PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. **9**: 269–279 (2010)
Published online 28 August 2009 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/pst.389

Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome

MAIN PAPER

Brenda J. Crowe[1,*,†], Ilya A. Lipkovich[1] and Ouhong Wang[2]
[1]*Eli Lilly and Company, Indianapolis, IN, USA*
[2]*Amgen, Thousand Oaks, CA, USA*

We performed a simulation study comparing the statistical properties of the estimated log odds ratio from propensity scores analyses of a binary response variable, in which missing baseline data had been imputed using a simple imputation scheme (Treatment Mean Imputation), compared with three ways of performing multiple imputation (MI) and with a Complete Case analysis. MI that included treatment (treated/untreated) and outcome (for our analyses, outcome was adverse event [yes/no]) in the imputer's model had the best statistical properties of the imputation schemes we studied. MI is feasible to use in situations where one has just a few outcomes to analyze. We also found that Treatment Mean Imputation performed quite well and is a reasonable alternative to MI in situations where it is not feasible to use MI. Treatment Mean Imputation performed better than MI methods that did not include both the treatment and outcome in the imputer's model. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: ***propensity scores; multiple imputation; observational study; imputation***

1. INTRODUCTION

Comparing binary outcomes of treated patients to untreated patients is often encountered in the analysis of observational studies. In order to account for differing baseline characteristics between the groups, various methods based on propensity scores [1,2] can be employed. Furthermore, patients are often missing some of the key baseline data needed in the propensity scores model, leading to the additional challenge of selecting the appropriate imputation method (otherwise, one would lose a significant portion of patients' data). Often baseline data are missing in groups (e.g. laboratory panels), leading to somewhat simple patterns of missing data. With simple missing data patterns it can be attractive to opt for simple methods of imputation (such as treatment mean imputation)

*Correspondence to: Brenda J. Crowe, Eli Lilly and Company, Indianapolis, IN, USA.
†E-mail: crowe_brenda_j@lilly.com

Pharmaceutical STATISTICS

that could be inadequate because they fail to utilize relationships between covariates, treatments and outcomes. Another approach used is to simply exclude covariates with missing data. This is generally unacceptable, as one of the key assumptions for propensity scores analyses to provide causal inferences is that all confounders have been included in the calculation of propensity scores.

This work was motivated by an analysis of safety data from the Hypopituitary Control and Complications Study (HypoCCS) [3], an observational cohort study in growth hormone-deficient adults that seeks to determine long-term safety outcomes in growth hormone-treated compared with untreated patients. Many (hundreds of) adverse events were analyzed by the method of subclassification on the propensity score [1,2]. Many patients were missing values of at least one baseline covariate, and it was not uncommon for entire groups of baseline data to be missing (e.g. systolic and diastolic blood pressure, laboratory panels). Missing quantitative data were imputed by a simple imputation scheme, replacing missing values for a variable by the respective treatment mean (Treatment Mean Imputation).

In this manuscript we investigate the statistical properties (in particular, the bias in computing odds ratios (ORs) of adverse events for treated versus untreated patients) of this single imputation scheme in comparison with three ways of implementing multiple imputation (MI) schemes. To investigate the statistical properties, we performed a simulation study comparing the bias and confidence interval (CI) coverage of the Treatment Mean Imputation scheme to various ways of implementing MI schemes, which varied by the scope of variables included in the imputation model (detailed in Section 3.3).

We first give a brief overview of the literature regarding analysis of propensity scores with missing values as well as MI. In Sections 3 and 4, we describe our simulation experiment and provide simulation results. Finally, we make recommendations about when and how the methods should be used.

2. BACKGROUND

The method of subclassification on the propensity score is a two-stage process. In the first stage, propensity scores (conditional probability of being treated) are generated without using any outcome information. The propensity scores are typically estimated by performing a logistic regression with treatment (Tx) as the response variable and baseline variables as the covariates. After generating propensity scores, patients are grouped into subgroups (five subgroups are often used, though this is not essential) within which the propensity scores are similar. The second stage is the analysis stage. We performed the analysis by using a Mantel–Haenszel (MH) OR [4] with associated CI [5], stratifying on the propensity score subgroup, to test for association of the Tx group with the AE in question.

Statisticians and epidemiologists use many different schemes to handle missing data. D'Agostino and Rubin [6] described a method for handling missing data via estimating generalized propensity scores (introduced in [2]) – pattern-mixture model for the joint distribution of covariates and binary indicators of missing values for each covariate. The number of parameters in the model was reduced by imposing log-linear type restrictions on cell probabilities of the joint distribution of all categorical variables (including missingness indicators) and the parameters of the model were estimated using the Expectation Conditional Maximization algorithm of Meng and Rubin [7] (see also [8], Section 8.5). Each patient was assigned a unique propensity score based on both observed covariates and the pattern of missingness associated with that patient. The resulting generalized propensity scores served as balancing scores for both treatment assignment and patterns of missingness. D'Agostino *et al.* [9] have described an experiment that introduced missing data using a non-ignorable missing data mechanism and compared treatment effect estimates using three propensity score methods: using only subjects with complete case data; incorporating missing value indicators into the model; and fitting separate propensity scores for each pattern of missing data (a simplified version of

DOI: 10.1002/pst

pattern-mixture model). The authors concluded that one should account for missing data in these models, but discouraged investigators from using the missing value indicator approach unless no other option is available. Furthermore, they mentioned that there is a growing literature that indicates that using the missing value indicator approach typically leads to biased results [10]. There are many examples of likelihood-based approaches for estimating parameters of generalized linear models in the presence of missing covariates (see Rubin and Little [11] for missingness at random and [10] for non-ignorable missing data). However, it is not obvious how these likelihood-based methods, when applied to estimating propensity scores (the first stage of analysis) would affect the estimation of the treatment effect in the second stage of propensity-based analysis. In particular, for our case, the second stage of analysis (MH OR) is not likelihood-based. Therefore, it would be difficult to implement it in the general likelihood-based scheme. Furthermore, we wanted to focus on methods that could be implemented with commonly available software. One easy-to-implement analytical approach that allows combining both stages in a single inferential procedure and that accounts for uncertainty in missing covariates throughout is MI.

MI was originally introduced by Rubin [12] to handle missing responses in sample surveys and has been recognized as a unified approach for the analysis of incomplete data [8,13]. It has been recently implemented in SAS® (as procedures MI and MIANALYZE [14]). Unlike various methods of single imputation, MI allows the uncertainty due to missingness to be incorporated into the final inferential statements about quantities of interest (e.g. CIs for ORs). In MI, several completed data sets are generated with missing values imputed using a statistical model (*imputation model*). These data sets are analyzed as if they were observed using an appropriate *analysis model*, and finally the resulting statistical estimates (e.g. ORs) are combined in a single estimate and associated CI by using simple combination rules ([13]). The attractiveness of MI is that it provides a universal and easy-to-apply method for accounting for the uncertainty due to missing data that is consistent with the data generation mechanism as long as probability of missingness does not depend on unobserved (missing) values themselves ([11,13]). The fact that MI allows different models to be used at imputation and analysis stages adds to both the flexibility and challenges of the method. A user may incorporate additional background variables to help impute missing values, but may want to omit those in the analysis model applied to the completed data sets. As suggested in literature ([8], pp. 139–143, [15,16]), MI inference would be valid even if the imputation model contains some unimportant predictors, as long as imputations are obtained by a proper method (in the sense that imputations are drawn from the Bayesian predictive distribution of missing data given the observed data). On the other hand, selection of an imputation model incompatible with the analysis model may significantly influence the properties of MI estimates, sometimes rendering them inefficient or even invalid ([17–20]).

Rubin's formulas for combining imputed data sets (given in [13], p. 76), in general, lead to valid, though conservative, inferences under so-called 'proper imputation', which roughly means that (i) imputation is done by drawing values from the correct predictive distribution of missing data given observed data [13,15], and (ii) the analysis model is correctly specified and (as was recently demonstrated in [18]) is likelihood-based. While in our simulations – which reflect analyses based on propensity scores typically encountered in clinical practice – we likely satisfied the first condition (i), the second condition is problematic. First, note that in our case the imputation model is likely to be incompatible with the analysis model, as the former is based on multivariate normal distribution of baseline covariates and the latter models occurrence of AEs post-baseline given baseline covariates. Also, the propensity score-based MH estimate is not likelihood-based and therefore might lead to an invalid inference [18] when used as the analysis method for MI. For this reason, we performed a simulation experiment; we did not know of a theoretical result to draw on for our situation.

To the best of our knowledge, the properties of propensity score methodology under MI have not

DOI: 10.1002/pst

Pharmaceutical STATISTICS

been extensively studied. Song *et al.* [21] used MI to impute missing data for propensity scores analysis, but those authors used it for a single study and did not investigate the statistical properties as we have.

3. SIMULATION EXPERIMENT

The idea of the simulation experiment was to simulate a reasonably realistic situation; however, we wanted to keep the number of covariates small. We therefore included six correlated covariates. Of these six, three (named $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$) were subsequently made to have some missing values, and three (named $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$) had no missing values. The latter three allowed us to introduce data missing at random (MAR) in a simple fashion. To simplify the true model, we allowed only one covariate ($\mathbf{z}_1$) to predict treatment assignment. Outcome depended on treatment assignment and on the value of the same covariate ($\mathbf{z}_1$). However, we assumed that the analyst would not necessarily know which covariates were important and might include several covariates when constructing propensity scores. Missing covariate values were introduced after the treatments were assigned, which made treatment assignment dependent on unobserved values of covariates. This methodology represents a common situation where, at the time of treatment assignment, all relevant information is available.

Outcome was never missing, as our focus was on the impact of missing covariates only. The simulation details are explained below.

3.1. Generating data

Simulation models for baseline outcome data were as follows.

We generated a random $n \times p$ data matrix ($n = 2000$ 'subjects', and $p = 6$ baseline explanatory variables) from a multivariate normal distribution with all means = 0, and correlation coefficient $\rho_{ij} = 0.3$ for all $i \neq j$, i, $j = 1,\ldots,6$. Columns correspond to six explanatory variables, $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$.

'Patients' were randomly assigned to one of two treatment groups (e.g. treated Tx = 1 and untreated Tx = 0) in an unbalanced fashion using the model

$$\text{Logit}(P(\text{Tx} = 1|\mathbf{z_1})) = \alpha_1 \mathbf{z_1} \tag{1}$$

Patients were randomly assigned to either experience an adverse event (AE) or not, according to the model

$$\text{Logit}(P(\text{AE}|\text{Tx}, \mathbf{z_1})) = -2.944 + \beta_1 \text{Tx} + \beta_2 \mathbf{z_1}, \quad \text{Tx} = 1 \text{ if treated, 0 otherwise} \tag{2}$$

This assignment corresponds to a base AE rate of approximately 5% and allows P(AE) to depend on treatment group (through β_1) and on the value of $\mathbf{z}_1$ (through β_2). Exp(β_1) is the true OR for treatment effect.

Values of α_1, β_1, and β_2 were arranged in a factorial layout to create input data sets for the simulation. The true values of α_1 were 0.3 and 1.0 (corresponding to moderate and strong imbalance, respectively, between the treatment groups with respect to $\mathbf{z}_1$). For β_1, the true values were 0 and log(1.5), corresponding to no treatment effect and a treatment effect with OR = 1.5, respectively. For β_2, true values were 0 and log(2.0), corresponding to P(AE) not depending on $\mathbf{z}_1$, and P(AE) depending strongly on $\mathbf{z}_1$, respectively. Three thousand data sets were simulated for each of the eight combinations of α_1, β_1, and β_2, yielding a maximum simulation standard error for CI coverage of less than 1%.

3.2. Missing data mechanisms

The data set with no missing values (hereafter referred to as 'gold standard,) was used as a hypothetical scenario (i.e. the gold standard to compare different analysis methods applied to missing data).

Data missing completely at random (MCAR) [11] were generated by setting approximately 12% of $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$ to be missing based on pseudo-random uniform values. For our simulations, either all of z_{1j}, z_{2j}, z_{3j} were set to be missing, or none were.

DOI: 10.1002/pst

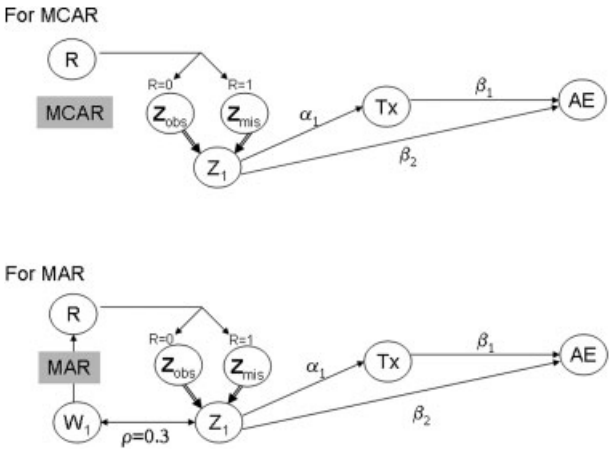

Figure 1. Simulation Schematic. R is a missing value indicator. $R = 1$ if the covariate value is missing and 0 otherwise.

Data MAR [11] were generated by setting approximately 12% of $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$ to be MAR by making $\mathbf{z}$'s missing based on completely observed values of $\mathbf{w}_1$ as follows: set z_{1j}, z_{2j}, z_{3j}, to be missing with probability 24% if $w_{1j} > 0$. Otherwise, z_{ij} are not missing. As with MCAR, either all of z_{1j}, z_{2j}, z_{3j} were set to be missing, or none were.

Figure 1 shows a schematic of the data simulation under two missing data mechanisms. The described simulation scenarios were combined with several *analysis strategies* as outlined below.

3.3.1. Analysis strategies. All of the analyses were performed using SAS® Version 9.1. MI analyses were performed with proc MI. Imputed values were generated from posterior predictive distributions for missing data given observed data using the Markov Chain Monte Carlo (MCMC) method for multivariate normal data with a single chain to create $m = 5$ imputations [14]. Variables used for imputation varied across scenarios as explained below, which allowed us to evaluate the robustness of MI when a larger number of potentially related variables were incorporated in the imputation model: continuous covariates and binary indicators for treatment and treatment outcome. The default options were used: posterior mode computed from the Expectation Maximization algorithm with a non-informative Jeffreys prior for initial value in MCMC. The major assumption of MI is that the missingness mechanism is MAR, which agrees with our simulations (see Section 3.2.).

The following analysis strategies were evaluated and compared with a 'gold standard' (analysis with propensity score adjustment of the full data set prior to generating missing data).

- Analysis without propensity scores adjustment (used as a benchmark of an obviously unacceptable approach when there is a Tx imbalance and P(AE) depends on Tx).
- Analysis with propensity score adjustment of complete cases (i.e. only 'patients' with no missing data).
- Propensity scores adjustment with missing baseline scores imputed by:
  - Treatment Mean Imputation, replacing the missing value with the mean of remaining data for that variable, within treatment group
  - MI using all observed **z**'s and **w**'s, but without Tx and AE in the imputation model
  - MI with observed **z**'s, **w**'s, and Tx in the imputation model, but no AE
  - MI with observed **z**'s, **w**'s, Tx, and AE in the imputation model

Note that in our simulation scenarios, we do not allow missing data for the outcome (AE), only for baseline covariates; therefore, when no propensity scores adjustment is used, the missing data are not

DOI: 10.1002/pst

Pharmaceutical STATISTICS

an issue. Also, for the last two methods, we apply a method for multivariate normal distribution to binary variables (Tx and AE), which we deemed appropriate based on the fact that methods of MI based on normal distribution are robust to non-normal data [8]; besides, the binary variables Tx and AE were free of missing values and therefore were essentially used as covariates in imputing missing baseline scores.

For each imputation scheme, we performed propensity score analysis and got estimated ORs and lower and upper 95% confidence limits. ORs are estimated by stratified MH estimator of common (across five propensity score bin strata) OR with associated 95% CIs. We ran MI with $m = 5$ completed data sets (see Schafer [8] for explanation of reasonable choice of m) to which the propensity score analysis was applied, resulting in five ORs. Those were combined (using PROC MIANALYZE) into the final OR with associated CI using so-called Rubin's rules [13] as follows: for ith imputed data set, compute the estimated log MH odds ratio and associated variance, $\log(\mathrm{OR}_{\mathrm{MH},i})$ and $\hat{\sigma}_i^2 = \mathrm{var}\{\log(\mathrm{OR}_{\mathrm{MH},i})\}$, respectively. Then the final estimate of common odds ratio is $\mathrm{OR}_{\mathrm{MH}} = \exp\left(m^{-1}\sum_{i=1}^{m}\log(\mathrm{OR}_{\mathrm{MH},i})\right)$ and its approximate $(1-\alpha/2)100\%$ level CI can be constructed by exponentiating the limits of the associated CI for log(OR), based on its total variance (V_{T}): $(\mathrm{OR}_{\mathrm{MH}}\exp(-t_{v,1-\alpha/2}\sqrt{V_{\mathrm{T}}}), \mathrm{OR}_{\mathrm{MH}}\exp(t_{v,1-\alpha/2}\sqrt{V_{\mathrm{T}}}))$, where the total variance, $V_{\mathrm{T}}=\bar{\sigma}^2+(1+m^{-1})B$ combines between-imputation variance, $B = (1/(m-1))\sum_{i=1}^{m}(\log(\mathrm{OR}_{\mathrm{MH},i})-\log(\mathrm{OR}_{\mathrm{MH}}))^2$ and within-imputation variance, $\bar{\sigma}^2 = m^{-1}\sum_{i=1}^{m}\hat{\sigma}_i^2$. Using a t-distribution with the degrees of freedom determined as $v = (m-1)[1+\bar{\sigma}^2/(1+m^{-1})B]^2$, rather than a normal distribution, accounts for the fact that for a finite number of imputations (m), the V_{T} is an inconsistent variance estimator having a non-degenerate chi-squared limiting distribution; hence a standard Wald-type inference is invalid [13]. (See discussion and critique of resulting 'inefficiency' of Rubin's estimator in [15,19,20].)

Simulation summaries given in Tables I–III are as follows:

- Standardized bias in log OR for treatment effect $= (\tilde{\beta}_1 - \beta_1)/\sqrt{\mathrm{Var}_{\mathrm{MC}}(\hat{\beta}_1)}$, where $\tilde{\beta}_1 = K^{-1}$

Table I. Standardized bias (%) in log odds ratio for the case of $\beta_1 = 0$ (no treatment effect).

		$\beta_2 = 0$		$\beta_2 = \log(2)$	
Analysis Strategy	Missing data mechanism	$\alpha_1 = 0.3$	$\alpha_1 = 1.0$	$\alpha_1 = 0.3$	$\alpha_1 = 1.0$
No PS	Not applicable	1.35	1.81	102.7*	281.4*
PS (Gold Standard)	No missing data	1.37	2.55	14.18	31.49
Complete Case	MCAR	1.65	2.50	13.29	29.32
PS+Treatment Mean Imputation	MCAR	1.23	2.41	14.97*	31.80*
PS+MI(no Tx or AE)	MCAR	1.52	2.38	31.16*	83.25*
PS+MI(Tx, no AE)	MCAR	1.45	2.78	23.07*	55.73*
PS+MI(Tx and AE)	MCAR	1.39	2.60	14.06	31.06
Complete Case	MAR	2.19	2.64	13.46	30.01
PS+Treatment Mean Imputation	MAR	1.45	3.00	17.40*	35.96*
PS+MI(no Tx or AE)	MAR	1.55	2.81	32.33*	86.99*
PS+MI(Tx, no AE)	MAR	1.50	2.82	23.89*	57.90*
PS+MI(Tx and AE)	MAR	1.50	3.09	14.11	31.26

*$P<0.05$ comparing mean log odds ratio for that method to the 'gold standard' (P-value from paired t-test). 'No PS' refers to analyzing the full data set without any accounting for baseline variables. α_1 is the coefficient of $\mathbf{z}_1$ in the model (1) and dictates the amount of imbalance between Tx groups with respect to $\mathbf{z}_1$. β_2 is the coefficient of $\mathbf{z}_1$ in the model (2) and dictates how much P(AE) depends on $\mathbf{z}_1$. MCAR = missing completely at random. MAR = missing at random. PS = propensity scores. MI = multiple imputation. All MI methods also include all six explanatory variables ($\mathbf{z}$'s and $\mathbf{w}$'s) in the imputation model. Results for each row are the average of 3000 simulated data sets.

DOI: 10.1002/pst

Table II. Actual coverage (%) of 95% model-based confidence intervals for the case of $\beta_1 = 0$ (no treatment effect); based on 3000 simulated data sets for each analysis strategy.

		$\beta_2 = 0$		$\beta_2 = \log(2)$	
Analysis Strategy	Missing data mechanism	$\alpha_1 = 0.3$	$\alpha_1 = 1.0$	$\alpha_1 = 0.3$	$\alpha_1 = 1.0$
No PS	Not applicable	94.87	95.47	80.60	17.83
PS (Gold Standard)	No missing data	94.93	94.93	94.80	93.50
PS+Complete Case	MCAR	95.03	94.87	95.33	94.00
PS+Treatment Mean Imputation	MCAR	94.90	94.90	94.90	93.67
PS+MI(no Tx or AE)	MCAR	95.03	95.40	94.57	87.33
PS+MI(Tx, no AE)	MCAR	95.10	95.33	94.90	91.40
PS+MI(Tx and AE)	MCAR	95.17	95.17	95.00	93.73
PS+Complete Case	MAR	94.90	95.10	95.33	93.90
PS+Treatment Mean Imputation	MAR	94.97	94.50	94.57	92.90
PS+MI(no Tx or AE)	MAR	95.07	95.17	94.50	86.63
PS+MI(Tx, no AE)	MAR	95.10	95.37	94.93	91.13
PS+MI(Tx and AE)	MAR	95.00	94.90	94.83	93.50

'No PS' refers to analyzing the full data set without any accounting for baseline variables. α_1 is the coefficient of $\mathbf{z_1}$ in the model (1) and dictates the amount of imbalance between Tx groups with respect to $\mathbf{z_1}$. β_2 is the coefficient of z_1 in the model (2) and dictates how much P(AE) depends on z_1. MCAR = missing completely at random. MAR = missing at random. PS = propensity scores. MI = multiple imputation. All the MI methods also include all six explanatory variables (**z**'s and **w**'s) in the imputation model.

$\sum_{i=1}^{K} \hat{\beta}_1$ is the average of $K = 3000$ Monte Carlo estimates of $\log(\mathrm{OR_{MH}})$, and $\mathrm{Var_{MC}}(\hat{\beta}_1)$ is the simulation variance of the estimator $\log(\mathrm{OR_{MH}})$.

- Type-I error rate is rejection rate for scenarios when the null hypothesis is correct (rejection rate is proportion of simulated samples when 0 is outside the 95% CI for log OR).

4. RESULTS

Unless otherwise specified, the following results correspond to the case of no true treatment effect ($\beta_1 = 0$); simulation scenarios under non-zero treatment effect led to essentially the same findings.

4.1. Bias

Bias in estimating treatment effect was present even when there were no missing data, and when there was no association between baseline covariate and outcome (no confounding) (see Table I, $\beta_2 = 0$). The bias is positive in this particular scenario because both the coefficients α_1 and β_2 are positive. Negative bias would have been induced if α_1 and β_2 had had opposite signs. When there was a strong association between $\mathbf{z}_1$ and outcome, bias increased (see Table I, $\beta_2 = \log(2)$). Not surprisingly, when there was no association between baseline covariate and outcome ($\beta_2 = 0$), all of the methods had a similar amount of bias compared with the 'gold standard' (see Table I, $\beta_2 = 0$). Under both MCAR and MAR, with an association between baseline covariate and outcome (Table I, $\beta_2 = \log(2)$), the only methods that produced estimates that were not significantly different from those obtained under 'No missing data' (which was used as the gold standard for comparison throughout) was the Complete Case method and MI with all covariates and Tx and AE indicator in the model. Treatment Mean Imputation produced only slightly larger bias than that of the gold standard under MCAR, but was slightly worse under MAR. However, the other two MI methods resulted in substantial bias.

4.2. CI coverage

In general, MI inference is conservative [13,19] (i.e. coverage of estimated CIs exceeds the

DOI: 10.1002/pst

Pharmaceutical STATISTICS

Table III. Mean width (of 3000 simulated datasets for each analysis strategy) of model-based confidence intervals for log odds ratio (top number) and relative to their simulation-based estimates (in parentheses). For this table, $\beta_1 = 0$ (no treatment effect).

		$\beta_2 = 0$		$\beta_2 = \log(2)$	
Analysis Strategy	Missing data mechanism	$\alpha_1 = 0.3$	$\alpha_1 = 1.0$	$\alpha_1 = 0.3$	$\alpha_1 = 1.0$
No PS	Not applicable	0.81 (0.96)	0.81 (1.00)	0.74 (0.98)	0.76 (0.95)
PS (Gold Standard)	No missing data	0.82 (0.99)	0.89 (0.97)	0.76 (0.98)	0.84 (0.98)
PS+Complete Case	MCAR	0.87 (0.97)	0.95 (0.98)	0.81 (0.99)	0.89 (0.96)
PS+Treatment Mean Imputation	MCAR	0.82 (0.98)	0.92 (0.98)	0.76 (0.97)	0.86 (0.96)
PS+MI(no Tx or AE)	MCAR	0.82 (0.99)	0.88 (1.00)	0.76 (0.98)	0.83 (0.97)
PS+MI(Tx, no AE)	MCAR	0.82 (0.98)	0.90 (0.99)	0.76 (0.98)	0.85 (0.99)
PS+MI(Tx and AE)	MCAR	0.82 (0.99)	0.90 (0.98)	0.76 (0.98)	0.85 (0.98)
PS+Complete Case	MAR	0.88 (0.98)	0.95 (0.98)	0.82 (1.01)	0.90 (0.97)
PS+Treatment Mean Imputation	MAR	0.82 (0.98)	0.92 (0.97)	0.76 (0.98)	0.86 (0.98)
PS+MI(no Tx or AE)	MAR	0.82 (0.98)	0.88 (0.99)	0.76 (0.97)	0.83 (0.98)
PS+MI(Tx, no AE)	MAR	0.82 (0.99)	0.90 (1.00)	0.76 (0.98)	0.85 (1.00)
PS+MI(Tx and AE)	MAR	0.82 (0.98)	0.90 (0.97)	0.76 (0.99)	0.85 (0.98)

'No PS' refers to analyzing the full dataset without any accounting for baseline variables. α_1 is the coefficient of $\mathbf{z}_1$ in the model (1) and dictates the amount of imbalance between Tx groups with respect to $\mathbf{z}_1$. β_2 is the coefficient of $\mathbf{z}_1$ in the model (2) and dictates how much P(AE) depends on $\mathbf{z}_1$. MCAR = missing completely at random. MAR = missing at random. PS = propensity scores. MI = multiple imputation. All the MI methods also include all six explanatory variables ($\mathbf{z}$'s and $\mathbf{w}$'s) in the imputation model.

nominal level). We looked at this in two ways when the null hypothesis was true ($\beta_1 = 0$).

First, we estimated the proportion of samples for which the 95% CI captured the true underlying log OR. Results from this estimation are given in Table II. Table II shows that, overall, the methods that gives coverage closest to the 'gold standard' method are MI with both Tx and AE (as well as all baseline covariates) in the imputer's model and the Complete Case method. When P(AE) depended on $\mathbf{z}_1$ ($\beta_2 = \log(2)$) and there was a large imbalance ($\alpha_1 = 1$) between treatment groups for $\mathbf{z}_1$, the MI methods that did not include both Tx and AE in the imputer's model had substantially poorer coverage than each of the other imputation schemes, including the Treatment Mean Imputation scheme. This was true for data MCAR and MAR. The Treatment Mean Imputation had lower coverage under MAR than MCAR, but not as low as MI when we did not include both Tx and AE in the imputer's model.

In order to provide a benchmark for how accurately the CIs were reported by analysis methods, compared with the underlying distributions, we compared them with the simulation-based intervals. Specifically, we looked at the ratio of (A) the average width of estimated CIs for log OR, divided by (B) the width of interval comprising

DOI: 10.1002/pst

95% of Monte Carlo estimates of log OR (2.5% on each side). Results are given in Table III. We found that these ratios were generally close to 1, although there was a tendency for the numerator (A) to be slightly smaller on average than B. This gave us some assurance that the model-based inference was capturing the true variability in the estimated parameters.

Table III also reports the mean width of the CIs for the log OR. The CIs for the Complete Case methods (both MCAR and MAR) were uniformly wider than for the other methods for the same parameters. This is to be expected, and suggests lower power for this method than for comparably performing methods.

5. DISCUSSION

Because the study that motivated this research involved analyzing hundreds of adverse events via propensity scores, we were particularly interested in the validity of imputation schemes that would not involve the response variable (as it is unsatisfying to have different propensity scores for each adverse event and hard to convince readers that one has succeeded in achieving balance). The Complete Case method is easy to implement and appears to provide valid inferences, though it is less efficient than other methods. Our results are in agreement with a fact noted in Little [22], that the Complete Case method provides valid inference when missingness depends on the regressors. We agree with Little's assessment that 'the rejection of incomplete cases seems an unnecessary waste of information' [22].

The Treatment Mean Imputation method performed quite well in our simulation study; however, it is unclear if this finding would hold in more complex scenarios. To understand why such a simplistic approach may work in this context, one should bear in mind that in the PS-based analysis of treatment effect only a relatively small portion of initial uncertainty due to missing covariate data would propagate into uncertainty about MHOR for treatment effect (our target parameter), compared with the uncertainty about parameters of logistic regression for estimating propensity scores (that are not of direct importance for us). This can be seen by looking at the estimated fraction of missing information (FMI) about a parameter of interest. While the missing values in z-covariates resulted in about 10–15% of FMI about the associated parameters of logistic regression for estimating PS (roughly corresponding to the probability of missing values set in our simulations at 12%), the FMI about estimated MHOR was in most cases under 1%. This explains why using a relatively simple method of single imputation did not result, contrary to our initial expectation, in substantially degraded coverage of associated CIs. Given low levels of FMI, one can think of mean imputation in the context of PS-based estimation of treatment effect as essentially adding a very small portion of probability mass to the data at the centroid of multivariate normal distribution in the covariate space, which should of course have a minimal impact on the inference about the treatment difference. (This may also shed some light on findings reported in the case study [21], where the authors were 'somewhat surprised' by the closeness of results of the available case analysis [in which they used various ad hoc imputation choices] and the MI analysis.) However, when using MI, one should be cautious about what imputation model is used, since using a misspecified model would introduce additional bias by placing a portion of data in the wrong part of the covariate space. Specifically, our results show that if one chooses to use MI as the imputation scheme for missing baseline data in a propensity scores analysis, it is necessary to include all the baseline variables as well as Tx and AE in the imputation model. That means that all explanatory variables as well as treatment group and response – i.e. a different MI set for *each* outcome studied – need to be included. While this may appear obvious to experts in MI (as it makes the *imputation model* compatible/congenial with the *analysis model* [15,19,20,22]), it may appear counterintuitive to practitioners of propensity scores analysis, where propensity scores are typically constructed without knowledge of outcome [23]. As further argued in [24], propensity scores methodology aims to

DOI: 10.1002/pst

Pharmaceutical STATISTICS

construct pseudo-randomization that, like randomization, should precede any observed outcome. However, our findings add credibility to the results of [21], where they combined MI and propensity scores methodologies in a similar way to our simulation setup. While their data structure was more complex than ours (in particular, the response was longitudinal), they still included a function of the outcome variable in the imputation model, to 'capture some of the association between outcomes and covariates'.

Given the results of our simulation study, the HypoCCS study team continued to use treatment mean imputation for analysis of adverse events with propensity scores.

Surprisingly, the MI methods using Rubin's rule for combined inference are slightly liberal, not conservative (i.e. estimated CIs had slightly smaller width, on average, than the 'true' distribution) for the situation we studied. However, as we mentioned earlier, the common characterization of MI estimator as 'conservative but valid' should not be taken for granted, since when the analysis model is not likelihood-based (clearly the case for propensity score-based MH estimate), the Rubin's estimator can be biased downward as well as upward [18]. Fortunately, as our simulation demonstrated, MI inference was valid for the settings that were used.

As with any simulation study, there are obvious limitations. The results may not apply to scenarios with a larger fraction of missing data or a different relationship between outcome/treatment assignment and covariates, or more complex missingness mechanisms (e.g. when missingness depends on unobserved covariates or outcomes).

We also ran limited simulations where missingness was not at random and our results were similar to the MAR scenarios presented in this manuscript. Note that in our simulation we assumed a linear relationship between the logit of probability treatment assignment and baseline covariate. Therefore, even when missingness depended on unobserved covariates, it should not have influenced estimated propensity scores very much, and therefore should not have influenced the treatment effect for the outcome. However, one could consider more complex scenarios where either relationship between the missing data and covariates is more complicated than what we studied or missingness depends on the outcome, leading to a potentially larger impact of missingness that might require different methodology to handle (see Ibrahim *et al*. [10] and references therein).

6. CONCLUSIONS

MI that included Tx and AE in the imputer's model had the best statistical properties of the imputation schemes we studied. MI is feasible to use in situations where one has just a few outcomes to analyze. We found that Treatment Mean Imputation performed fairly well and is a reasonable alternative to MI in situations where it is not feasible to use MI. Although the Complete Case strategy showed little bias, the larger standard errors (shown by uniformly wider confidence intervals as compared with other methods) limits its power in detecting relevant treatment differences. Additionally, when there are many explanatory variables, even a sparse missing data pattern can result in dropping a large number of cases. Especially in the context of analysis of safety outcomes, where event frequency is often low and retaining power is of utmost importance, we do not feel it is a better choice than the treatment mean imputation.

We found that when we used MI with an imputation model that was different from the analysis model, there was a big increase in bias compared with the gold standard, even exceeding the bias of the Treatment Mean Imputation. Thus, it is critically important, if one uses MI to impute missing baseline data, to include both the treatment and outcome in the imputer's model. As this was true in our relatively simple simulation scenarios, it is reasonable to believe that more complex scenarios would make it even more the case. The implication for practitioners is that different imputed data sets must be generated for each outcome. However, when used properly, MI had the best statistical properties of the schemes we studied.

DOI: 10.1002/pst

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
2. Rosenbaum PR, Rubin DB. Reducing bias in observation studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
3. Hartman ML, Crowe BJ, Kleinberg DL, Chipman JJ, Melmed S. Prospective multicenter surveillance of GH replacement in 2033 GH deficient (GHD) adults. *Program of the 83rd Annual Meeting of the Endocrine Society*, Denver, CO, 2001; pp. 145–146 (Abstract OR57–6).
4. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**:719–748.
5. Robins JM, Breslow N, Greenland S. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; **42**:311–323.
6. D'Agostino Jr RB, Rubin DB. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 2000; **95**:749–759.
7. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993; **80**:267–278.
8. Schafer JL. *Analysis of incomplete multivariate data*. Chapman and Hall: London, 1997.
9. D'Agostino R, Lang W, Walkup M, Morgon T. Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG) 2001. *Health Services & Outcomes Research Methodology* 2001; **2**:291–315.
10. Ibrahim J, Lipitz S, Chen M. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society Series B* (*Statistical Methodology*) 1999; **61**:173–190.
11. Little RJ, Rubin DB. *Statistical analysis with missing data*. Wiley: New York, 2002.
12. Rubin DB. Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Washington, DC, 1978; pp. 20–34.
13. Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley: New York, 1987.
14. SAS Institute Inc. *SAS/STAT® User's Guide, Version 9.1*. SAS Institute Inc.: Cary, NC; 2003.
15. Meng XL. Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science* 1994; **9**:538–574.
16. Rubin DB. Multiple imputation after 18+years. *JASA* 1996; **91**:473–489.
17. Fay RE. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Alexandria, VA, 1992; pp. 227–232.
18. Nielson SF. Proper and improper multiple imputation. *International Statistical Review* 2003; **71**:593–607.
19. Robins JM, Wang N. Inference for imputation estimators. *Biometrika* 2000; **87**:113–124.
20. Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika* 1998; **85**:935–948.
21. Song J, Belin TR, Lee MB, Gao X, Rotheram-Borus MJ. Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services & Outcomes Research Methodology* 2001; **2**:317–329.
22. Little RJA. Regression with missing X's: a review. *American Statistical Association* 1992; **87**: 1227–1237.
23. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2001; **2**:169–188.
24. Rubin DB. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 2007; **26**:20–36.

DOI: 10.1002/pst