Estimating treatment effects under untestable assumptions with non-ignorable missing data

Manuel Gomes[*] Michael G Kenward[†] Richard Grieve[‡] James Carpenter[†§]

Abstract

Non-ignorable missing data poses key challenges for estimating treatment effects because the substantive model is typically not identifiable without imposing further assumptions. For example, the Heckman selection model has been widely used for handling non-ignorable missing data but requires the study to make correct assumptions, both about the joint distribution of the missingness and outcome and that there is a valid exclusion restriction. Recent studies have revisited how alternative selection model approaches, for example estimated by multiple imputation (MI) and maximum likelihood, relate to Heckman-type approaches in addressing the first hurdle. However, the extent to which these different selection models rely on the exclusion restriction assumption with non-ignorable missing data is unclear.

Motivated by an interventional study (REFLUX) with non-ignorable missing outcome data in half of the sample, this paper critically examines the role of the exclusion restriction in Heckman, MI and full-likelihood selection models when addressing non-ignorability. We explore the implications of the different methodological choices concerning the exclusion restriction for relative bias and root-mean-squared error in estimating treatment effects. We find that the relative performance of the methods differs in practically important ways according to the relevance and strength of the exclusion restriction. The full-likelihood approach is less sensitive to alternative assumptions about the exclusion restriction than Heckman-type models and appears an appropriate method for handling non-ignorable missing data. We illustrate the implications of method choice for inference in the REFLUX study, which evaluates the effect of laparoscopic surgery on long-term quality of life for patients with gastro-oseophageal reflux disease.

Selection models, Heckman model, multiple imputation, full-information maximum likelihood, missing not at random, average treatment effects

[*]Department of Applied Health Research, University College London, London, UK.
[†]Department of Medical Statistics, LSHTM, London, UK.
[‡]Department of Health Services Research and Policy, University College London, London, UK.
[§]MRC Clinical Trials Unit at UCL, London, UK.

1 Introduction

Common approaches taken to handling missing data, such as inverse probability weighting and multiple imputation, assume that the reasons for the missing data are independent of unobserved values, conditional on the observed data. Under this assumption, the missing data mechanism can be said to be missing at random (MAR) or *ignorable*. Once we establish that the probability of observing the data is independent of missing values, and assuming that the model for the observed data is correctly specified, the functional form of the missing mechanism can be 'ignored' when making inferences from the observed data. However, in many cases, the reasons for the missing data are likely to be associated with unobserved values. In such cases, the missingness mechanism is said to be missing not at random (MNAR) or *non-ignorable* and must be modelled together with the substantive model for the observed data.

Non-ignorable missing data is of particular concern in the evaluation of treatment effects because non-response is often associated with the underlying unobserved values of the outcome of interest and tends to differ by treatment group[1,2,3]. This problem is illustrated in the REFLUX study, which evaluated the average treatment effect of laparoscopic surgery (versus usual medical management) for treating patients with gastro-oesophageal reflux disease that adversely affects their well-being. The primary outcome of interest was the patient-reported health-related quality of life (HRQL), measured using the EQ-5D-3L questionnaire, but this was missing for about half of the patients. MNAR was a major concern in this study because the chances of completing the HRQL questionnaire were likely to be related to patient's (unobserved) health status and treatment assignment, after adjusting for the observed data. For example, patients whose outcomes did not improve following medical management (control group) were anticipated to be more likely to drop out of the study or fail to return the health questionnaire. Failure to take into account the contextually plausible missing data mechanism may have resulted in misleading inferences about the treatment effect.

One of the most commonly used approaches to handle non-ignorable missing data in health and social sciences is the Heckman selection model[4]. As with other selection models, Heckman's approach addresses MNAR data by jointly modelling the outcome and missingness models, and typically assuming these are drawn from a bivariate Normal distribution. To avoid some of the problems of direct likelihood maximisation, Heckman proposed a 2-stage least-squares estimation procedure. This involves combining a probit model for the probability of observing the outcome (1st stage) with

a linear regression model for the outcome (2nd stage), which is a function of the estimates obtained in the 1st stage. An alternative estimating approach to selection models is to use a single-step full maximum likelihood to jointly estimate the outcome and missing data models. For example, Diggle and Kenward[5] combined a marginal model for the outcome with a logistic regression for the missing data mechanism, allowing the latter to be a function of the unobserved outcome. This approach requires some form of integration over the unobserved outcomes. The study used the Nelder-Mead optimisation algorithm, but in practice such selection models are often estimated by MCMC techniques in a Bayesian framework [6]. Alternatively, multiple imputation (MI) has been widely recommended for addressing missing data, and it has been recently advocated for handling MNAR mechanisms[7]. Essentially, MI imputes a set of plausible values for each missing observation, which are drawn from the posterior distribution of the missing values given the observed data. To handle MNAR, the imputed values can be drawn from a selection model, such as the Heckman model, to recognise that the missing data may be related to unobserved values.

Non-ignorable missing data poses key challenges for drawing statistical inferences, because the model of interest is typically not identifiable without imposing further assumptions. For example, while a model for the joint distribution of non-response and the partially-observed outcome is required, the 'true' form of the model is typically unknown. Alternative parametric, semi-parametric and non-parametric methods for dealing with this challenge have been subject to extensive debate in the last two decades[8,9,10,11]. Another important assumption that is general to the alternative selection approaches, relates to the exclusion restriction. In this context, a valid exclusion restriction requires the presence of variables that are both predictive of non-response and conditionally independent of the partially-observed outcome. In our motivating example, this would require identifying variables that helped explain questionnaire non-response but were unrelated to patient's health status conditional on their characteristics. The key concern is that these variables may be associated with unobserved prognostic factors that predict both non-response and outcome, i.e the exclusion restriction is generally untestable[1]. With the Heckman approach it is generally recommended that at least one such variable is included in the selection model to help identification[13,14]. However, it is unclear whether the relative performance of this approach depends on the strength of association between these variables and non-response. More generally, the extent to which alternative selection approaches rely on

[1]Mohan and Pearl[12] have shown that at least independence between Z and Y is testable if Z is fully observed and Y has missing values, which is not the same as showing the exclusion restriction is testable, but at least intuitively it would appear that the conditional independence assumption is testable but requires information for every strata of X.

the exclusion restriction is not well understood, and has received little attention in the development of methods for handling MNAR data[15].

The aim of this paper is to investigate the role of the exclusion restriction in Heckman, MI and full-likelihood selection models across a range of typical MNAR mechanisms. In doing so, the paper seeks to clarify the implications of differences in the way this assumption is formulated across alternative methods used to estimate treatment effectiveness. For the purposes of this study, we focused on the MNAR problem in which a variable causes its own missingness; this is sometimes called self-masking missingness. The plan for the remainder of the paper is as follows. In section 2 we describe our motivating example. Section 3 describes each method for estimating selection models and clarifies the underlying exclusion restriction. Section 4 presents the design and results of a simulation study that evaluates the relative merits of the alternative methods across different MNAR settings. Section 5 reports the results from applying the methods to the case-study, and Section 7 discusses the findings and provides directions for further research.

2 Motivating example

Gastro-Oseophageal Reflux Disease (GORD) develops when reflux of the stomach acid causes troublesome symptoms or complications which adversely affect patients' well-being. About 20-30% of adult 'Western' populations experience heartburn or reflux intermittently, and these patients are often treated with Proton Pump Inhibitors (PPIs) to suppress acid reflux. While PPIs are effective, there is the concern that long-term acid suppression with PPIs may be associated with increased risk of chronic hypergastrinaemia and gastric cancer. An alternative to long-term medication is to have laparoscopic surgery, which is a minimally invasive procedure but carries some risk of side effects. Previous studies comparing these interventions suggested that laparoscopic surgery is associated with better health-related quality of life (HRQL) compared to medical management[16]. However, in all these studies HRQL responses were missing for a large proportion of the patients, and the key concern was that the relative effectiveness of laparoscopic surgery compared to medical management may be sensitive to alternative assumptions about the missing data.

Our study is motivated by the REFLUX study, which compares these interventions for treating patients with GORD in the UK, and illustrates the challenges of drawing inferences from the analysis of patient-reported outcome measures (PROMs) when data are liable to be MNAR[2,17]. The REFLUX

study included two components (full details in[16]): a randomised controlled trial in which patients were randomised to surgery and medical management, and a non-randomised study which contrasted those patients who did not wish to be randomised, and were assigned to a policy either of surgery or medical management according to their preferences. Our study focused on the non-randomised comparison as the patients who did not want to be randomised are likely to be more typical of those patients with GORD in routine clinical practice. For example, current NICE guidelines for the management of GORD recommend providing GORD treatment according to patient preferences. The non-randomised study included 261 patients in the group with a preference for surgery and 192 in the group with a preference for medical management. Self-reported HRQL was measured at baseline, 3 months and then annually up to 5 years, using the EQ-5D-3L questionnaire[18]. The HRQL endpoint was then combined with survival to report quality-adjusted life-years (QALYs) at 5 years. The parameter of interest was the treatment effect of surgery on the QALY, which had a causal interpretation under the 'no unobserved confounding' assumption.

A significant proportion of patients failed to complete the EQ-5D-3L questionnaire (Table 1). This proportion increased over time, resulting in missing 5-year QALYs for 55% (106 out of 192) of patients in the medical management group and 48% (125 out of 261) in the surgery group. Baseline covariate information was mostly complete. Among the complete cases in the surgery group, average EQ-5D-3L increased rapidly in the first 3 months after the operation (0.68 to 0.80) but then remained very similar up to 5 years. For the medical management group, average EQ-5D-3L decreased slightly up to 2 years and then increased in the last 3 years of follow-up. Five years after the intervention, complete cases in the surgery group had, on average, slightly higher QALYs compared to that in the medical management group (unadjusted differences). A key concern to the study investigators was whether the increase in mean EQ-5D-3L in the latest part of the follow up for the control group was driven by the fact that patients in worse health might have dropped out of the study.

Table 2 provides a description of the main baseline covariates and their association with the missingness indicator (R). One set of baseline variables (X) were imbalanced between the comparison groups, and another set of variables (Z) were predictive of the missingness, but were assumed to be conditionally independent of the outcome. That is, these variables (Z) were assumed to meet the criteria for the exclusion restriction. The probability of whether or not a patient had missing outcome was positively (and significantly) associated with both clinical (physical symptom score, previous

Table 1: Mean (SD) health-related quality of life (measured as EQ-5D-3L) over time and quality-adjusted life years (QALYs) by intervention group.

	Medical management (N=192)		Surgery (N=261)	
	Complete data, N (%)	Mean (SD)	Complete data, N (%)	Mean (SD)
EQ-5D-3L				
Baseline	186 (97%)	0.750 (0.22)	253 (97%)	0.682 (0.26)
3 months	181 (94%)	0.763 (0.23)	232 (89%)	0.806 (0.25)
Year 1	181 (94%)	0.740 (0.25)	232 (89%)	0.791 (0.26)
Year 2	156 (81%)	0.736 (0.24)	203 (78%)	0.796 (0.26)
Year 3	159 (83%)	0.763 (0.23)	196 (75%)	0.803 (0.25)
Year 4	142 (74%)	0.773 (0.21)	168 (64%)	0.806 (0.25)
Year 5	136 (71%)	0.794 (0.21)	176 (67%)	0.800 (0.25)
QALY (5-year)	106 (55%)	3.594 (0.83)	125 (48%)	3.777 (0.94)

Notes: The EQ-5D-3L is a health-related quality of life measure anchored on a scale that includes 0 (death) and 1 (perfect health)

gastro-oseophageal hernia) and socio-demographic (age, education) baseline characteristics.

The validity of the exclusion restriction cannot generally be tested from the data, but can be supported or refuted by external evidence, for example from expert opinion. In the REFLUX study, the clinical investigators suggested that the number of patients recruited (centre size), and the patient's general views about medicine might well be conditionally independent of the patient's health status, conditional on the observed data (e.g. patient characteristics). For example, hospitals with higher recruitment rates were anticipated to be more actively engaged in data collection (e.g. by reminding patients to return questionnaires). Each of the centres were experienced in providing both surgery and medical management to patients with GORD, and so there was no obvious concern that the number recruited would have a direct effect on health status. In addition, the outcome was not expected to have a direct effect on the exclusion restriction variables (reverse causality). Hence, while the 'conditional independence' assumption could not be tested, expert opinion did support this criterion for the exclusion restriction.

However, Table 2 also suggests that the association between these variables and non-response was relatively low (not statistically significant at 5% level). This raises a pertinent question that we sought to address in this study: whether the relative merits of alternative selection models differ according to the strength of association between the exclusion restriction variables and non-response, that is the extent to which the relevance assumption is met (see section 3.1 below). Using the REFLUX study, we then illustrate whether inferences about the effectiveness of laparoscopic surgery versus medical management are sensitive to the choice of selection model, and the extent to which it relies on the

exclusion restriction.

3 Methods

3.1 The exclusion restriction

Let Y_{1i} be a continuous outcome for individual i, and R_i an indicator of whether Y_{1i} is observed ($R_i = 1$) or missing ($R_i = 0$). Let Y_{2i} be a continuous latent variable representing the missingness process, whereby R_i equals to 1 if $Y_{2i} > 0$, and 0 otherwise. In addition, let X_{1i} be the set of prognostic variables, and X_{2i} the set of variables that predict missingness (selection), as described below

$$\begin{aligned} Y_{1i} &= \beta_1 X_{1i} + e_{1i} \\ Y_{2i} &= \beta_2 X_{2i} + e_{2i} \end{aligned} \qquad R_i = \begin{cases} 1, & \text{if } Y_{2i} > 0. \\ 0, & \text{if } Y_{2i} \leq 0. \end{cases} \tag{1}$$

where β_1 and β_2 are the vector of regression coefficients in the outcome (Y_1) and missingness (Y_2) models, respectively. Now assume that X_{2i} includes the prognostic variables (X_{1i}) and other variables Z_i, such that $X_{2i} = (X_{1i}, Z_i)$. Then Z_i is a valid exclusion restriction if it satisfies:

1. $cov(Z_i, Y_{1i}|X_{1i}) = 0$ (conditional independence assumption). Z_i does not either have a direct effect on Y_{1i}, or any effect through omitted variables. Any reverse effect of Y_{1i} on Z_i must also be ruled out.
2. $cov(Z_i, Y_{2i}|X_{1i}) \neq 0$ (relevance assumption). Z_i must independently predict Y_{2i}.

An invalid exclusion restriction may arise if either the conditional independence assumption or the relevance assumption (or both) are not met. For the purposes of the simulation study, we assumed that the conditional independence assumption was met. We assessed the relative performance of the methods according to whether or not the relevance assumption was met, and according to alternative strengths of the exclusion restriction. Within the REFLUX case study we critically assessed the plausibility of both assumptions.

3.2 Heckman's 2-step approach

The Heckman selection model[4] allows for the non-ignorable missing data by jointly estimating Y_1 and Y_2, typically assuming that these follow a bivariate Normal distribution,

Table 2: Descriptive statistics of the baseline prognostic factors, the exclusion restriction variables, and their correlation with the missingness indicator (R).

Variable	Medical management (N=192)	Surgery (N=261)	Standardised difference (%)	Correlation with R[†]
Baseline prognostic factors				
Male	111 (58%)	170 (65%)	15.1	0.06
Age	49.9 (11.8)	44.4 (12.0)	45.9	0.18***
BMI (kg/m2)	27.4 (4.1)	27.7 (3.9)	7.5	-0.05
REFLUX quality-of-life	76.1 (19.8)	55.9 (22.8)	94.7	0.06
Baseline EQ-5D-3L	0.75 (0.22)	0.68 (0.26)	27.5	0.06
Heart burn score	72.2 (20.9)	49.4 (24.2)	101	0.08
Gastro 1 symptom score	59.3 (22.2)	47.2 (21.1)	55.8	0.06
Gastro 2 symptom score	82.9 (17.5)	75.9 (21.7)	35.3	-0.01
Nausea symptom score	89.4 (13.5)	77.0 (19.7)	73.8	0.12
Activity symptom score	86.6 (12.8)	74.5 (15.9)	83.4	0.02**
Previous hiatus hernia	73 (38%)	76 (29%)	18.9	-0.09**
Smoker	39 (20%)	71 (27%)	16.2	0.1
Asthma	36 (19%)	30 (11%)	20.4	-0.04
Duration of REFLUX symptoms (days)	45.9 (53.7)	55.9 (67.3)	16.4	-0.03
Employment status				0.01
Full-time	101 (53%)	171 (66%)	26.5	
Part-time	20 (10%)	35 (13%)	9.3	
School leaving age				0.10**
16 year or younger	107 (56%)	154 (59%)	6.6	
20 years or older	40 (21%)	44 (17%)	10.2	
Exclusion restriction variables				
Centre size[‡]	27.1 (10.7)	27.7 (9.4)	6.2	0.09*
General views about medicine				
Doctors use too many medicines	32 (17%)	61 (23%)	16.8	0.03
People should pause treatments	42 (22%)	76 (29%)	16.7	-0.01
Medicines are addictive	22 (11%)	39 (15%)	10.3	-0.07*
Natural remedies are safer	30 (16%)	38 (15%)	3	0.07
Medicines do more harm than good	4 (2%)	5 (2%)	1.1	-0.02
All medicines are poisons	13 (7%)	7 (3%)	19.3	-0.03
Doctors trust medicines too much	26 (14%)	51 (20%)	16.2	0.09*
Doctors should spend more time with patients	69 (36%)	94 (36%)	0.2	0.06

Notes: Continuous covariates reported as Mean (SD) and binary covariates as N (%). Belief variables are dichotomised: 1 if patient agrees or strongly agrees with the statement, 0 otherwise.

[‡] Number of patients recruited per centre within the non-randomised study. [†] Pearson correlation coefficient. Statistical significance is based on the corresponding coefficients from the logistic model: *p<0.1, **p<0.05 ***p<0.001

$$\begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ & \sigma_2^2 \end{pmatrix} \right] \tag{2}$$

To help identification, the variance of the latent variable (σ_2^2) is set to 1. The correlation parameter (ρ) governs the strength of the MNAR mechanism. Under bivariate Normality, the conditional mean for the observed Y_1, given X_1 and X_2, can be expressed as:

$$E(Y_{1i}|X_{1i}, Y_{2i} > 0) = \beta_1 X_{1i} + \rho\sigma_1\lambda_i, \qquad \lambda_i = \frac{\phi(\beta_2 X_{2i})}{\Phi(\beta_2 X_{2i})} \tag{3}$$

Where λ is denoted as the inverse Mills ratio, ϕ is the standard Normal density and Φ is the standard Normal cumulative distribution function. To estimate the parameters of interest (β), Heckman proposed a two-step estimation approach[4]:

1. Regress Y_2 on X_2 by applying a probit model to the full sample to obtain estimates of $\hat{\beta}_2$ and construct $\hat{\lambda}_i$.
2. Estimate parameters of interest (e.g. $\hat{\beta}_1$) by applying a linear regression model on the observed sample:
   $Y_{1i} = \beta_1 X_{1i} + \beta_\lambda \hat{\lambda}_i + e_{1i}$.

The 2-step approach allows a direct test as to whether the data are MNAR (i.e. whether $\beta_\lambda \neq 0$). However, it should be noted that the validity of this test relies heavily on the parametric assumptions underlying both the substantive and missing data models (e.g. linearity and Normality). We will illustrate this in section 6. In addition, given that λ is estimated rather than known in step 2), e_{1i} will be heteroscedastic, and estimates of $\hat{\sigma}_1^2$ will not be valid. Heckman proposed the following consistent variance estimator:

$$Var(e_{1i}|X_{1i}, Y_{2i} > 0) = \sigma_1^2(1 - \rho^2(\lambda_i^2 + \beta_2 X_{2i}\lambda_i)) \tag{4}$$

In settings where all the variables in X_2 are predictive of Y_1 (i.e. X_2 and X_1 are the same), we say that the 'exclusion restriction' is not met. In this case, equation (3) depends critically on the non-linear form of the inverse Mills ratio (λ). As $\lambda_i(\beta_2 X_2)$ is approximately linear, this tends to cause collinearity issues[13]. To provide stable, precise estimates, the 2-step estimator in practice requires that X_2 includes at least one variable, Z, that predicts Y_2 but is conditionally independent of Y_1 given X_1[14].

3.3 Full maximum likelihood approach

An equivalent way of specifying the selection model is to directly include the outcome in the selection equation (see full derivation in Appendix A of the Online Supplementary Material):

$$Pr(Y_{2i} > 0|X_{1i}, X_{2i}, Y_{1i}) = \Phi(\beta_2 X_{2i} + \beta_y Y_{1i}) \tag{5}$$

Given that Y_1 is partially observed, we cannot directly estimate model (5) from the data by maximum likelihood. Full likelihood approaches will typically require this model to be jointly estimated with the outcome model. At this point, we should note that instead of a probit function, one can equally combine a logistic regression for the Y_2 with a marginal regression model for Y_1[5].

We now write the log likelihood function, which is a combination of the joint probability of Y_1 and $Y_2 > 0$, with the marginal probability that $Y_2 \leq 0$ (further details are provided in Appendix B, Supplementary Material):

$$\begin{gathered} log\ L\left(\beta_1, \beta_2, \sigma_1, \rho\right) = \sum_{Y_2 \leq 0} log\ \left[1 - \Phi\left(\beta_2 X_{2i}\right)\right] + \\ \sum_{Y_2 > 0} \left[- log\ \sigma_1 + log\ \phi\left(\frac{Y_{1i} - \beta_1 X_{1i}}{\sigma_1}\right) + log\ \Phi\left(\frac{\beta_2 X_{2i} + \frac{\rho}{\sigma_1}(Y_{1i} - \beta_1 X_{1i})}{\sqrt{1-\rho^2}}\right)\right] \end{gathered} \tag{6}$$

where ϕ and Φ are defined above. Recommendations for maximum likelihood estimators suggest that the joint model includes those variables that are associated with the outcome, even if these are not independent predictors of the missing data[14]. However, the role that variables that are not predictive of the outcome (such as Z_i) play in the selection model in unclear[19,14]. We investigate this issue in our simulation study.

3.4 Multiple imputation based on Heckman's 2-step approach

While multiple imputation (MI) is a recommended approach for handling missing data in many settings, it typically relies on the validity of the MAR assumption[20]. In this section, we introduce an imputation model based on the Heckman's selection model that allows for MNAR mechanisms that are compatible with the Heckman model[7]. This approach uses predictions from Heckman's first step (probit model) to develop the imputation model. In other words, the imputed values are drawn from the posterior conditional distribution of missing values given the observed data, and the unobserved

determinants of missingness via the inverse Mills ratio:

$$Y_{1i}^{miss} \sim N\Big(\beta_1 X_{1i} + \beta_\lambda \lambda_i^{'},\ \sigma_e^2\Big) \tag{7}$$

where $\lambda_i^{'} = \frac{-\phi\ (\beta_2 X_{2i})}{1-\Phi\ (\beta_2 X_{2i})}$ is the inverse Mills ratio derived from the conditional expectation of $Y_{2i} \leq 0$. The main steps of the imputation are as follows:

1. Fit a probit model to Y_{2i} (Heckman's approach, step 1) and compute $\hat{\lambda}_i^{'} = \frac{-\phi\ (\hat{\beta}_2 X_{2i})}{1-\Phi\ (\hat{\beta}_2 X_{2i})}$
2. Fit a OLS regression to the observed Y_{1i} (Heckman's approach, step 2) to estimate $\hat{\beta}_1$, $\hat{\beta}_\lambda$ and $\hat{\sigma}_e$
3. Compute Bayesian posterior draws for β_1^*, β_λ^* and σ_e^*, in the standard way for MI with this particular linear outcome model[20] (page 77-89).
4. Draw e_1 from $N \sim (0, \sigma_e^{2*})$.
5. For each missing observation (Y_{1i}^{miss}), impute Y_1^* using model (7): $Y_{1i}^* = \beta_1^* X_{1i} + \beta_\lambda^* \hat{\lambda}_i^{'} + e^*$.
6. Repeat steps 1) to 5) M times to obtain M imputed datasets.

Then for each imputed dataset, we fit the outcome model of interest, and the resultant estimates $(\hat{\beta}_1, \hat{\sigma}_e)$ can be combined using Rubin's formulae[21].

With standard MI, if X_2 includes more variables than X_1, this will improve the precision of the estimates provided that these variables are predictive of the outcome[19,22]. However, the extent to which this Heckman-based MI approach relies on the exclusion restriction, and more specifically the relevance assumption, is unclear. We hypothesise that the this imputation approach may be less reliant on the exclusion restriction compared to the original 2-step Heckman[23], while being more robust to departures from bivariate normality compared to the full maximum likelihood approach. We now investigate the implications of the alternative exclusion restriction assumptions, for the bias and efficiency in the estimates of treatment effectiveness following the alternative selection approaches.

4 Simulation study

4.1 Data-generating process

The data generating process was informed by our motivating example and previous empirical studies [24,25] in order to reflect a wide range of non-ignorable missing data settings that could arise in prac-

tice. For example, we considered scenarios that differ according to MNAR mechanism, proportion of missing data, strength of the exclusion restriction and distribution of the outcome data. Overall, we simulated the joint model[2] for outcome and missing data as a product of the corresponding conditional and marginal models[26,27].

Let Y be a partially observed continuous outcome, X a continuous prognostic factor, and Z a exclusion restriction variable. These variables were defined as follows:

$$\begin{aligned} Z &\sim N(0,\ 1) \\ X &\sim N(0.2 + \alpha Z,\ 1) \\ Y &\sim N(0.1 + 0.1X,\ 1) \end{aligned} \tag{8}$$

We assumed the same model parameters in (8) across all scenarios, except α (see scenarios A to B below). For simplicity, we assumed linear additive outcome models relating X to Y and Z to X throughout. Throughout we assumed that the CIA assumption was met for the exclusion restriction variable, i.e. Z is independent of Y conditional on X. The parameter of interest was the true treatment effect represented by the effect of X on Y (true value is 0.1). To further investigate the role of exclusion restriction assumption in settings with departures from the bivariate Normal distribution, we have considered scenarios with a skewed (Gamma-distributed) outcome: $Y \sim G(\mu_y = 0.1 + 0.1X,\ \eta)$, where the mean ($\mu$) and skewness ($\eta$) are simple functions of the usual shape and scale parameters.

Next we describe the framework to simulate the missing data. Let $P(R = 1)$ denote the probability that the response is missing. We simulated MNAR mechanisms as

$$\text{probit}P(R = 1|X, Z, Y) = \theta_0 + \theta_1 X + \theta_2 Z + \theta_3 Y \tag{9}$$

where the probability of missingness may be a function of the outcome as well as the prognostic factor and exclusion restriction. We fixed the value of θ_1 so that the correlation of R with X was about 0.3. We varied θ_0 across scenarios to allow for different proportions of missing data. We have considered a range of values fro θ_2 and θ_3 to reflect alternative strengths of association between R and Z, and between R and Y, respectively, where $\theta_2 = \frac{\sigma_R}{\sigma_Z}\sqrt{\frac{\varphi^2}{1-\varphi^2}}$ and $\theta_3 = \frac{\sigma_R}{\sigma_Y}\sqrt{\frac{\rho^2}{1-\rho^2}}$. In the scenarios with

[2]We have also considered an alternative data generating process whereby outcome and missing data were simulated directly from a joint distribution (e.g. bivariate normal). We found that this data generating approach resulted in similar findings.

Gamma outcome, $\theta_3 = \frac{\sigma_R}{\mu_Y}\sqrt{k\frac{\rho^2}{1-\rho^2}}$, where $k = \frac{\mu_Y^2}{\sigma_Y^2}$ is the shape parameter. Through this simulation framework, we are able to control for: 1) the strength of MNAR by varying ρ; 2) the presence/absence and strength of the exclusion restriction using θ_2; and 3) the distributional assumptions by modifying the marginal model; while maintaining a distinct conditional model that incorporates the exclusion restriction.

We considered four broad MNAR mechanisms, described in Figure 1, which represented alternative forms of the exclusion restriction. Our starting point is that any of the proposed selection approaches obtains formal identification from the normality assumption. MNAR 1 and 2 considered scenarios with an invalid exclusion restriction (relevance assumption is not met, i.e. Z does not have an independent effect on R), whereas MNAR 3 and 4 included possible settings with a valid exclusion restriction (relevance assumption is met). Within each of these mechanisms, we considered four distinct scenarios of increasing complexity:

- Scenario A: % missing data is 'low' (20%); correlation between Y and R is 0.2 ('weak' MNAR); and correlation between Z and X was 0.7.
- Scenario B: % missing data is 'low' (20%); correlation between Y and R is 0.4 ('strong' MNAR); and correlation between Z and X was 0.7.
- Scenario C: % missing data is 'high' (40%); correlation between Y and R is 0.4 ('strong' MNAR); and correlation between Z and X was 0.7.
- Scenario D: % missing data is 'high' (40%); correlation between Y and R is 0.4 ('strong' MNAR); and correlation between Z and X was 0.3.

Further, within the scenario MNAR4-D with a valid exclusion restriction we considered two sub-scenarios: i) different associations between the Z and R to represent alternative strengths of the exclusion restriction: 'weak' ($cor(Z,R) = 0.1$), 'moderate' ($cor(Z,R) = 0.3$) and 'strong' ($cor(Z,R) = 0.5$); and ii) scenarios with skewed outcome data, considering gamma distributed outcome with increasing levels of the skewness parameter: 0.5 (slightly skewed), 1, 1.5, and 2 (extremely skewed).

4.2 Implementation

This simulation study compared the following six approaches:

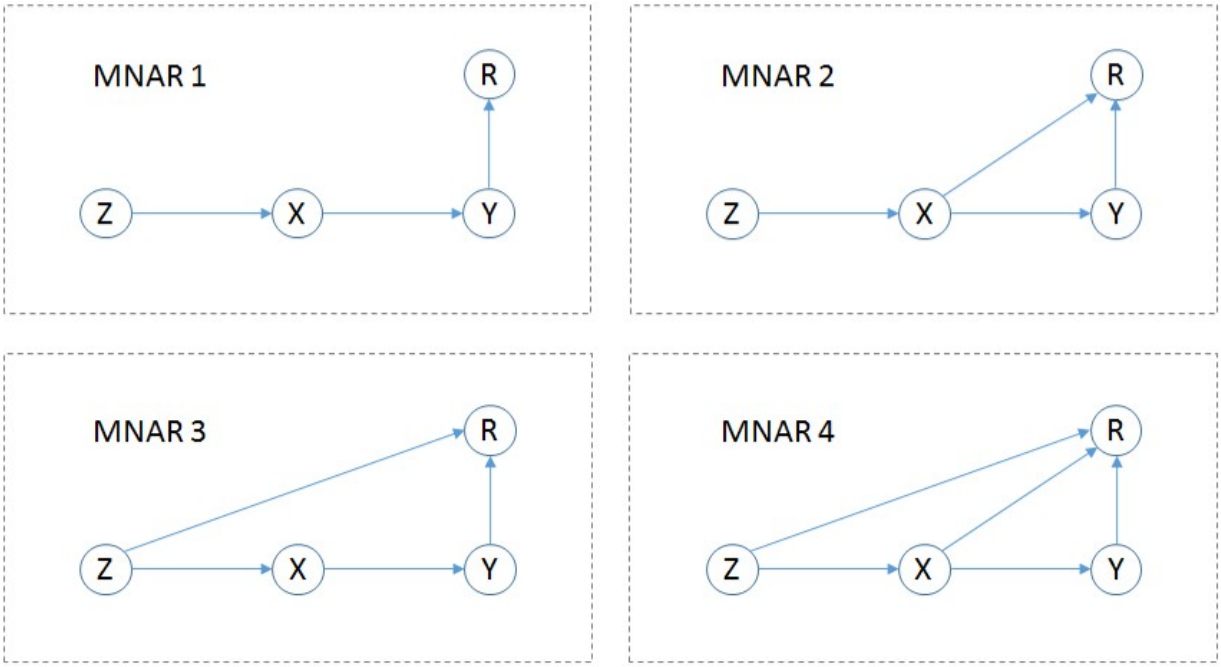

Figure 1: Different missing not at random (MNAR) mechanisms considered in the simulation study. Y is missing, X and Z are fully observed.

1. Full-data analysis (True values): analysis before any data deletion ('benchmark' for the other methods).
2. Complete-case analysis (CCA): analysis based on the individuals with complete outcome data
3. Standard MI assuming MAR (MI-MAR)
4. Standard 2-step Heckman selection model (HECK)
5. Multiple imputation based on the Heckman model (MI-H)
6. Full maximum likelihood selection model (FML)

For each scenario, we applied the methods to 1000 simulated datasets, each with 1000 individuals, and compared the bias, root mean squared error (rMSE), and confidence interval (CI) coverage for estimating the treatment effect, β_1 in the following linear outcome model:

$$Y_{1i} = \beta_1 X_{1i} + e_{1i} \tag{10}$$

Standard MI (under MAR) was estimated using the chained-equations approach[22] with 10 imputations and 10 iterations between each imputation, and was implemented in R using the $mice$ package. The algorithm to implement the MI based on the Heckman model is described in section 3.4 (R code

available from the corresponding author). After imputation, we applied the analysis model (10) to each imputed dataset using the glm package in R. The standard two-step Heckman approach was implemented using the $sampleSelection$ package in R[28]. This packages produces Heckman's consistent variance estimator as per equation (4). The full maximum likelihood selection model was estimated using MCMC methods[29]. We considered 50,000 MCMC iterations, after which convergence was good for both regression coefficients and variance/covariance parameters (Gelman-Rubin scale reduction factor was smaller than 1.1), and assumed vague priors throughout. This approach was implemented in JAGS, which can be used via interface with R $rjags$ package[30].

5 Results

Table 3 shows results for scenarios where the probability of missingness depends only on the outcome Y (MNAR1 in Figure 1). Both CCA and MI under MAR provided biased estimates and coverage below the nominal level (95%) across all scenarios. In these scenarios, MI under MAR (MI-MAR) performs similarly to CCA because both X and Z are not predictive of R.

The 2-step Heckman approach led to biased estimates (9-13% bias) and CIs that were too wide (CI coverage around 0.99) across all MNAR 1 scenarios. The MI approach based on the Heckman model (MI-Heckman) provided similar CI coverage but lower percent bias and rMSE compared to the 2-step Heckman approach across these scenarios. The full maximum likelihood selection model provided the lowest bias and rMSE, and CI coverage close to nominal levels.

Figure 2 shows the distribution of the parameter estimates of interest ($\hat{\beta}_1$) across the 1000 simulations in MNAR 2 and MNAR 4 settings (results for the MNAR 3 setting are very similar to those of MNAR 4, and are available in Appendix C, Supplementary Material). In MNAR 2 settings where missing data is related to X and Y but not Z, both the 2-step Heckman model and the MI-Heckman led to biased results and highly variable estimates, particularly for scenarios C and D, which included large proportions of missing data. The full-likelihood selection model provided unbiased, precise estimates across all MNAR 2 scenarios. In MNAR 4 scenarios with a valid exclusion restriction, all selection models provided unbiased results, although the full likelihood selection model provided estimates that were slightly more precise, and hence provided the lowest rMSE (see Figure 3). The relative superior performance of the full likelihood approach was observed across alternative strengths of the exclusion restriction (full results in Appendix D, Supplementary Material). In

Table 3: Percent bias, rMSE, and confidence interval coverage for the estimated treatment effect (true β_1 is 0.1) across the alternative methods when the probability of missingness depends only on the outcome Y (MNAR1).

Scenario	% Missing	cor(Y, R)	cor(Z, X)	Method	$\hat{\beta}_1$	Bias (%)	Coverage	rMSE
MNAR1 A	20	0.2	0.7	TRUE	0.099	1%	0.947	0.023
				CCA	0.095	5%	0.939	0.025
				MI-MAR	0.095	5%	0.942	0.026
				HECK	0.090	10%	0.998	0.155
				MI-H	0.096	4%	0.998	0.089
				FML	0.099	1%	0.945	0.026
MNAR1 B	20	0.4	0.7	TRUE	0.099	1%	0.947	0.023
				CCA	0.087	13%	0.917	0.027
				MI-MAR	0.087	13%	0.913	0.028
				HECK	0.088	12%	0.991	0.151
				MI-H	0.090	10%	0.993	0.107
				FML	0.098	2%	0.945	0.026
MNAR1 C	40	0.4	0.7	TRUE	0.099	1%	0.947	0.023
				CCA	0.084	16%	0.902	0.031
				MI-MAR	0.084	16%	0.901	0.032
				HECK	0.087	13%	0.994	0.253
				MI-H	0.093	7%	0.991	0.118
				FML	0.098	2%	0.945	0.029
MNAR1 D	40	0.4	0.3	TRUE	0.099	1%	0.947	0.023
				CCA	0.083	17%	0.915	0.040
				MI-MAR	0.084	16%	0.924	0.041
				HECK	0.091	9%	0.990	0.359
				MI-H	0.095	5%	0.997	0.129
				FML	0.096	4%	0.934	0.040

Notes: TRUE: full-data analysis (True values), CCA: complete-case analysis, MI-MAR: multiple imputation assuming MAR, HECK: 2-step Heckman model, MI-H: multiple imputation based on the Heckman model, FML: full maximum likelihood selection model.

particular, when the correlation between the exclusion restriction variable and non-response was low ($corr(R, Z) = 0.1$), the Heckman-based approaches provided higher bias and rMSE compared to the full-likelihood approach. While all methods provided unbiased estimates when the strength of association was moderate ($corr(R, Z) = 0.3$) or strong ($corr(R, Z) = 0.5$), the full likelihood approach provided the lowest rMSE.

Figure 4 reports rMSE for each of the selection models across different (MNAR 4) scenarios with increasing levels of skewness and alternative strengths of the exclusion restriction. In scenarios with a 'weak' (correlation between Z and R was about 0.1) exclusion restriction, the 2-step Heckman approach and the MI-Heckman led to substantially higher rMSE compared to the full likelihood selection model. The high rMSE of the Heckman-based approaches was driven mostly by the large

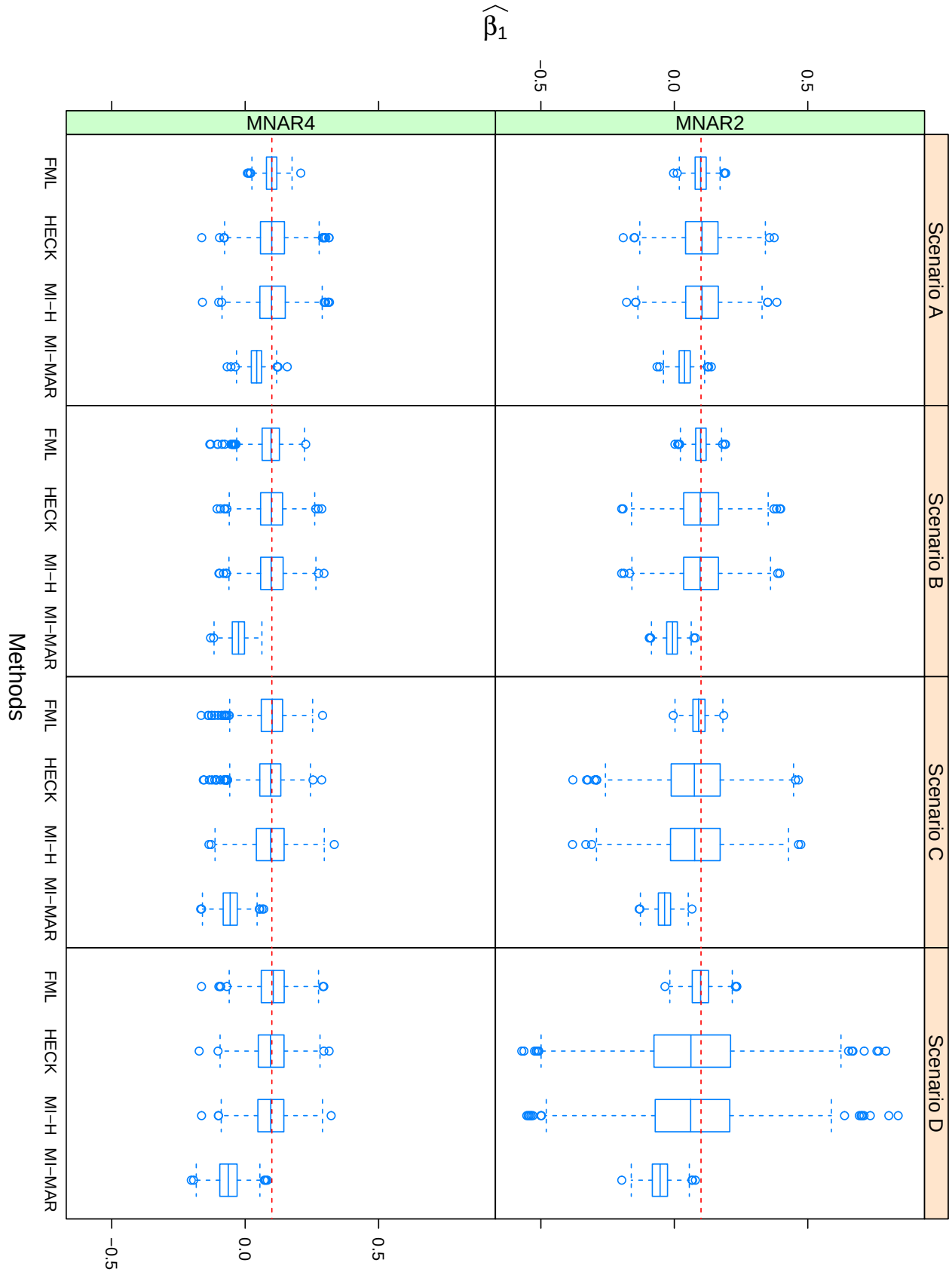

Figure 2: Estimated parameter of interest ($\hat{\beta}_1$) according to method for MNAR2 and MNAR4 scenarios. The boxplots show bias and variation, as median, quartiles and 1.5 times interquartile range for the estimated parameter across the 1000 replications. The dashed lines are the true values. MI-MAR: multiple imputation assuming MAR, HECK: 2-step Heckman model, MI-H: multiple imputation based on the Heckman model, FML: full maximum likelihood selection model

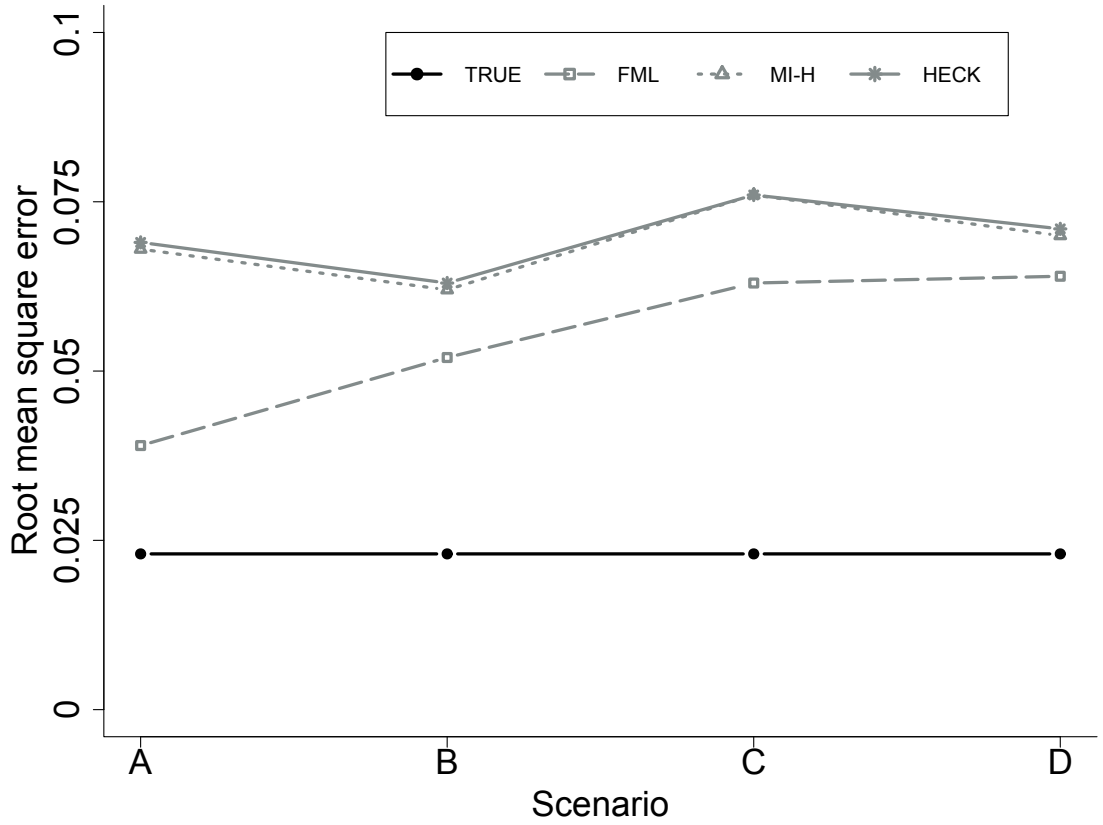

Figure 3: Root mean square error for each selection model in MNAR4 scenarios. TRUE: true values (from applying the substantive model to the fully observed data), HECK: 2-step Heckman model, MI-H: multiple imputation based on the Heckman model, FML: full maximum likelihood selection model

standard errors of the bias, whereas the rMSE of the full likelihood approach followed more closely the absolute level of bias. When the strength of the exclusion restriction was moderate or strong, both the 2-step Heckman model and the MI-Heckman approach provided more stable estimates, and hence lower rMSE compared to the full likelihood approach.

6 Application to REFLUX data

To investigate the implications of method choice in practice, we now apply each of the methods to the REFLUX case study, given the potential concerns about whether the possible exclusion restrictions are valid in this study. Under the bivariate Normality assumption, there seems to be some evidence that QALYs were MNAR. The coefficient (SE) of the inverse Mills ratio (β_λ in equation 3) from the two-step Heckman approach was $-0.13(0.056)$, and the correlation between the residuals of the outcome and missing data models was $\rho = -0.20$. This suggests that the probability of completing

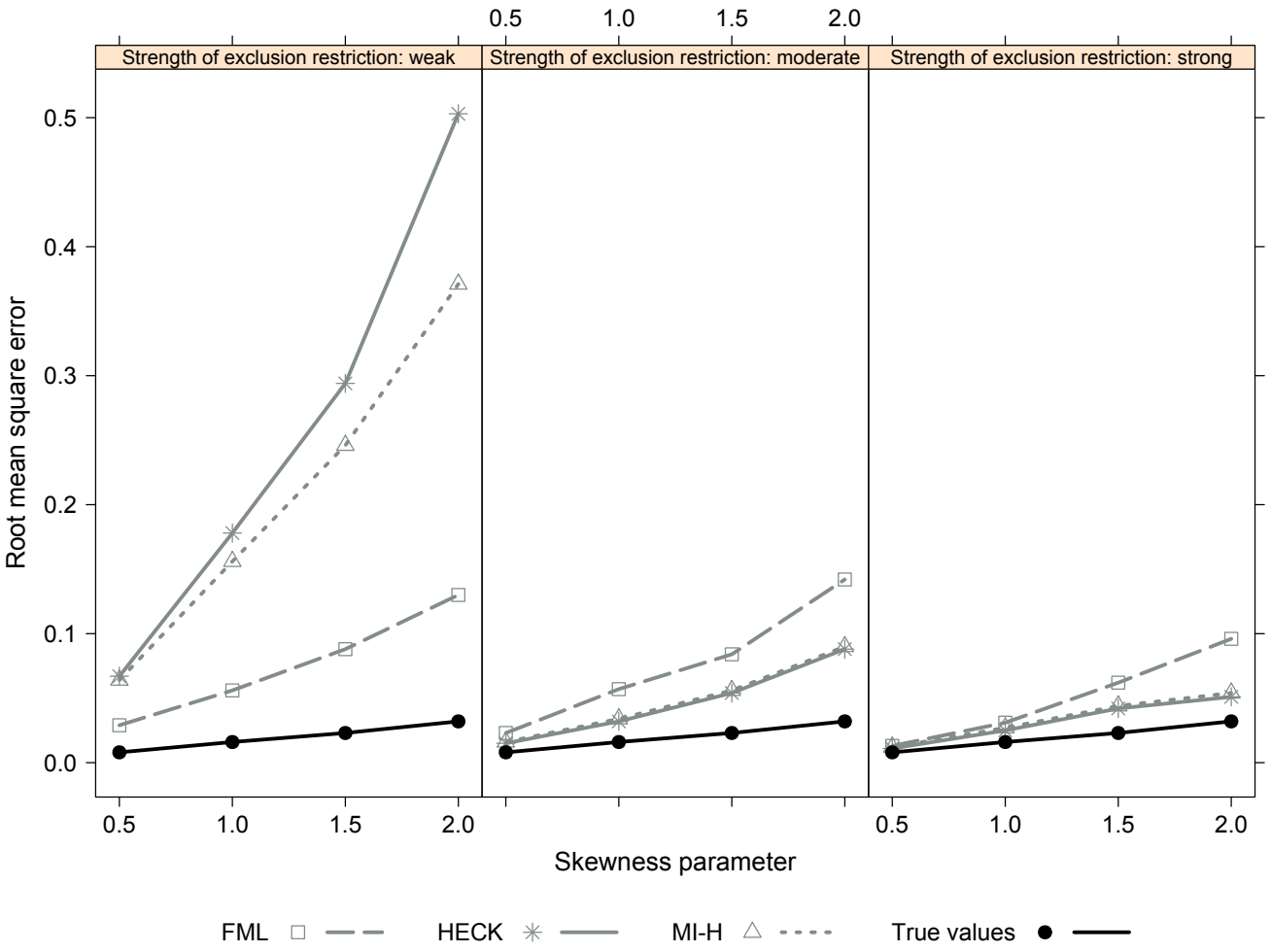

Figure 4: Root mean square error for each method across scenarios that vary according to the strength of the exclusion restriction and skewness parameter (Gamma distributed outcome). TRUE: true values (from applying the substantive model to the fully observed data), HECK: 2-step Heckman model, MI-H: multiple imputation based on the Heckman model, FML: full maximum likelihood selection model.

the questionnaire was relatively higher for patients in better health (higher HRQL). This is illustrated in Figure 5, which reports the predicted outcomes against the inverse Mills ratio values, extracted from the Heckman 2-step model estimates. This offers some support for the suggestion that those patients who did not respond to medical management chose to leave the study early or not complete HRQL questionnaires. Thus, a complete-case analysis may underestimate the average treatment effect.

The results from applying the various methods to the REFLUX data are reported in Table 4. We followed the same regression adjustment used in the primary analysis of this study[16]. The outcome linear model included key prognostic factors, such as age, gender, baseline HRQL (both REFLUX-specific and generic) and body mass index. The missing data model for each method included all the covariates used in the analysis model, and those variables anticipated to meet the criteria for a valid

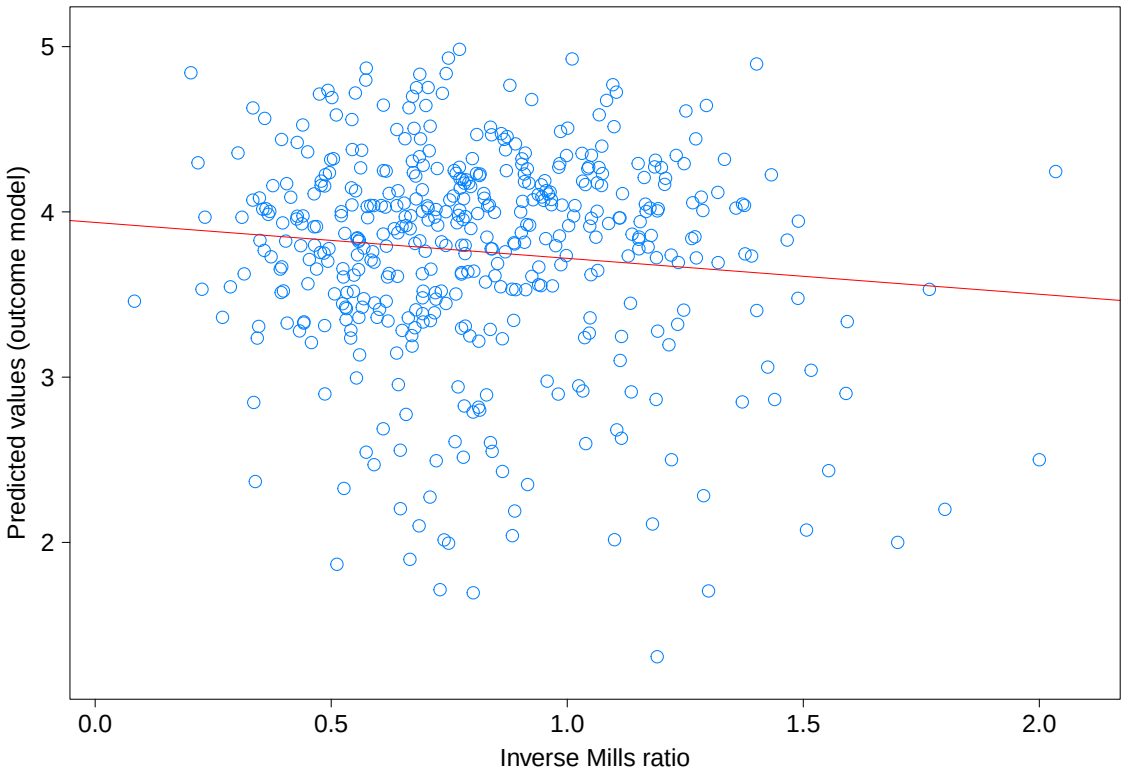

Figure 5: Predicted outcome values and estimated inverse Mills ratio by the two-step Heckman selection model. The straight line represents the linear association between these.

exclusion restriction. The parameter of interest was the average treatment effect ($\hat{\beta}_1$) of surgery versus medical management on the QALY over five years. Any causal interpretation for the results relies on the assumption that there are no unobserved confounders, and there is an inevitable concern that this assumption is not met.

There was strong evidence that patients receiving laparoscopic surgery had a higher average QALY over five years versus patients in the medical management group, after adjustment for baseline imbalances between the comparison groups. The estimated average gain in QALYs was larger under the MNAR approaches compared to both complete-case and MAR analyses. The three selection models led to similar treatment effects, although the MI based on the Heckman model and the full-likelihood approach yielded slightly lower standard errors. There was some evidence that the QALY estimates were associated with potential confounders, such as gender, baseline REFLUX and EQ-5D scores, and heart burn and symptoms scores. Regression coefficients for these parameters were very similar across the MNAR approaches, but again MI based on Heckman model and full maximum likelihood approaches had lower standard errors. The results from sensitivity analysis, including an analysis with no exclusion restriction variables, are reported in Appendix E. These sensitivity analyses

suggest that treatment effect estimates by the Heckman's approach are more sensitive to alternative specifications of the exclusion restriction compared to FIML. Partly, this may explained by: i) CIA unlikely to be tenable for the variables on patient's views about medicine, ii) exclusion restriction variables are weakly associated with missingness (see Table 2 and Appendix E). However, the main conclusion that laparoscopic surgery leads to greater improvements in QALYs versus medical management, was robust to alternative assumptions about the missing data.

7 Discussion

With missing data, inferences on treatment effects ultimately rest on untestable assumptions about the missing data mechanism. Selection models can make plausible assumptions regarding the missing data by allowing for departures from the standard MAR assumption. However, the use of selection models requires that the analyst recognises the additional, untestable assumptions imposed by these models. This paper clarifies the role of the exclusion restriction assumption across different selection models. We focused on this assumption for several reasons. First, while Heckman selection models are commonly used for handling MNAR data[24,31,26,25], little attention is given to assessing the validity and strength of the exclusion restriction. This is particularly concerning because, as our study shows, the practical advantages of this approach rely on the plausibility of this assumption. Second, strong exclusion restriction variables are rare in practice, and hence understanding the implications of a invalid and/or weak exclusion restriction to different selection model approaches and its impact on inferences is required. Third, other aspects of selection models such as model specification and distributional assumptions have received wider attention in the last few years[32,10,11,23].

The development of practical approaches for estimating treatment effects in the presence of non-ignorable missing data is an active area of research[29,1,27,3]. This paper adds to this literature by bringing together insights from econometrics and biostatistics to clarify the practical implications of the exclusion restriction assumption in selection models. We believe this study makes three distinctive contribution to this literature. Firstly, the paper considers a wide range of scenarios with both valid and invalid exclusion restrictions (predictive or not predictive of the missing data), together with alternative strengths of association of exclusion restriction. We find that under plausible distributional assumptions the FIML provides unbiased, precise estimates of treatment effects across a wide range of typical settings, and appears an appropriate method for handling MNAR data. Secondly, We have

Table 4: Regression coefficients and standard errors from applying the alternative methods to the REFLUX data. The parameter of interest is the average treatment effect ($\hat{\beta}_1$).

Covariates	CCA (N=231)	MI-MAR (N=453)	HECK (N=231)	MI-H (N=453)	FML (N=453)
Treatment	0.361 (0.101)***	0.410 (0.101)***	0.434 (0.098)***	0.442 (0.095)***	0.443 (0.095)***
Male	-0.153 (0.097)	-0.192 (0.099)	-0.212 (0.098)*	-0.211 (0.094)*	-0.197 (0.094)*
Age	-0.003 (0.004)	-0.006 (0.004)	-0.007 (0.004)	-0.006 (0.004)	-0.006 (0.004)
Baseline EQ-5D	2.066 (0.227)***	2.151 (0.223)***	2.171 (0.216)***	2.112 (0.197)***	2.174 (0.198)***
REFLUX score	-0.004 (0.003)	-0.007 (0.004)	-0.007 (0.004)	-0.007 (0.003)*	-0.007 (0.003)*
BMI (kg/m2)	-0.026 (0.013)*	-0.010 (0.012)	-0.006 (0.012)	-0.010 (0.011)	-0.008 (0.011)
Heart burn score	0.004 (0.003)	0.006 (0.003)	0.006 (0.003)*	0.006 (0.003)*	0.006 (0.003)*
Symptom score 1	0.007 (0.002)**	0.006 (0.002)*	0.006 (0.002)*	0.006 (0.002)*	0.006 (0.002)*
Symptom score 2	-0.005 (0.002)*	-0.003 (0.003)	-0.003 (0.002)	-0.003 (0.002)	-0.003 (0.002)
Nausea score	-0.001 (0.003)	0.002 (0.003)	0.001 (0.003)	0.001 (0.003)	0.002 (0.003)
Activity score	0.003 (0.004)	0.004 (0.004)	0.005 (0.004)	0.004 (0.004)	0.004 (0.004)
Intercept	2.958 (0.564)***	1.962 (0.552)***	2.202 (0.544)***	2.314 (0.487)***	2.699 (0.163)***

Notes: CCA: complete-case analysis, MI-MAR: multiple imputation assuming MAR, HECK: 2-step Heckman model, MI-H: multiple imputation based on the Heckman model, FML: full maximum likelihood selection model. Statistical significance is based on the corresponding coefficients from the outcome regression model: *p<0.05, **p<0.01 ***p<0.001

extended on previous work[33,34] by going beyond the bivariate normal distributed data, and considering skewed continuous outcomes. Our simulations suggest that the Heckman 2-step approach can be more robust to departures from the bivariate normal assumption compared to FIML, but it requires that the strength of the exclusion restriction is moderate/strong. Thirdly. in light of the common perception that multiple imputation (MI) is only valid under the MAR assumption, we have described and assessed a MI approach that addresses MNAR outcomes, and found that this approach performed no worse than the Heckman 2-step approach, but underperformed the FIML approach.

Our findings add to previous evidence on the performance of Heckman-based selection models [13,10,7] by showing that this approach requires not only a valid, but also a moderate or strong exclusion restriction variable to work well in practice. Our simulations also corroborate previous studies[13,23] that Heckman-based approaches may be less sensitive to departures from the assumed distributional assumptions compared to full-likelihood approaches, provided the exclusion restriction is not weak. In addition, a common criticism of multiple imputation is that MI is only valid (and hence useful) under the MAR assumption. This study builds on a previous study proposing a MI approach that allows for MNAR mechanisms compatible with the Heckman model[7], and showed that this method performed no worse than the 2-step Heckman approach. In particular, in settings without a valid exclusion restriction, the Heckman-based MI led to substantially more precise (lower rMSE) estimates compare to the original Heckman approach, partly because the estimation is based on the whole sample. However, overall the Heckman-based MI underperforms the full-likelihood approach across the scenarios considered.

This study sheds light on the role of exclusion restriction assumption in full-likelihood selection models. The findings from our simulation study suggested that the full-likelihood approach is less sensitive to alternative assumptions about the exclusion restriction compared to Heckman-based selection approaches. The former provided relatively low bias and rMSE consistently across all scenarios considered, where the assumptions about the joint distribution were plausible. In particular, the full-likelihood approach provided minimal biases and rMSE close to the 'true' values even when the exclusion restriction variable was invalid (scenarios MNAR1 and MNAR2). As the strength of the association between the exclusion restriction variables and non-response increases, the inclusion of these variables in the full-likelihood model helps make a weaker assumption about the missing data mechanism, and hence more precise estimates (see Appendix D). This is of practical relevance be-

cause existing methodological guidelines provide little insight on the role of the exclusion restriction variables on full-likelihood selection models[14,15].

Throughout the paper, we have assumed that the exclusion restriction variables were exogenous, i.e. CIA was met. However, we have investigated some scenarios where the CIA is violated by allowing alternative strengths of association between Z and Y (the results are reported in Appendix F), and found that even small violations of the CIA ($cor(Z, Y|X = 0.05)$ led to higher biases and rMSE for the Heckman acompared to FIML. This is in line with existing literature[13] that suggests that, by including Z as a predictor in the outcome model, the Heckman's approach provides more unstable estimates given the great overlap between X and Z. In the REFLUX study, we focused our discussion with clinical experts on the plausibility of the exclusion restriction. There were some concerns that the strength of association between patient's views about medicine variables and non-response was relatively weak, and that the CIA may not be tenable. For example, patients' perceptions about medicine may not be related to their chances of completing HRQL questionnaires. In addition, while the CIA would seem more reasonable for the centre size variable, this also had a weak association with non-response. In light of this, we were more confident about the full-likelihood approach to provide less biased, more precise estimates of treatment effect, although that did not change the conclusion that surgery improved patient's QALYs. As shown in the simulation study, the Heckman approach provides less precise estimates when the strength of the exclusion restriction is 'low'.

There are some aspects of selection models not addressed in this paper, that present interesting avenues for further research. Throughout, all selection models assumed the partially observed outcome and missingness followed a joint Normal distribution. However, the implications of the exclusion restriction for the choice of method are likely to be similar across selection models that allow for non-normal data[32,11], as well as with discrete outcomes[26]. Bayesian approaches provide the flexibility to accommodate joint models beyond the bivariate normal case, but the implementation of such models is not straightforward, not least because identifying uninformative, conjugate priors is challenging [6]. The development of flexible, user-friendly software tools for implementing selection models to handle non-normal outcomes with non-ignorable missing data is warranted. For example, Gomes and colleagues[23] are exploring the flexibility of copula-based approaches in this context, which provides an encouraging starting point.

Another area that warrants further consideration is longitudinal data. The longitudinal setting has

additional implications for the exclusion restriction assumption and choice of method. First, it is harder to find plausible exclusion restrictions in this setting; i.e. variables that predict non-response over time, but are unrelated to the longitudinal, partially-observed outcome. Second, longitudinal selection models are increasingly challenging to implement, and often require additional assumptions (e.g. about longitudinal correlation structure) and sophisticated estimation procedures[27]. In addition, we did not consider scenarios with both outcome and covariates missing. In such settings, multiple imputation approaches such as those advocated in this paper can accommodate the missingness both in covariates (typically assuming MAR) and MNAR outcomes.

In conclusion, this paper explores the implications of the different methodological choices concerning the exclusion restriction across alternative selection models, and finds that the relative performance of the methods differs according to the relevance and strength of the exclusion restriction. Under plausible distributional assumptions, the full-likelihood approach provides unbiased, precise estimates of treatment effects across a wide range of settings that could arise in practice, and appears an appropriate method for handling MNAR data. As illustrated in the REFLUX study, exclusion restriction variables are often weakly associated with non-response, and in these settings full-likelihood approaches are less sensitive to alternative assumptions about the exclusion restriction than Heckman-type selection models. This comes at the expense of an assumption about the joint distribution of the outcome and missingness, and we should routinely investigate the robustness of the study's conclusions to departures from the joint Normality[35]. Even in settings where the exclusion restriction assumption is plausible, likelihood approaches are typically more efficient than Heckman-type models, and therefore the former approach followed by appropriate sensitivity analysis ought to be the way forward.

Acknowledgments

The authors would like to thank Prof Mark Sculpher for the access to the REFLUX data.

References

1. Mattei A, Mealli F, Pacini B. Identification of causal effects in the presence of nonignorable missing outcome values *Biometrics.* 2014;70:278-288.

2. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials *Pharmacoeconomics.* 2014;32:1157-70.
3. Mason A, Gomes M, Grieve R, Ulug P, Powell J, Carpenter J. Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE Trial *Clinical Trials.* 2017;14:357-367.
4. Heckman JJ. Sample Selection Bias as a Specification Error *Econometrica.* 1979;47:153-161.
5. Diggle P, Kenward MG. Informative Drop-out in Longitudinal Data-Analysis *Journal of the Royal Statistical Society Series C-Applied Statistics.* 1994;43:49-93.
6. Daniels M, Hogan J. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: Chapman and Hall CRC 2008.
7. Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model *Stat Med.* 2016;35:2907-20.
8. Vella F. Estimating models with sample selection bias: A survey *Journal of Human Resources.* 1998;33:127-169.
9. Das M, Newey WK, Vella F. Nonparametric estimation of sample selection models *Review of Economic Studies.* 2003;70:33-58.
10. Pigini C. Bivariate Non-Normality in the Sample Selection Model *Journal of Econometric Methods.* 2015;4:123–144.
11. Zhelonkin M, Genton MG, Ronchetti E. Robust inference in sample selection models *Journal of the Royal Statistical Society Series B.* 2016;78:805-827.
12. Mohan K, Pearl J. On the Testability of Models with Missing Data *Artificial Intelligence and Statistics.* 2014;33:643-50.
13. Puhani PA. The Heckman correction for sample selection and its critique *Journal of Economic Surveys.* 2000;14:53-68.
14. Little RJ, Rubin DB. *Statistical analysis with missing data*. Wiley series in probability and mathematical statisticsNew York, US: Wiley 2002.
15. Molenberghs G, Fitzmaurice GM, Kenward M, Tsiatis AA, Verbeke G. *Handbook of missing data methodology*. Boca Raton, US: Chapman Hall/CRC 2014.
16. Grant AM, Boachie C, Cotton SC, Faria R, al . Clinical and economic evaluation of laparoscopic surgery compared with medical management for gastro-oesophageal reflux disease: 5-year

follow-up of multicentre randomised trial (the REFLUX trial) *Health Technol Assess.* 2013;17:1-167.

17. Gomes M, Gutacker N, Bojke C, Street A. Addressing Missing Data in Patient-Reported Outcome Measures (PROMs): Implications for the Use of PROMs for Comparing Provider Performance *Health Economics.* 2016;25:515-28.
18. EuroQol . EuroQol-a new facility for the measurement of health-related quality of life *Health Policy.* 1990;16:199-208.
19. Meng XL. Multiple-Imputation Inferences with Uncongenial Sources of Input *Statistical Science.* 1994;9:538-558.
20. Carpenter J, Kenward M. *Multiple Imputation and its Application*. Statistics in PracticeChichester, UK.: Wiley 2013.
21. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley series in probability and mathematical statisticsNew York, US: Wiley 1987.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice *Stat Med.* 2011;30:377-399.
23. Gomes M, Rosalba R, Camarena Brenes J, Giampiero M. Copula selection models for non-Gaussian outcomes that are missing not at random *Statistics in Medicine.* 2019;38:480-96.
24. Sales AE, Plomondon ME, Magid DJ, Spertus JA, Rumsfeld JS. Assessing response bias from missing quality of life data: the Heckman method *Health Qual Life Outcomes.* 2004;2:49.
25. Alva M, Gray A, al . The effect of diabetes complications on health-related quality of life: the importance of longitudinal data to address patient heterogeneity *Health Econ.* 2014;23:487-500.
26. Washbrook E, Clarke PS, Steele F. Investigating non-ignorable dropout in panel studies of residential mobility *Journal of the Royal Statistical Society Series C-Applied Statistics.* 2014;63:239-266.
27. Tseng CH, Elashoff R, Li N, Li G. Longitudinal data analysis with non-ignorable missing data *Statistical Methods in Medical Research.* 2016;25:205-220.
28. Toomet O, Henningsen A. Sample selection models in R: Package sampleSelection *Journal of Statistical Software.* 2008;27:1-23.
29. Mason A, Richardson S, Plewis I, Best N. Strategy for Modelling Nonrandom Missing Data

Mechanisms in Observational Studies Using Bayesian Methods *Journal of Official Statistics.* 2012;28:279-302.

30. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling 2003.
31. Del Bianco P, Borgoni R. Handling dropout and clustering in longitudinal multicentre clinical trials *Statistical Modelling.* 2006;6:141-157.
32. Marchenko YV, Genton MG. A Heckman Selection-t Model *Journal of the American Statistical Association.* 2012;107:304-317.
33. McGovern M, BÃd'rnighausen T, Marra G, Radice R. On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence. *Epidemiology.* 2015;26:229-37.
34. Clarke S, Houle B. Evaluation of Heckman Selection Model Method for Correcting Estimates of HIV Prevalence from Sample Surveys. *Center for Statistics and the Social Sciences.* 2012;Working paper no. 120.
35. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity *Statistics in Medicine.* 1998;17:2723-32.