://doi.org/10.1002/9780470316696

Scharfstein, D. O., Daniels, M. J., & Robins, J. M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics* (Oxford, England), *4*(4), 495–512. https://doi.org/10.1093/biostatistics/4.4.495

Shen, Z. (2000). *Nested multiple imputations*. (PhD thesis, Harvard University, United States).

Siddique, J., Harel, O., & Crespi, C. M. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *Annals of Applied Statistics*, *6*(4), 1814–1837. https://doi.org/10.1214/12-AOAS555

Siddique, J., Harel, O., Crespi, C. M., & Hedeker, D. (2014). Binary variable multiple–model multiple imputation to address missing data mechanism uncertainty: Application to a smoking cessation trial. *Statistics in Medicine*, *33*(17), 3013–3028. https://doi.org/10.1002/sim.6137

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal* (BMJ), *339*(7713), 157–160. https://doi.org/10.1136/bmj.b2393

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. https://doi.org/10.1177/0962280206074463

van Buuren, S. (2012). *Flexible imputation of missing data* (1st ed.). Hoboken, NJ: Taylor and Francis. https://doi.org/10.1201/b11826

van Buuren, S. (2015). Fully conditional specification. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis & G. Verbeke (Eds.). *Handbook of missing data methodology* (1st edn. ed., pp. 267–294). Chapman & Hall/CRC.

van Buuren, S., Boshuizen, H. C., & Boshuizen, H. C. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681–694.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., & Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, *16*(3), 953–980.

Varni, W. J. (2006). The PedsQL(pediatric quality of life inventory). Retrieved from http://www.pedsql.org/

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, *42*(1), 105–138. https://doi.org/10.1177/0049124112464866

White, I. R. (2015). The elicitation and use of expert opinion. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis & G. Verbeke (Eds.). *Handbook of missing data methodology* (1st ed., pp. 471–490). Chapman & Hall/CRC.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *29*(28), 2920–2931. https://doi.org/10.1002/sim.3944

White, I. R., Carpenter, J. R., Evans, S., & Schroter, S. (2007). Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials*, *4*(2), 125–139. https://doi.org/10.1177/1740774507077849

White, I. R., Horton, N. J., Carpenter, J. R., & Pocock, S. J. (2011). Strategy for intention to treat analysis in randomised trials with missing outcome data. *British Medical Journal* (BMJ), *342*(7803), 910–912. https://doi.org/10.1136/bmj.d40

White, I. R., Kalaitzaki, E., & Thompson, S. G. (2011). Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an internet-based alcohol trial. *Statistics in Medicine*, *30*(27), 3192–3207. https://doi.org/10.1002/sim.4360

Yan, X., Lee, S., & Li, N. (2009). Missing data handling methods in medical device clinical trials. *Journal of Biopharmaceutical Statistics*, *19*(6), 1085–1098. https://doi.org/10.1080/10543400903243009

Yuan, Y. (2014). *Sensitivity analysis in multiple imputation for missing data*. SAS Global Forum. Accessed from http://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf

Supplementary material

Online appendices – see website.

Appendix A: Characteristics of children with and without missing data in the LSAC case study (DOCX)
Appendix B: Elicitation questionnaire (DOCX)
Appendix C: Elicitation using SHELF package (DOCX)
Appendix D: MNAR analysis results for the LSAC case study (DOCX)