A mean score method for sensitivity analysis to departures from the missing at random assumption in randomised trials

Ian R. White[1,2,*], James Carpenter [2,3], and Nicholas J. Horton [4]

[1] MRC Biostatistics Unit, Cambridge, UK

[2] MRC Clinical Trials Unit at UCL, London, UK

[3] London School of Hygiene and Tropical Medicine, UK

[4] Amherst College, Amherst, MA, USA

[*] Email: ian.white@ucl.ac.uk

April 29, 2017

Abstract

Most analyses of randomised trials with incomplete outcomes make untestable assumptions and should therefore be subjected to sensitivity analyses. However, methods for sensitivity analyses are not widely used. We propose a mean score approach for exploring global sensitivity to departures from missing at random or other assumptions about incomplete outcome data in a randomised trial. We assume a single outcome analysed under a generalised linear model. One or more sensitivity parameters, specified by the user, measure the degree of departure from missing at random in a pattern mixture model. Advantages of our method are that its sensitivity parameters are relatively easy to interpret and so can be elicited from subject matter experts; it is fast and non-stochastic; and its point estimate, standard error and confidence interval agree perfectly with standard methods when particular values of the sensitivity parameters make those standard methods appropriate. We illustrate the method using data from a mental health trial.

Keywords: Intention-to-treat analysis, Longitudinal data analysis, Mean score, Missing data, Randomised trials, Sensitivity analysis.

Running title: Mean score method for sensitivity analysis.

1 Introduction

Missing outcome data are a threat to the validity of randomised controlled trials, and they usually require untestable assumptions to be made in the analysis. One common assumption is that data are missing at random (MAR) (Little and Rubin, 2002). Other possible assumptions may be less implausible in particular clinical settings. For example, in smoking cessation trials, the outcome is binary, indicating whether an individual quit over a given period, and it is common to assume that missing values are failures — "missing=failure" (West *et al.*, 2005); while in weight loss trials, missing data is sometimes assumed to be unchanged since baseline — "baseline observation carried forward" (Ware, 2003).

The US National Research Council (2010) suggested measures that should be taken to minimise the amount of missing outcome data in randomised trials, and described analysis strategies based on various assumptions about the missing data. This report recommended that "Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting." However, among "several important areas where progress is particularly needed", the first was "methods for sensitivity analysis and principled decision making based on the results

from sensitivity analyses". Sensitivity analysis is also an essential part of an intention-to-treat analysis strategy, which includes all randomised individuals in the analysis strategy (White *et al.*, 2011a, 2012): even if the main analysis is performed under MAR and hence draws no information from individuals with no outcome data, such individuals are included in sensitivity analysis and hence in the analysis strategy.

Sensitivity analysis is often done by performing two different analyses, such as an analysis assuming MAR and an analysis by last observation carried forward, and concluding that inference is robust if the results are similar (Wood *et al.*, 2004). Better is a principled sensitivity analysis, where the data analyst typically formulates a model including unidentified 'sensitivity parameter(s)' that govern the degree of departure from the main assumption (e.g. from MAR), and explores how the estimate of interest varies as the sensitivity parameter(s) are varied (Rotnitzky *et al.*, 1998; Kenward *et al.*, 2001). We consider global sensitivity analyses where the sensitivity parameter(s) are varied over a range of numerical values that subject-matter experts consider plausible.

Likelihood-based analyses assuming MAR can usually ignore the missing data mechanism and simply analyse the observed data (Little and Rubin, 2002). Under a missing not at random (MNAR) assumption, however, it is usually necessary to model the data of interest jointly with the assumed miss-

ing data mechanism. The joint model can be specified as a pattern-mixture model, which explicitly describes the differences between profiles of patients who complete and drop out (Little, 1993, 1994), or as a selection model, which relates the chance of drop-out to the (possibly missing) response values either directly (Diggle and Kenward, 1994; Kenward, 1998) or indirectly through a random effect (Follmann and Wu, 1995; Roy, 2003). Rotnitzky *et al.* (1998) proposed a selection model for incomplete repeated measures data and showed how to estimate it by inverse probability weighting, given values of the sensitivity parameters. Scharfstein *et al.* (2003) adopted a non-parametric Bayesian approach to analysing incomplete randomised trial data, and argued that sensitivity parameters are more plausibly a priori independent of other parameters of interest in a selection model than in a pattern-mixture model. Scharfstein *et al.* (2014) proposed a fully parametric approach based on a selection model. On the other hand, Daniels and Hogan (2000) advocated a pattern-mixture framework as "a convenient and intuitive framework for conducting sensitivity analyses".

We use the pattern-mixture model in this paper because its sensitivity parameters are usually more easily interpreted (White *et al.*, 2007, 2008). For a binary outcome, a convenient sensitivity parameter is the informative missing odds ratio (IMOR), defined, conditional on covariates, as the odds of positive outcome in missing values divided by the odds of positive outcome in

3

observed values (Higgins *et al.*, 2008; Kaciroti *et al.*, 2009). For a continuous outcome, a convenient sensitivity parameter is the covariate-adjusted mean difference between missing and observed outcomes (Mavridis *et al.*, 2015).

Most estimation procedures described for general pattern-mixture models are likelihood-based (Little, 1993, 1994; Little and Yau, 1996; Hedeker and Gibbons, 2006), while the National Research Council (2010) describes point estimation using sensitivity parameters with bootstrap standard errors. In this paper we propose instead using the mean score method, a computationally convenient method which was originally proposed under a MAR assumption for incomplete outcome data (Pepe *et al.*, 1994) and for incomplete covariates (Reilly and Pepe, 1995). The method is particularly useful to allow for auxiliary variables, so that outcomes can be assumed MAR given model covariates and auxiliary variables but not necessarily MAR given model covariates alone (Pepe *et al.*, 1994). We are not aware of the mean score method having been used for sensitivity analysis.

The aim of this paper is to propose methods for principled sensitivity analysis that are fast, non-stochastic, available in statistical software, and agree exactly with standard methods in the special cases where standard methods are appropriate. We focus on randomised trials with outcome measured at a single time, allowing for continuous or binary outcomes, or indeed any generalised linear model, and for covariate adjustment.

4

The paper is organised as follows. Section 2 describes our proposed method. Section 3 proposes small-sample corrections which yield exact equivalence to standard procedures in special cases. Section 4 illustrates our method in QUATRO, a mental health trial with outcome measured at a single time. Section 5 describes a simulation study. Section 6 discusses the implementation of our method, possible alternatives, limitations and extensions.

2   Mean score approach

Assume that for the ith individual ($i = 1$ to n) in an individually randomised trial, there is an outcome variable y_i, and let r_i be an indicator of y_i being observed. Let n_{obs} and $n_{mis} = n - n_{obs}$ be the numbers of observed and missing values of y respectively. Let $\mathbf{x}_i$ be a vector of covariates including the p_S-dimensional fully-observed covariates $\mathbf{x}_{Si}$ in the substantive model, comprising an intercept, an indicator z_i for the randomised group, and (optionally) baseline covariates. $\mathbf{x}_i$ may also include fully-observed auxiliary covariates $\mathbf{x}_{Ai}$ that are not in the substantive model but that help to predict y_i, and/or covariates $\mathbf{x}_{Ri}$ that are only observed in individuals with missing y_i and describe the nature of the missing data: for example, the reason for missingness.

The aim of the analysis is to estimate the effect of randomised group, adjusting for the baseline covariates. We assume the substantive model is a generalised linear model (GLM) with canonical link,

$$E\left[y_i|\mathbf{x}_{Si}; \boldsymbol{\beta}_S\right] = h(\boldsymbol{\beta}_S^T \mathbf{x}_{Si}) \tag{1}$$

where $h(.)$ is the inverse link function. We are interested in estimating β_{Sz}, the component of the p_S-dimensional vector $\boldsymbol{\beta}_S$ corresponding to z.

If we had complete data, we would estimate $\boldsymbol{\beta}_S$ by solving the estimating equation $U_S^*(\boldsymbol{\beta}_S) = 0$ where $U_S^*(\boldsymbol{\beta}_S) = \sum_i U_{Si}^*(\boldsymbol{\beta}_S)$ and

$$U_{Si}^*(\boldsymbol{\beta}_S) = \{y_i - h(\boldsymbol{\beta}_S^T \mathbf{x}_{Si})\}\mathbf{x}_{Si}. \tag{2}$$

The mean score approach (Pepe *et al.*, 1994; Reilly and Pepe, 1995) handles missing data by replacing $U_S^*(\boldsymbol{\beta}_S)$ with $U_S(\boldsymbol{\beta}_S)$, its expectation over the distribution of the missing data given the observed data. We write $U_S(\boldsymbol{\beta}_S) = \sum_i U_{Si}(\boldsymbol{\beta}_S)$ and $U_{Si}(\boldsymbol{\beta}_S) = \mathrm{E}\left[U_{Si}^*(\boldsymbol{\beta}_S)|\mathbf{x}_i, r_i, r_i y_i\right]$. Then by the repeated expectation rule, $\mathrm{E}\left[U_{Si}(\boldsymbol{\beta}_S)|\mathbf{x}_{Si}\right] = \mathrm{E}\left[U_{Si}^*(\boldsymbol{\beta}_S)|\mathbf{x}_{Si}\right]$ since $\mathbf{x}_i$ includes $\mathbf{x}_{Si}$, so $U_S(\boldsymbol{\beta}_S) = 0$ is an unbiased estimating equation if $U_S^*(\boldsymbol{\beta}_S) = 0$ is.

To compute $U_{Si}(\boldsymbol{\beta}_S)$, we need only $\mathrm{E}\left[y_i|\mathbf{x}_i, r_i = 0\right]$, because (2) is linear in y_i. We estimate this using the pattern-mixture model

$$E\left[y_i|\mathbf{x}_i, r_i; \boldsymbol{\beta}_P\right] = h\left(\boldsymbol{\beta}_P^T \mathbf{x}_{Pi} + \Delta(\mathbf{x}_i)(1 - r_i)\right) \tag{3}$$

where $\mathbf{x}_{Pi} = (\mathbf{x}_{Si}, \mathbf{x}_{Ai})$ of dimension p_P; the subscript P distinguishes the

6

parameters $\boldsymbol{\beta}_P$ and covariates $\mathbf{x}_{Pi}$ of the pattern-mixture model from the parameters $\boldsymbol{\beta}_S$ and covariates $\mathbf{x}_{Si}$ of the substantive model. Models (1) and (3) are typically not both correctly specified: we return to this issue in the simulation study.

In equation (3), $\Delta(\mathbf{x}_i)$ is a user-specified departure from MAR for individual i. MAR in this setting means that $p(r_i = 1|\mathbf{x}_i, y_i) = p(r_i = 1|\mathbf{x}_i)$, which implies $\mathrm{E}\,[y_i|\mathbf{x}_i, r_i] = \mathrm{E}\,[y_i|\mathbf{x}_i]$ and hence $\Delta(\mathbf{x}_i) = 0$ for all i. A simple choice of $\Delta(\mathbf{x}_i)$ that expresses MNAR is $\Delta(\mathbf{x}_i) = \delta$, where the departure from MAR is the same for all individuals. Differences in departure from MAR between randomised groups are often plausible and can have strong impact on the estimated treatment effect (White $et\ al.$, 2007), so an alternative choice is $\Delta(\mathbf{x}_i) = \delta_{z_i}$. The departure $\Delta(\mathbf{x}_i)$ could also depend on reasons for missingness coded in $\mathbf{x}_{Ri}$: for example, it could be 0 for individuals lost to follow-up (if MAR seemed plausible for them), and δ_{z_i} for individuals who refused follow-up.

Putting it all together, the mean score method solves

$$\sum_i \{\tilde{y}_i(\boldsymbol{\beta}_P) - h(\boldsymbol{\beta}_S^T \mathbf{x}_{Si})\}\mathbf{x}_{Si} = \mathbf{0} \tag{4}$$

where $\tilde{y}_i(\boldsymbol{\beta}_P)$ is defined as y_i if $r_i = 1$ and $h\left(\boldsymbol{\beta}_P^T \mathbf{x}_{Pi} + \Delta(\mathbf{x}_i)(1 - r_i)\right)$ if $r_i = 0$.

2.1 Estimation using full sandwich variance

The parameter $\boldsymbol{\beta}_P$ in (3) is estimated by regressing y_i on $\mathbf{x}_{Pi}$ in the complete cases ($r_i = 1$). Once $\boldsymbol{\beta}_P$ is estimated, we calculate the $\tilde{y}_i(\hat{\boldsymbol{\beta}}_P)$ using the known values $\Delta(\mathbf{x}_i)$ and solve (4) for $\boldsymbol{\beta}_S$. The whole procedure amounts to solving the set of estimating equations $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ where $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, \boldsymbol{\beta}_P^T)^T$, $\mathbf{U}(\boldsymbol{\beta}) = (\mathbf{U}_S(\boldsymbol{\beta})^T, \mathbf{U}_P(\boldsymbol{\beta}_P)^T)^T$, $\mathbf{U}_S(\boldsymbol{\beta}) = \sum_i \mathbf{U}_{Si}(\boldsymbol{\beta})$, $\mathbf{U}_P(\boldsymbol{\beta}_P) = \sum_i \mathbf{U}_{Pi}(\boldsymbol{\beta}_P)$,

$$\left. \begin{aligned} \mathbf{U}_{Si}(\boldsymbol{\beta}) &= \left\{ r_i y_i + (1-r_i)h(\boldsymbol{\beta}_P^T \mathbf{x}_{Pi} + \Delta(\mathbf{x}_i)) - h(\boldsymbol{\beta}_S^T \mathbf{x}_{Si}) \right\} \mathbf{x}_{Si}, \\ \mathbf{U}_{Pi}(\boldsymbol{\beta}_P) &= r_i \left\{ y_i - h(\boldsymbol{\beta}_P^T \mathbf{x}_{Pi}) \right\} \mathbf{x}_{Pi}. \end{aligned} \right\} \tag{5}$$

Pepe *et al.* (1994) derived a variance expression for $\boldsymbol{\beta}_S$ assuming categorical $\mathbf{x}_{Si}$. To accommodate any form of $\mathbf{x}_{Si}$, we instead obtain standard errors by the sandwich method, based on both estimating equations. The sandwich estimator of var $\left(\hat{\boldsymbol{\beta}} \right)$ is

$$\mathbf{V} = \mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-T} \tag{6}$$

where $\mathbf{B} = -d\mathbf{U}/d\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\mathbf{C} = \sum_i \mathbf{U}_i(\hat{\boldsymbol{\beta}})\mathbf{U}_i(\hat{\boldsymbol{\beta}})^T$ and $\mathbf{U}_i(\boldsymbol{\beta}) = (\mathbf{U}_{Si}(\boldsymbol{\beta})^T, \mathbf{U}_{Pi}(\boldsymbol{\beta}_P)^T)^T$. $\mathbf{B}$ and $\mathbf{C}$ are given in Section A of the Supplementary Materials.

2.2 Estimation using two linear regressions

A special case arises if there are no auxiliary variables, so $\mathbf{x}_{Pi} = \mathbf{x}_{Si}$ for all i, and $h(.)$ is the identity function, as in linear regression. Then we can

rearrange (5) to give

$$\mathbf{U}_{Si}(\boldsymbol{\beta}) - \mathbf{U}_{Pi}(\boldsymbol{\beta}_P) = \left\{ (1 - r_i)\Delta(\mathbf{x}_i) - (\boldsymbol{\beta}_S - \boldsymbol{\beta}_P)^T \mathbf{x}_{Si} \right\} \mathbf{x}_{Si}.$$

Thus $(\boldsymbol{\beta}_S - \boldsymbol{\beta}_P)$ may be estimated by linear regression of $(1 - r_i)\Delta(\mathbf{x}_i)$ on $\mathbf{x}_{Si}$. This estimate is uncorrelated with $\hat{\boldsymbol{\beta}}_P$ because cov $\left(\mathbf{U}_{Si}(\boldsymbol{\beta}) - \mathbf{U}_{Pi}(\boldsymbol{\beta}_P), \mathbf{U}_{Pi}(\boldsymbol{\beta}_P) \right) = \mathbf{0}$. This gives a direct way to estimate $\boldsymbol{\beta}_S$, and its variance var $\left(\hat{\boldsymbol{\beta}}_P \right) +$ var $\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P \right)$, from standard linear regressions.

In particular, consider a two-arm trial with no covariates, and write the coefficients of z_i in (1) and (3) as β_{Sz} and β_{Pz}. Then $\beta_{Sz} - \beta_{Pz}$ is estimated as the difference between arms in the mean of $(1 - r_i)\Delta(\mathbf{x}_i)$, which is $a_1\delta_1 - a_0\delta_0$ where a_j is the proportion of missing data in arm $j = 0, 1$ and δ_j is the average of $\Delta(\mathbf{x}_i)$ over individuals with missing data in arm $j = 0, 1$. Therefore the estimated parameter of interest is $\hat{\boldsymbol{\beta}}_{Sz} = \hat{\boldsymbol{\beta}}_{Pz} + a_1\delta_1 - a_0\delta_0$ as in White *et al.* (2007); the same result can be derived in other ways.

3   Equivalence to standard procedures

We now consider two special cases which can be fitted by standard procedures: (1) when MAR is assumed and there are no auxiliary variables, so incomplete cases contribute no information and the standard procedure is an analysis of complete cases, and (2) when 'missing = failure' is assumed for a binary outcome, so the standard procedure is to replace missing values with

9

failures. Our aim is that point estimates, standard errors and confidence intervals produced by the mean score procedure should agree exactly with those produced by the standard procedures in these cases.

Equality of point estimates is easy to see. In case (1), we have $\Delta(\mathbf{x}_i) = 0$ and $\mathbf{x}_{Pi} = \mathbf{x}_{Si}$ for all i, so if $\boldsymbol{\beta}_P$ solves $\mathbf{U}_P(\boldsymbol{\beta}_P) = \mathbf{0}$ then $\boldsymbol{\beta} = (\boldsymbol{\beta}_P, \boldsymbol{\beta}_P)$ solves $\mathbf{U}_S(\boldsymbol{\beta}) = \mathbf{0}$. In case (2), 'missing = failure' can be expressed as $\Delta(\mathbf{x}_i) = -\infty$ for all i, so the mean score procedure gives $y_i^* = 0$ whenever $r_i = 0$, and solving $\mathbf{U}_S(\boldsymbol{\beta}_S, \boldsymbol{\beta}_P) = \mathbf{0}$ gives the same point estimate as replacing missing values with failures.

Exact equality of variances between mean score and standard procedures depends on which finite sample corrections (if any) are applied. Many such corrections have been proposed to reduce the small-sample bias of the sandwich variance estimator and to improve confidence interval coverage (Kauermann and Carroll, 2001; Lu *et al.*, 2007). Here, we assume that the standard procedures use the commonly used small-sample correction factor for the sandwich variance $\widehat{\mathrm{var}}\left(\hat{\boldsymbol{\beta}}_S\right) = f\mathbf{V}$ where $f = n/(n - p^*)$, n is the sample size and p^* is the number of regression parameters (in linear regression) or 1 (in other GLMs): this is for example the default in Stata (StataCorp, 2011).

Exact equality of confidence intervals between mean score and standard procedures additionally depends on the distributional assumptions used to construct confidence intervals. Here, we assume that standard procedures for

linear regression construct confidence intervals from the t distribution with $n - p^*$ degrees of freedom, and that standard procedures for other GLMs construct confidence intervals from the Normal distribution.

With missing data, we propose using the same small-sample correction factor, distributional assumptions and degrees of freedom, but replacing n by an effective sample size n_{eff} as shown below. Thus we propose forming confidence intervals for linear regression by assuming $\hat{\boldsymbol{\beta}}_S \sim t_{n_{eff}-p^*}\left(\boldsymbol{\beta}_S, \frac{n_{eff}}{n_{eff}-p^*}\mathbf{V}\right)$ and for other GLMs by assuming $\hat{\boldsymbol{\beta}}_S \sim N\left(\boldsymbol{\beta}_S, \frac{n_{eff}}{n_{eff}-p^*}\mathbf{V}\right)$.

3.1 Full sandwich method

For the full sandwich method of Section 2.1, we propose computing n_{eff} as $n_{obs} + (I_{mis}/I_{mis^*})n_{mis}$ where I_{mis} is the influence of the individuals with missing values, and I_{mis^*} is the same individuals' influence if the missing values had been observed. The comparison of the same individuals is crucial, because missing individuals, if observed, would have different influence from observed individuals.

To determine I_{mis}, we consider weighted estimating equations $\sum_i w_i \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0}$ with solution $\boldsymbol{\beta}_{\mathbf{w}}$. Differentiating with respect to $\mathbf{w} = (w_1, \ldots, w_n)^T$ at $\mathbf{w} = \mathbf{1}$ gives $\mathbf{U}^T + \frac{d\mathbf{U}(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\frac{d\boldsymbol{\beta}_{\mathbf{w}}}{d\mathbf{w}} = \mathbf{0}$ where $\mathbf{U}$ is a $n \times (p_S + p_P)$ matrix with ith

row $\mathbf{U}_i(\boldsymbol{\beta})$. Since also $\frac{d\mathbf{U}(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -\mathbf{B}$, we get

$$\frac{d\boldsymbol{\beta}_\mathbf{w}}{d\mathbf{w}} = -\mathbf{B}^{-1}\mathbf{U}^T. \tag{7}$$

We now define the influence of observation i as

$$I_{mis,i} = \frac{d\boldsymbol{\beta}_{\mathbf{w}S}}{dw_i}^T \mathbf{V}_S^{-1} \frac{d\boldsymbol{\beta}_{\mathbf{w}S}}{dw_i}$$

where the S subscript denotes the elements corresponding to $\boldsymbol{\beta}_S$. Hence we define the influence of the individuals with missing values as $I_{mis} = \sum_i (1 - R_i)I_{mis,i}$.

To determine I_{mis^*}, we let $\boldsymbol{\beta}_{\mathbf{w}S}^*$ be the (unknown) parameter estimate that would be obtained if the complete data had been observed, following the pattern-mixture model (3). In this case the influence would be

$$I_{mis,i}^* = \frac{d\boldsymbol{\beta}_{\mathbf{w}S}^*}{dw_i}^T \mathbf{V}_S^{-1} \frac{d\boldsymbol{\beta}_{\mathbf{w}S}^*}{dw_i}.$$

We define the "full-data influence" as the expectation of $I_{mis,i}^*$ over the distribution of the complete data given the observed data, under the pattern-mixture model (3). From (7) we get $\frac{d\boldsymbol{\beta}_\mathbf{w}^*}{dw_i} = -\mathbf{B}^{-1}\mathbf{U}_i^{*T}$ and $\mathbf{U}_i^{*T} = \left\{ y_i^* - h(\boldsymbol{\beta}_s^T \mathbf{x}_{Si}) \right\} \mathbf{x}_{Si}$, so that

$$\mathrm{E}\left[I_{mis,i}^* \right] = \mathrm{E}\left[\left\{ y_i^* - h(\boldsymbol{\beta}_s^T \mathbf{x}_{Si}) \right\}^2 \right] \mathbf{x}_{Si}^T \mathbf{B}_{SS}^{-T} \mathbf{V}_S^{-1} \mathbf{B}_{SS}^{-1} \mathbf{x}_{Si} \tag{8}$$

and $\mathrm{E}\left[\left\{ y_i^* - h(\boldsymbol{\beta}_s^T \mathbf{x}_S) \right\}^2 \right]$ is evaluated as the squared residual plus the residual variance from model (1) for individual i. Finally, $I_{mis^*} = \sum_i (1-r_i)\mathrm{E}\left[I_{mis,i}^* \right]$.

12

In case (1), $I_{mis} = 0$ so $n_{eff} = n_{obs}$, as in standard analysis. In case (2), $I_{mis*} = I_{mis}$ so $n_{eff} = n$, again as in standard analysis.

3.2 Two linear regressions method

For the two linear regressions method of Section 2.2, the small-sample correction to the variance is naturally applied separately to each variance in $\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_P\right) + \widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right) = V_{small}$ say. We can derive the corresponding variances without small-sample correction as $\frac{n_{obs}-p}{n_{obs}}\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_P\right) + \frac{n-p}{n}\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right) = V_{large}$ say. To compute n_{eff}, we use the heuristic $V_{small} \approx \frac{n_{eff}}{n_{eff}-p}V_{large}$, and hence we estimate n_{eff} by solving $|V_{small}| = \left(\frac{n_{eff}}{n_{eff}-p}\right)^p |V_{large}|$. In case (1), $\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right) = \mathbf{0}$ so $n_{eff} = n_{obs}$, as in standard analysis. Case (2) does not apply to linear regression.

4 Example: QUATRO trial

The QUATRO trial (Gray *et al.*, 2006) was a randomised controlled trial in people with schizophrenia, to evaluate the effectiveness of a patient-centred intervention to improved adherence to prescribed antipsychotic medications. The trial included 409 participants in four European centres. The primary outcome, measured at baseline and 1 year, was participants' quality of life, expressed as the mental health component score (MCS) of the SF-36 (Ware,

1993). The MCS is designed to have mean 50 and standard deviation 10 in a standard population, and a higher MCS score implies a better quality of life. The data are summarised in Table 1.

We first estimate the intervention effect on MCS, adjusted for baseline MCS and centre. Thus in the substantive model (1), $h(.)$ is the identity link, y_i is MCS at 1 year for participant i, and $\mathbf{x}_{Si}$ is a vector containing 1, randomised group z_i (1 for the intervention group and 0 for the control group), baseline MCS and dummy variables for three centres. We have no $\mathbf{x}_{Ai}$ or $\mathbf{x}_{Ri}$.

We also estimate the intervention effect on a binary variable, MCS dichotomised at an arbitrary value of 40, where for illustration we use baseline MCS and dummy variables for centre as auxiliary variables $\mathbf{x}_{Ai}$. Thus in the substantive model (1), $h(\eta) = 1/\{1 + \exp(-\eta)\}$ is the inverse of the logit link, y_i is dichotomised MCS at 1 year for participant i, and $\mathbf{x}_{Si}$ is a vector containing 1 and z_i.

In both analyses, we replace the 23 missing values of baseline MCS with the mean baseline MCS: while such mean imputation is not valid in general, it is appropriate and efficient in the specific case of estimating intervention effects with missing baseline covariates in randomised trials (White and Thompson, 2005; Groenwold *et al.*, 2012).

As expected (results not shown), the point estimate, standard error and

confidence interval from the mean score method agree exactly with standard methods under MAR and under missing=failure, using the small-sample corrections of Section 3.

We consider three sets of sensitivity analyses using the mean score method around a MAR assumption, with departures from MAR (1) in the intervention arm only ($\Delta(\mathbf{x}_i) = \delta z_i$), (2) in both arms ($\Delta(\mathbf{x}_i) = \delta$), or (3) in the control arm only ($\Delta(\mathbf{x}_i) = \delta(1 - z_i)$) (White *et al.*, 2012). For the quantitative outcome, the investigators suggested that the mean of the missing data could plausibly be lower than the mean of the observed data by up to 10 units (equal to nearly one standard deviation of the observed data), so we allow δ to range from 0 to -10 (Jackson *et al.*, 2010). The investigators were not asked about missing values of the dichotomised outcome, so for illustrative purposes we allow δ to range from 0 to -6, which is close to "missing=failure".

Figure 1 shows the results of these sensitivity analysis using the two linear regressions method for the quantitative outcome (upper panel) and using the full sandwich method for the dichotomised outcome (lower panel). Results are more sensitive to departures from MAR in the intervention arm because there are more missing data in this arm. However, the finding of a non-significant intervention effect is unchanged over these ranges of sensitivity analyses. This means that the main results of the trial are robust even to quite strong departures from MAR.

15

Figure 2 shows the effective sample size for these two analyses. Effective sample size increases from 367 at MAR for both analyses. For the dichotomised outcome it is near the total sample size of 409 when $\delta = -6$ in both arms. For the quantitative outcome it does not pass 370 because the range of δ is more moderate.

5   Simulation study

We report a simulation study aiming (i) to evaluate the performance of the mean score method when it is correctly specified, (ii) to compare the mean score method with alternatives, and (iii) to explore the impact of the incompatibility of models (1) and (3). We assume the sensitivity parameters $\Delta(\mathbf{x}_i)$ are correctly specified.

5.1   Data generating models

We generate data under four data generating models (DGMs), each with four choices of parameters. We focus on the case of a binary outcome. In DGMs 1-3, we generate data under a pattern-mixture model. In DGM 1, there are no baseline covariates, so $\mathbf{x}_i = \mathbf{x}_{Si} = \mathbf{x}_{Pi} = (1, z_i)$. We generate a treatment indicator $z_i \sim Bern(0.5)$; a missingness indicator r_i with logit $\mathrm{P}(r_i|z_i) = \alpha_1 + \alpha_z z_i$; and a binary outcome y_i following model (3) with

16

logit $P\left(y_i|z_i, r_i\right) = \beta_{P1} + \beta_{Pz}z_i + \beta_{Pr}(1 - r_i)$. The substantive model (1) is then logit $P\left(y_i|z_i\right) = \beta_{S1} + \beta_{Sz}z_i$. Because this substantive model contains only a single binary covariate, it is saturated and cannot be mis-specified. Therefore both substantive model and pattern-mixture model are correctly specified

DGM 2 extends DGM 1 by including $x_i \sim N(0, 1)$ as a single baseline covariate independent of z_i, so $\mathbf{x}_i = (1, x_i, z_i)$. The missingness indicator follows logit $P\left(r_i|x_i, z_i\right) = \alpha_1 + \alpha_x x_i + \alpha_z z_i$ and the binary outcome follows model (3) with logit $P\left(y_i|x_i, z_i, r_i\right) = \beta_{P1} + \beta_{Px}x_i + \beta_{Pz}z_i + \beta_{Pr}(1 - r_i)$. The substantive model is as in DGM 1 with $\mathbf{x}_{Si} = (1, z_i)$, and $\mathbf{x}_{Ai} = (x_i)$ is an auxiliary variable in the analysis. Thus the substantive model and pattern-mixture model are again both correctly specified.

DGM 3 is identical to DGM 2, but now x_i is included in the substantive model, which is therefore logit $P\left(y_i|x_i, z_i\right) = \beta_{S1} + \beta_{Sx}x_i + \beta_{Sz}z_i$ with $\mathbf{x}_{Si} = \mathbf{x}_{Pi} = (1, x_i, z_i)$. Now the substantive model is incorrectly specified while the pattern-mixture model remains correctly specified.

DGM 4 is a selection model. Here z_i and x_i are generated as in DGM 2, then y_i is generated following the substantive model logit $P\left(y_i|x_i, z_i\right) = \beta_{S1} + \beta_{Sx}x_i + \beta_{Sz}z_i$ and r_i is generated following logit $P\left(r_i|x_i, z_i, y_i\right) = \alpha_1 + \alpha_x x_i + \alpha_z z_i + \alpha_y y_i$. Here the substantive model is correctly specified while the pattern mixture model is mis-specified.

For the parameter values, we consider scenarios a-d for each DGM. In scenario a, the sample size is $n_{obs} = 500$; the missingness model has $\alpha_x = \alpha_z = \alpha_y = 1$ and we choose α_1 to fix $\pi_{obs} = \mathrm{P}(r = 1) = 0.75$; and the pattern mixture model has $\beta_{P1} = 0$, $\beta_{Px} = \beta_{Pz} = 1$, $\beta_{Pr} = -1$. (α_x and β_{Px} are ignored in DGM 1, α_y is ignored in DGM 1-3 and β_{Pr} is ignored in DGM 4.) Scenarios b-d vary scenario a by setting $n_{obs} = 2000$, $\pi_{obs} = 0.5$, and $\beta_{Pr} = -2$ respectively. 1000 data sets were simulated in each case. Table 2 summarises the simulation design.

5.2 Analysis methods

The mean score (MS) method is implemented as described in sections 2 and 3, with logit link. x_i is used as an auxiliary in DGM 2. In DGM 1-3, $\Delta(\mathbf{x}_i)$ is taken to equal the known value $-\beta_{Pr}$ for all individuals; in DGM 4, $\Delta(\mathbf{x}_i)$ is not known but (for the purposes of the simulation study) is estimated by fitting the pattern-mixture model to a data set of size 1,000,000 before data deletion.

The MS method is compared with analysis of data before data deletion (Full); analysis of complete cases (CC), which wrongly assumes MAR; and two alternative methods that allow for MNAR, multiple imputation (MI) and selection model with inverse probability weighting (SM).

In the MI approach (Rubin, 1987; White *et al.*, 2011b), the imputation model is equation (3), and data are imputed with an offset $\Delta(\mathbf{x}_i)$ in the imputation model. The number of imputations is fixed at 30.

In the SM approach, we use the response model logit $p(r_i = 1|y_i, \mathbf{x}_i) = \boldsymbol{\alpha}^T \mathbf{x}_{Pi} + \Delta^*(\mathbf{x}_i) y_i$ where the sensitivity parameter $\Delta^*(\mathbf{x}_i)$ expresses departure from MAR as the log odds ratio for response per 1-unit change in y_i. In DGM 4, $\Delta^*(\mathbf{x}_i)$ is taken to equal the known value α_y; in DGM 1-3, $\Delta^*(\mathbf{x}_i)$ is not known but (for the purposes of the simulation study) is estimated by fitting the selection model to a data set of size 1,000,000 before data deletion. The parameters $\boldsymbol{\alpha}$ cannot be estimated by standard methods, since some y_i are missing, so we use a weighted estimating equation which does not involve the missing y_i's (Rotnitzky *et al.*, 1998; Dufouil *et al.*, 2004; National Research Council, 2010):

$$\sum_i \mathbf{x}_{Pi} \left\{ \frac{r_i}{h(\boldsymbol{\alpha}^T \mathbf{x}_{Pi} + \Delta^*(\mathbf{x}_i) y_i)} - 1 \right\} = 0.$$

The substantive model is then fitted to the complete cases with stabilised weights $\hat{p}(r_i = 1|\mathbf{x}_{Si})/\hat{p}(r_i = 1|y_i, \mathbf{x}_i)$, where $\hat{p}(r_i = 1|\mathbf{x}_{Si})$ is estimated by the same procedure as $\hat{p}(r_i = 1|y_i, \mathbf{x}_i)$ but with no $\Delta^*(\mathbf{x}_i) y_i$ or $\mathbf{x}_{Ai}$ terms (Robins *et al.*, 2000). Variances are computed by the sandwich variance formula, ignoring uncertainty in $\hat{\boldsymbol{\alpha}}$.

19

5.3   Estimand

The estimand of interest is the coefficient β_{Sz} in the substantive model. It is computed by fitting the substantive model to the data set of size 1,000,000 before data deletion. We explore bias, empirical and model-based standard errors, and coverage of estimates $\hat{\beta}_{Sz}$.

5.4   Results

Results are shown in Table 3. CC is always biased, often inefficient, and poorly covering. Small bias (at most 3% of the true value) is observed in the "Full" analysis (i.e. before data deletion) in some settings: this is a small-sample effect (Nemes *et al.*, 2009), since as noted above, the true value and "Full" are calculated in the same way with large and small samples respectively. Taking "Full" as a gold standard, MS, MI and SM methods all have minimal bias (at most 2% of true values). Precisions of MS and MI are similar, with SM slightly inferior in DGM 2. Coverages are near 95%, with some over-coverage for SM in some settings, as a consequence of slightly overestimated standard errors (results not shown) due to ignoring uncertainty in $\hat{\boldsymbol{\alpha}}$ (Lunceford and Davidian, 2004). The performance of MS is not appreciably worse when the selection model is mis-specified (DGM 3) or when the pattern-mixture model is mis-specified (DGM 4). Computation

times for MI are 15-18 times longer than for MS, which is 10-30% longer than SM.

6   Discussion

We have proposed a mean score method which works well when the sensitivity parameters are known. In practice, of course, the sensitivity parameters are unknown, and a range of values will be used in a sensitivity analysis.

The main practical difficulty in implementing any principled sensitivity analysis is choosing the value(s) of the sensitivity parameters. This is a subjective process requiring subject-matter knowledge and is best done by discussion between statisticians and suitable 'experts', typically the trial investigators. By using the pattern-mixture model, we use a sensitivity parameter $\Delta(\mathbf{x}_i)$ that is easier to communicate with 'experts' than the corresponding parameters in selection models or shared parameter models. The procedure has been successfully applied in several trials (White, 2015). Special attention is needed to the possibility that $\Delta(\mathbf{x}_i)$ varies between randomised groups, because estimated treatment effects are highly sensitive to such variation (White et $al.$, 2007). As with all aspects of trial analysis, plausible ranges of the $\Delta(\mathbf{x}_i)$ parameters should be defined before the data are collected or before any analysis. An alternative approach would report the

"tipping point", the value of $\Delta(\mathbf{x}_i)$ for which the main results are substantively affected, leaving the reader to make the subjective decisions about the plausibility of more extreme values (Liublinska and Rubin, 2014). Presenting this information could be complex without subjective decisions about the difference in $\Delta(\mathbf{x}_i)$ between randomised groups. Effective methods are therefore needed for eliciting sensitivity parameters.

Our method does not incorporate data on discontinuation of treatment, unless this can be included as an auxiliary variable. Our method would however be a suitable adjunct to estimation of effectiveness in a trial with good follow-up after discontinuation of treatment. Further work is needed to combine our sensitivity analysis with models for outcome before and after discontinuation of treatment (Little and Yau, 1996). In a drug trial in which follow-up ends on discontinuation of treatment, we see our method as estimating efficacy or a *de jure* estimand; if effectiveness or a *de facto* estimand is required then post-discontinuation missing data in each arm may be imputed by the methods of Carpenter *et al.* (2013). Further work is needed to perform a full sensitivity analysis in this setting.

Our use of parametric models makes our results susceptible to model misspecification, and indeed in many cases models (1) and (3) cannot both be correctly specified except under MAR. However, the simulation study shows that the impact of such model inconsistency is small relative to the impact of

22

assumptions about the missing data and the difficulty of knowing the values of the sensitivity parameters.

We compared the mean score method with multiple imputation and inverse probability weighting. In the case of a quantitative outcome, MI can be simplified by imputing under MAR and then adding the offset to the imputed data before fitting the substantive model and applying Rubin's rules. We could also impute under MAR and then use a weighted version of Rubin's rules to allow for MNAR (Carpenter *et al.*, 2007). Both MI methods are subject to Monte Carlo error and so seem inferior to the mean score method. A full likelihood-based analysis of the selection model would also be possible, and a Bayesian analysis could directly allow for uncertainty about $\Delta(\mathbf{x}_i)$ in a single analysis (Mason *et al.*, 2012). These alternative approaches are both more computationally complex.

The proposed mean score method can be extended in various ways. We have illustrated the method for departures from a MAR assumption, but it can equally be used if the primary analysis with a binary outcome assumed "missing = failure", by varying $\Delta(\mathbf{x}_i)$ from $-\infty$ rather than from 0. The method is also appropriate in observational studies, except that mean imputation for missing covariates is not appropriate in this context. The method can be applied to a cluster-randomised trial as described in Section B of the Supplementary Materials. Further work could allow the imputation and

23

substantive models to have different link functions, non-canonical links to be used with suitable modification to $\mathbf{U}_S$, and extension to trials with more than two arms.

Sensitivity analysis should be more widely used to assess the importance of departures from assumptions about missing data. The proposed mean score approach provides data analysts with a fast and fully theoretically justified way to perform the sensitivity analyses. It is implemented in a Stata module `rctmiss` available from Statistical Software Components (SSC) at `https://ideas.repec.org/s/boc/bocode.html`.

Supplementary materials

The supplementary materials give details of the sandwich variance in equation (6) and sketch an extension to clustered data.

Acknowledgements

References

Carpenter, J. R., Kenward, M. G. and 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., Roger, J. H. and Kenward, M. G. (2013) Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, **23**(3), 1352–71.

Daniels, M. J. and Hogan, J. W. (2000) Reparameterizing the Pattern Mixture Model for Sensitivity Analyses Under Informative Dropout. *Biometrics*, **56**(4), 1241–1248.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis. *Applied Statistics*, **43**(1), 49–93.

Dufouil, C., Brayne, C. and Clayton, D. (2004) Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine*, **23**(14), 2215–2226.

Follmann, D. and Wu, M. (1995) An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

Gray, R., Leese, M., Bindman, J., Becker, T., Burti, L., David, A., Gournay, K., Kikkert, M., Koeter, M., Puschner, B., Schene, A., Thornicroft, G. and Tansella, M. (2006) Adherence therapy for people with schizophrenia: European multicentre randomised controlled trial. *The British Journal of Psychiatry*, **189**(6), 508–514.

Groenwold, R. H. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G. and Moons, K. G. M. (2012) Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, **184**(11), 1265–1269.

Hedeker, D. R. and Gibbons, R. D. (2006) *Longitudinal data analysis.* John Wiley and Sons.

Higgins, J. P., White, I. R. and Wood, A. M. (2008) Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials*, **5**(3), 225–239.

Jackson, D., White, I. R. and Leese, M. (2010) How much can we learn about missing data?: an exploration of a clinical trial in psychiatry. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173**(3), 593–612.

Kaciroti, N. A., Schork, M. A., Raghunathan, T. and Julius, S. (2009) A

Bayesian sensitivity model for intention-to-treat analysis on binary outcomes with dropouts. *Statistics in Medicine*, **28**(4), 572–585.

Kauermann, G. and Carroll, R. J. (2001) A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, **96**(456), 1387–1396.

Kenward, M., Goetghebeur, E. and Molenberghs, G. (2001) Sensitivity analysis for incomplete categorical data. *Statistical Modelling*, **1**(1), 31–48.

Kenward, M. G. (1998) Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**(23), 2723–2732.

Little, R. and Yau, L. (1996) Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, **52**(4), 1324–1333.

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. (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**(3), 471–483.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. Hoboken, N. J.: Wiley, second edition.

Liublinska, V. and 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.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I. and Wolfson, M. (2007) A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics*, **63**(3), 935–941.

Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, 2937–2960.

Mason, A., Richardson, S., Plewis, I. and Best, N. (2012) Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics*, **28**(2), 279–302.

Mavridis, D., White, I. R., Higgins, J. P. T., Cipriani, A. and Salanti, G. (2015) Allowing for uncertainty due to missing continuous outcome data in pair-wise and network meta-analysis. *Statistics in Medicine*, **34**, 721–741.

National Research Council (2010) *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: Panel on Handling Missing Data

in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press.

Nemes, S., Jonasson, J. M., Genell, A. and Steineck, G. (2009) Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, **9**(1), 56.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994) Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, **42**, 137–160.

Reilly, M. and Pepe, M. S. (1995) A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**(2), 299–314.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Rogers, W. H. (1993) Regression standard errors in clustered samples. *Stata Technical Bulletin*, **13**, 19–23.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93**(444), 1321–1339.

Roy, J. (2003) Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, **59**(4), 829–836.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Scharfstein, D., McDermott, A., Olson, W. and Wiegand, F. (2014) Global sensitivity analysis for repeated measures studies with informative dropout: a fully parametric approach. *Statistics in Biopharmaceutical Research*, **6**(4), 338–348.

Scharfstein, D. O., Daniels, M. J. and Robins, J. M. (2003) Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, **4**, 495–512.

StataCorp (2011) *Stata Statistical Software: Release 12.* College Station, TX: StataCorp LP.

Ware, J. E. (1993) *SF-36 Health Survey: Manual and Interpretation Guide.* Boston: The Health Institute, New England Medical Center.

Ware, J. H. (2003) Interpreting incomplete data in studies of diet and weight loss. *New England Journal of Medicine*, **348**(21), 2136–2137.

West, R., Hajek, P., Stead, L. and Stapleton, J. (2005) Outcome criteria in smoking cessation trials: proposal for a common standard. *Addiction*, **100**, 299–303.

White, I. R. (2015) Sensitivity analysis: The elicitation and use of expert opinion. In *Handbook of Missing Data Methodology* (Eds G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis and G. Verbeke). Chapman and Hall.

White, I. R. and Thompson, S. G. (2005) Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*, **24**(7), 993–1007.

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

White, I. R., Higgins, J. P. T. and Wood, A. M. (2008) Allowing for uncertainty due to missing data in meta-analysis - Part 1: Two-stage methods. *Statistics in Medicine*, **27**(5), 711–727.

White, I. R., Horton, N. J., Carpenter, J. and Pocock, S. J. (2011a) Strategy for intention to treat analysis in randomised trials with missing outcome data. *British Medical Journal*, **342**, d40.

White, I. R., Wood, A. and Royston, P. (2011b) Tutorial in biostatistics: Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*, **30**, 377–399.

White, I. R., Carpenter, J. and Horton, N. J. (2012) Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical Trials*, **9**, 396–407.

Wood, A. M., White, I. R. and Thompson, S. G. (2004) Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, **1**(4), 368–376.

Table 1: QUATRO trial: data summary.

		Intervention (n=204)	Control (n=205)
Centre	Amsterdam (%)	50 (25%)	50 (24%)
	Leipzig (%)	49 (24%)	48 (23%)
	London (%)	45 (22%)	47 (23%)
	Verona (%)	60 (29%)	60 (29%)
MCS at baseline	Mean (SD)	38.4 (11.2)	40.1 (12.1)
	Missing (%)	13 (6%)	10 (5%)
MCS at 1 year	Mean (SD)	40.2 (12.0)	41.3 (11.5)
	> 40 (%)	99 (57%)	104 (54%)
	Missing (%)	29 (14%)	13 (6%)

Table 2: Simulation study: data generating models.

DGM	Type	PM correct?	SM correct?	Auxiliary variables?
1	Pattern-mixture	Yes	Yes	No
2	Pattern-mixture	Yes	Yes	Yes
3	Pattern-mixture	Yes	No	No
4	Selection	No	Yes	No

Parameters	Description
a	Base case
b	Larger sample size
c	More missing data
d	Larger departure from MAR

Table 3: Simulation results: bias, empirical standard error and coverage of nominal 95% confidence interval for methods Full (before data deletion), CC (complete cases analysis), MS (mean score), MI (multiple imputation), SM (selection model + IPW). Error denotes maximum Monte Carlo error.

DGM	Bias					Empirical SE					Coverage				
	Full	CC	MS	MI	SM	Full	CC	MS	MI	SM	Full	CC	MS	MI	SM
1a	0.007	-0.128	0.010	0.010	0.009	0.191	0.223	0.218	0.219	0.218	95.0	90.6	95.1	95.3	95.8
1b	0.004	-0.130	0.007	0.006	0.006	0.095	0.113	0.111	0.112	0.111	94.3	76.7	93.8	93.4	94.6
1c	0.007	-0.167	0.018	0.018	0.018	0.183	0.267	0.258	0.259	0.258	96.5	91.9	95.9	95.4	96.6
1d	0.000	-0.177	0.003	0.004	0.004	0.187	0.223	0.203	0.204	0.203	95.1	88.3	95.6	95.5	97.1
2a	0.009	-0.151	0.010	0.010	0.005	0.178	0.219	0.203	0.203	0.209	96.3	90.1	95.4	95.4	97.0
2b	0.004	-0.157	0.003	0.003	-0.001	0.092	0.109	0.101	0.101	0.104	94.6	71.5	95.1	95.3	96.9
2c	0.005	-0.206	0.002	-0.001	-0.004	0.185	0.296	0.256	0.257	0.290	94.9	88.1	95.0	94.7	97.2
2d	0.011	-0.175	0.011	0.010	0.007	0.186	0.227	0.200	0.200	0.202	95.0	86.1	94.5	95.2	98.2
3a	0.016	-0.110	0.021	0.020	0.015	0.208	0.249	0.245	0.246	0.249	96.5	92.7	95.4	94.7	95.3
3b	0.003	-0.127	0.003	0.003	-0.003	0.106	0.122	0.120	0.120	0.121	94.8	81.8	95.3	95.2	95.3
3c	0.012	-0.160	0.015	0.012	0.018	0.217	0.327	0.316	0.318	0.323	95.1	90.3	95.0	95.1	95.4
3d	0.019	-0.163	0.020	0.020	0.020	0.216	0.254	0.238	0.239	0.249	94.2	88.6	94.5	94.6	95.3
4a	0.005	-0.148	0.014	0.012	0.006	0.218	0.255	0.251	0.252	0.253	94.2	90.0	94.8	94.5	95.0
4b	0.006	-0.147	0.015	0.015	0.008	0.107	0.124	0.122	0.122	0.123	94.0	77.1	94.0	94.1	94.6
4c	0.008	-0.203	0.004	0.002	0.011	0.213	0.345	0.336	0.337	0.343	94.2	90.8	94.2	94.3	95.0
4d	0.026	-0.118	0.045	0.045	0.036	0.211	0.250	0.246	0.248	0.249	94.2	91.2	94.3	94.5	94.8
Error	0.007	0.011	0.011	0.011	0.011	0.005	0.008	0.008	0.008	0.008	0.8	1.4	0.8	0.8	0.7

Figure 1: QUATRO trial: sensitivity analysis for the estimated intervention effect on the MCS (with 95% confidence interval) over a range of departures from MAR.
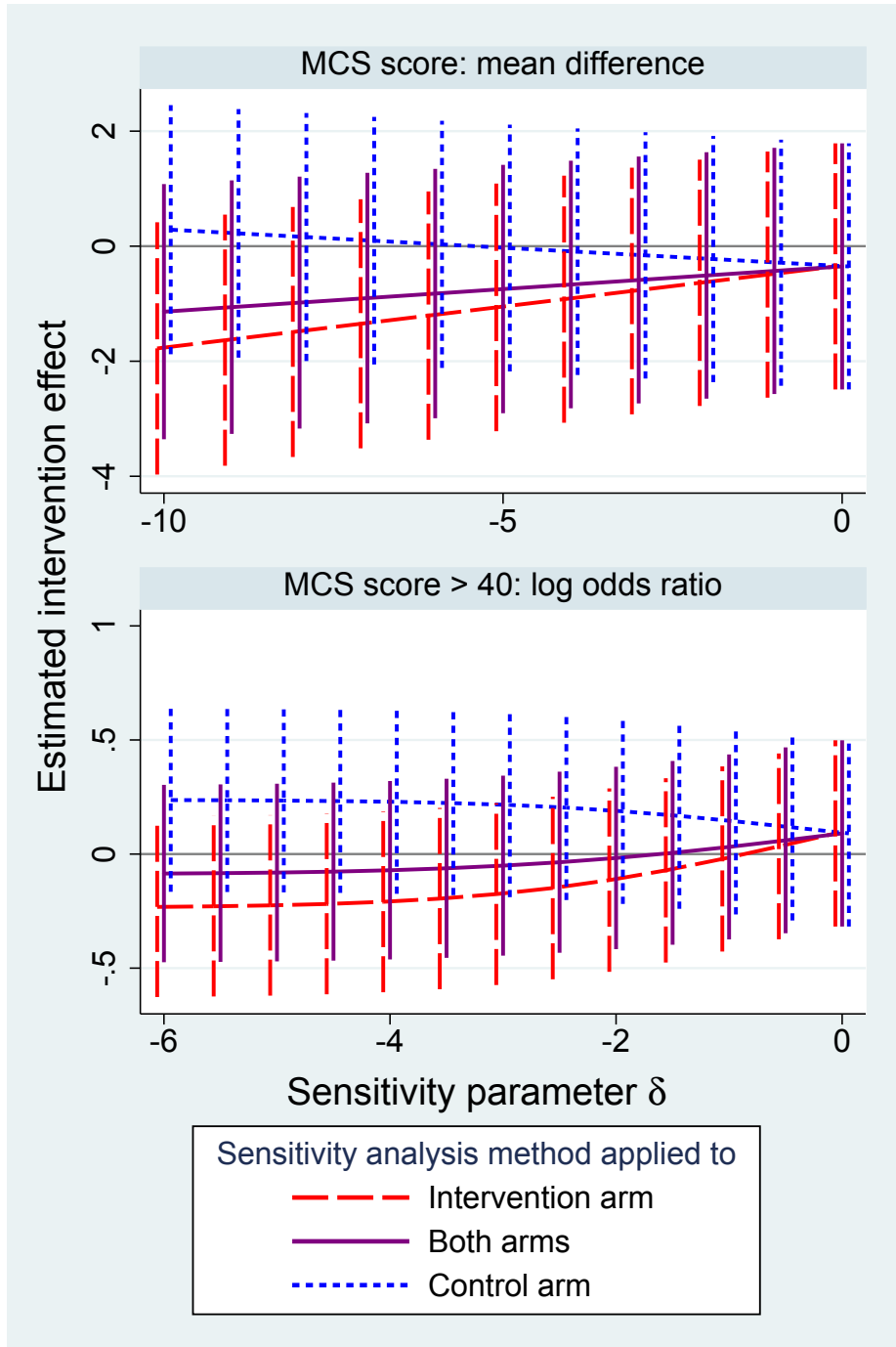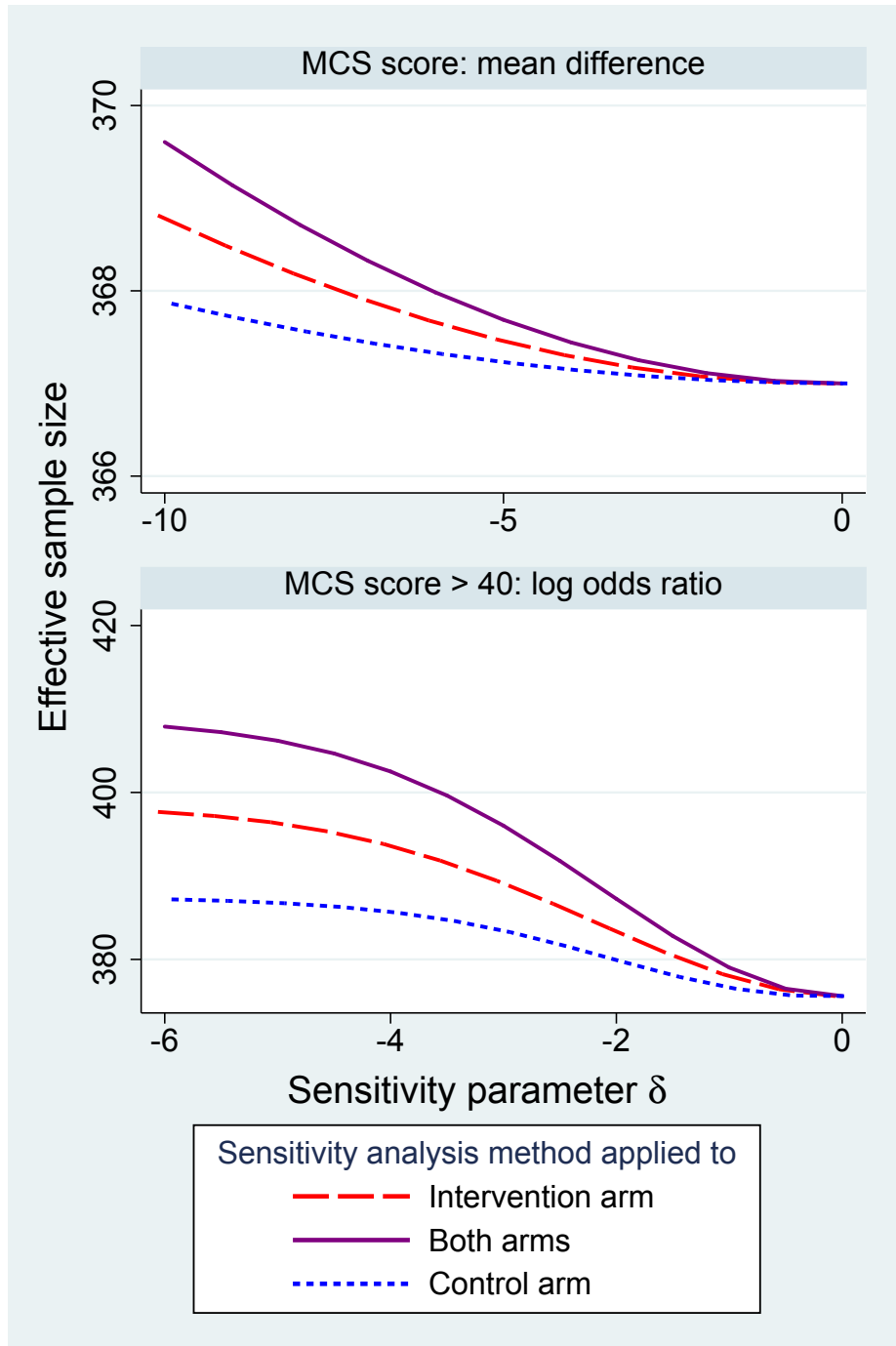
Figure 2: QUATRO data: effective sample size in sensitivity analysis.

Supplementary materials

A Details of mean score variance

The sandwich variance (6) is computed from $\mathbf{B}$ and $\mathbf{C}$, where $\mathbf{B}$ has components

$$
\begin{aligned}
\mathbf{B}_{SS} &= -d\mathbf{U}_S/d\boldsymbol{\beta}_S &&= \sum_i h'(\hat{\boldsymbol{\beta}}_S\mathbf{x}_{Si})\mathbf{x}_{Si}\mathbf{x}_{Si}^T \\
\mathbf{B}_{SP} &= -d\mathbf{U}_S/d\boldsymbol{\beta}_P &&= -\sum_i (1-r_i)h'(\hat{\boldsymbol{\beta}}_P\mathbf{x}_{Pi}+\Delta_i)\mathbf{x}_{Si}\mathbf{x}_{Pi}^T \\
\mathbf{B}_{PS} &= -d\mathbf{U}_P/d\boldsymbol{\beta}_S &&= \mathbf{0} \\
\mathbf{B}_{PP} &= -d\mathbf{U}_P/d\boldsymbol{\beta}_P &&= \sum_i r_i h'(\hat{\boldsymbol{\beta}}_P\mathbf{x}_{Pi})\mathbf{x}_{Pi}\mathbf{x}_{Pi}^T
\end{aligned}
$$

and $\mathbf{C}$ has components

$$
\begin{aligned}
\mathbf{C}_{SS} &= \sum_i e_{Si}^2 \mathbf{x}_{Si}\mathbf{x}_{Si}^T \\
\mathbf{C}_{SP} = \mathbf{C}_{PS}^T &= \sum_i e_{Si}e_{Pi}\mathbf{x}_{Si}\mathbf{x}_{Pi}^T \\
\mathbf{C}_{PP} &= \sum_i e_{Pi}^2 \mathbf{x}_{Pi}\mathbf{x}_{Pi}^T
\end{aligned}
$$

where $e_{Si} = y_i^*(\hat{\boldsymbol{\beta}}_P) - h(\hat{\boldsymbol{\beta}}_S^T\mathbf{x}_{Si})$ and $e_{Pi} = r_i\{y_i - h(\hat{\boldsymbol{\beta}}_P^T\mathbf{x}_{Pi})\}$.

B Modifications for clustered data

If data are clustered, as in a cluster-randomised trial, we need to modify the variance calculations in Sections 2.1 and 2.2 and the small-sample corrections in Sections 3.1 and 3.2. Let m be the total number of clusters, m_{obs} be the number of clusters with at least one observed outcome, and $m_{mis} =$

$m - m_{obs}$ be the number of clusters with no observed outcome. Let the data be subscripted by cluster membership $c = 1, \ldots, m$ as well as individual i.

For the full sandwich variance method of Section 2.1, we only need to re-define the matrix $\mathbf{C} = \sum_c \mathbf{U}_c(\hat{\boldsymbol{\beta}})\mathbf{U}_c(\hat{\boldsymbol{\beta}})^T$ where $\mathbf{U}_c(\hat{\boldsymbol{\beta}}) = \sum_i \mathbf{U}_{ci}(\hat{\boldsymbol{\beta}})$ (Rogers, 1993).

For the two linear regressions method of Section 2.2, we similarly take $\text{var}\left(\hat{\boldsymbol{\beta}}_P\right)$ and $\text{var}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right)$ as clustered sandwich variances.

For the small-sample methods of Section 3, we assume the standard methods use a small-sample correction factor $f = \frac{n-1}{n-p^*}\frac{m}{m-1}$, and use $m-1$ degrees of freedom for linear regression (StataCorp, 2011). We replace n and m by n_{eff} and m_{eff}, calculated by the two methods explained below.

For the full sandwich variance method, we compute n_{eff} as in Section 3.1, and compute $m_{eff} = m_{obs} + (I_{mis}/I_{mis^*})m_{mis}$.

For the two linear regressions method, the variance with small-sample correction is (as before) $\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_P\right) + \widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right) = V_{small}$. The corresponding variance without small-sample correction is $\frac{n_{obs}-p}{n_{obs}-1}\frac{m_{obs}-1}{m_{obs}}\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_P\right) +$ $\frac{n-p}{n-1}\frac{m-1}{m}\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right) = V_{large}$. The heuristic $V_{small} \approx \frac{n_{eff}-1}{n_{eff}-p}\frac{m_{eff}}{m_{eff}-1}V_{large}$ leads to the equation $|V_{small}| = \left(\frac{n_{eff}}{n_{eff}-p}\right)^p |V_{large}|$. However, we have two unknowns n_{eff} and m_{eff}, so we take a second equation representing the variance with small-sample correction only for the number of clusters: $\frac{m_{obs}-1}{m_{obs}}\widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_P\right)+$

$\frac{m-1}{m}\widehat{\mathrm{var}}\left(\hat{\boldsymbol{\beta}}_S - \hat{\boldsymbol{\beta}}_P\right) = V_{largen}$ say, with the heuristic $V_{small} \approx \frac{m_{eff}}{m_{eff}-1}V_{largen}$ and the second equation $|V_{small}| = \left(\frac{m_{eff}}{m_{eff}-1}\right)^p |V_{largen}|$. We solve the second equation for m_{eff} and then the first equation for n_{eff}.