



---

UW Biostatistics Working Paper Series

---

2-19-2009

# Multiple Imputation Methods for Treatment Noncompliance and Nonresponse in Randomized Clinical Trials

Leslie Taylor

*UW*, [taylorl@u.washington.edu](mailto:taylorl@u.washington.edu)

Xiao-Hua (Andrew) Zhou

*University of Washington*, [azhou@u.washington.edu](mailto:azhou@u.washington.edu)

---

## Suggested Citation

Taylor, Leslie and Zhou, Xiao-Hua (Andrew), "Multiple Imputation Methods for Treatment Noncompliance and Nonresponse in Randomized Clinical Trials" (February 2009). *UW Biostatistics Working Paper Series*. Working Paper 312. <http://biostats.bepress.com/uwbiostat/paper312>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

# Multiple Imputation Methods for Treatment Noncompliance and Nonresponse in Randomized Clinical Trials

L. Taylor\* and X. H. Zhou\*\*

University of Washington, Seattle, Washington 98195, U.S.A.

\* *email:* taylorl@u.washington.edu

\*\* *email:* azhou@u.washington.edu

**SUMMARY.** Randomized clinical trials are a powerful tool for investigating causal treatment effects, but in human trials there are oftentimes problems of noncompliance which standard analyses, such as the intention-to-treat or as-treated analysis, either ignore or incorporate in such a way that the resulting estimand is no longer a causal effect. One alternative to these analyses is the complier average causal effect (CACE) which estimates the average causal treatment effect among a subpopulation that would comply under any treatment assigned. We focus on the setting of a randomized clinical trial with crossover treatment noncompliance (e.g., control subjects could receive the intervention and intervention subjects could receive the control) and outcome nonresponse. In this article, we develop estimators for the CACE using multiple imputation methods, which have been successfully applied to a wide variety of missing data problems, but have not yet been applied to the potential outcomes setting of causal inference. Using simulated data we investigate the finite sample properties of these estimators as well as of competing procedures in a simple setting. Finally we illustrate our methods using a real randomized encouragement design study on the effectiveness of the influenza vaccine.

**KEY WORDS:** Causal inference; Complier average causal effect; Missing data; Multiple imputation; Noncompliance; Nonresponse; Principal stratification.

## 1. Introduction

The focus of empirical studies in medicine is often to estimate the causal effect of treatments, where randomized clinical trials are considered the most acceptable tool for investigating these causal relationships. But in trials involving human subjects there are oftentimes problems of patient noncompliance, where the patient does not adhere to the treatment assigned. In addition, there is also the problem of nonignorable missing data, where the missing-data mechanism may depend on unobserved data. Standard methods and analyses either ignore these complications, which can lead to biased estimates of causal treatment effect, or account for them in such a way that the resulting estimand can no longer be considered a causal effect of treatment. A statistical framework for causal inference that deals with the issue of noncompliance is based on potential outcomes and was first introduced by Neyman (1923) for randomized studies and later developed by Rubin (1974, 1978) for nonrandomized studies and other forms of inference. Rubin's approach, sometimes referred to as the Rubin Causal Model (Holland, 1986), provides a framework for defining the parameters of interest and correctly attributing the data observed between different treatment groups to the causal effects of the treatment.

As a motivating example, we focus on the setting of a randomized clinical trial for the influenza vaccine (McDonald, Hui, and Tierney, 1992). Observational studies suggest that among patients with a high risk of pulmonary disease,

those vaccinated with the influenza vaccine have better outcomes, including fewer hospitalizations. Clinical trials have never been performed because of ethical problems that come from withholding the vaccine from patients in the control arm. As an alternative, a three-year clinical trial was performed where the intervention arm increased the use of the influenza vaccine without changing its use in the control arm (McDonald et al., 1992). For doctors in the intervention arm, computer reminders were sent out when a patient with a scheduled visit was eligible for a flu shot. Unfortunately, noncompliance is a large problem in this type of study where encouragement to take the treatment, rather than the treatment itself, is randomized.

This article attempts to address noncompliance using a principal stratification framework (Frangakis and Rubin, 2002) which focuses on the subpopulation of compliers, who are not fully identifiable from the observed data. Alternative methods of modeling noncompliance, such as conditioning on the potential outcomes in selection models (Heitjan, 1999) or the use of structural equation models (Robins, Greenland, and Hu, 1999) are not considered here.

In the principal stratification framework, existing methods for estimation of the treatment effect include Bayesian (Imbens and Rubin, 1997; Hirano, Rubin, and Zhou, 2000; Frangakis, Rubin, and Zhou, 2002), likelihood (O'Malley and Normand, 2005; Zhou and Li, 2005), and moment methods (Frangakis and Rubin, 1999; Levy, O'Malley, and

Normand, 2004; O'Malley and Normand, 2005; Taylor and Zhou, 2008). An interesting alternative is multiple imputation (Rubin, 1987) which has been successfully applied to a wide variety of missing data problems but has not yet been applied to the potential outcomes setting of causal inference.

Multiple imputation (MI) is attractive for a number of reasons (see Rubin, 1996; Schafer and Olsen, 1998). An important advantage of the MI analysis (and the Bayesian analysis) over and above the likelihood methods is that neither the monotonicity nor the compound exclusion restriction, which have been used to develop and justify estimation procedures, are essential, and violations of these assumptions can be addressed. However, for both Bayesian and MI methods, identification is based entirely on a carefully thought out (or mathematically tractable) prior distribution. MI has statistical properties that closely approach the optimality of maximum likelihood (ML) methods (although not exactly as optimal since ML involves no simulation) with the principal advantage to MI being that it can be used in almost any situation, whereas ML is much more restricted in its applications, perhaps requiring specially designed EM algorithms and sometimes difficult analytic or numerical integration on the log likelihood for interval estimates.

An important advantage of the MI analysis over the Bayesian and ML methods is that MI works in conjunction with standard complete-data methods and software so that after the imputations are generated, data analysts who are not professional statisticians can apply standard complete-case analyses to the multiply imputed data sets. And since the imputation phase is operationally distinct from subsequent analyses, given a set of  $m$  imputations, many different analyses can be performed making it unnecessary to re-impute when a new analysis is considered. By using a Bayesian framework for imputation and a frequentist approach to examining the estimator, one is also able to take advantage of the Bayesian framework, for example by relaxing certain assumptions, while having the ability to make direct comparisons to other frequentist methods. MI is especially promising when standard complete-case analyses are difficult to modify analytically in the presence of nonresponse, as in the case of nonignorable missing data or multivariate outcomes. Finally, it can be very efficient with as few as 3–5 imputations needed to obtain valid results.

Barnard et al. (1998) describe a basic template for obtaining inferences using multiple imputations in a setting where only intervention subjects could receive the new treatment, although they stop short of implementing the imputation techniques or comparing them to existing methods. This article extends the template for obtaining inferences using multiple imputation to a setting of crossover noncompliance (e.g., control subjects could receive the intervention and intervention subjects could receive the control), and compares these methods to existing methods. Section 2 and 3 define the causal inference notation, assumptions, and parameters of interest. Section 4 introduces the multiple imputation framework. Section 5 provides additional assumptions used in this framework. Section 6 provides simulation results for the setting of a binary outcome with noncompliance and outcome nonresponse. In Section 7 these methods are applied to a reanalysis

of a data set on the influenza vaccine previously studied by McDonald et al. (1992).

## 2. Setting and Notation

The setting consists of a randomized clinical trial where  $N$  subjects are randomized to treatment  $\mathbf{Z}$  where  $\mathbf{Z}$  is an  $N$ -vector of treatment assignments with  $i$ th element  $Z_i$ . For subject  $i$ ,  $Z_i = 1$  if assigned to the new treatment and  $Z_i = 0$  if assigned to the control. Let  $\mathbf{D}(\mathbf{Z})$  be the vector of potential treatment receipts with  $i$ th element  $D_i(\mathbf{Z})$ . Here  $D_i(\mathbf{Z}) = 1$  if subject  $i$  receives the new treatment and  $D_i(\mathbf{Z}) = 0$  if subject  $i$  receives the control, under treatment assignment vector  $\mathbf{Z}$ . Let  $Y_i(\mathbf{Z})$  and  $R_i(\mathbf{Z})$  be, respectively, the potential outcome and potential indicator for response, equal to 1 for response, and 0 for nonresponse, on outcome  $Y_i$ , for subject  $i$  under treatment assignment  $\mathbf{Z}$ . A random subset of the  $N$  subjects are assigned to treatment arm  $Z$ .

## 3. Definition of Causal Estimands

We make the stable unit treatment value assumption (SUTVA) which limits the number of potential outcomes and allows us to write the potential outcomes as functions of  $Z_i$  rather than of the vector  $\mathbf{Z}$ . Formally the SUTVA states that if, for all  $\mathbf{Z}, \mathbf{Z}'$  where  $Z_i = Z'_i$  (i.e., under treatment assignments which may differ for some subjects but not for subject  $i$ ),  $D_i(\mathbf{Z})$  equals  $D_i(\mathbf{Z}')$ ,  $Y_i(\mathbf{Z})$  equals  $Y_i(\mathbf{Z}')$ , and  $R_i(\mathbf{Z})$  equals  $R_i(\mathbf{Z}')$ , which means that we can write  $D_i(\mathbf{Z}), Y_i(\mathbf{Z}),$  and  $R_i(\mathbf{Z})$  as  $D_i(Z_i), Y_i(Z_i),$  and  $R_i(Z_i)$ , respectively. Under the SUTVA, we can define the intention-to-treat (ITT) average causal effect of  $Z$  on  $Y$  as  $E[Y_i(1) - Y_i(0)]$ .

We can stratify the population into four compliance principal strata—never-takers, always-takers, compliers, and defiers—as determined by the value of the vector of potential treatment receipts  $[D_i(0), D_i(1)]$  where  $C_i = n$  (never-taker) if  $D_i(0) = D_i(1) = 0$ ;  $C_i = a$  (always-taker) if  $D_i(0) = D_i(1) = 1$ ;  $C_i = c$  (complier) if  $D_i(0) = 0$  and  $D_i(1) = 1$ ; and  $C_i = d$  (defier) if  $D_i(0) = 1$  and  $D_i(1) = 0$ . Note that without additional assumptions, compliance type is not identified from the observable data on treatment receipt  $D$ . In addition to the SUTVA, we chose to assume the *monotonicity assumption* which states that  $D_i(1) \geq D_i(0)$  for all subjects (i.e., no defiers) which means compliance type is observable when  $Z_i \neq D_i$ . We also assume that compliance is all-or-none meaning that any switching of treatments was done soon after randomization so that the subject is assumed to have completely taken the new treatment or the control. Note that unlike membership to the *observed* compliance strata, membership to these *principal* compliance strata is unaffected by assigned treatment and therefore can be considered as a baseline covariate (Frangakis and Rubin, 2002).

Define  $\eta_{zt} = E[Y_i(z)|Z_i = z, C_i = t]$  and  $\gamma_{zt} = E[R_i(z)|Z_i = z, C_i = t]$  to be the conditional expectation of outcome and indicator for response, respectively, given treatment assignment  $z$  and principal compliance type  $t$ , and let  $\omega_t$  be the proportion of the population with compliance type  $t$  for  $t \in \{n, c, a, d\}$ . Then we can define the ITT effect as  $ITT = \sum_{t \in \{n, a, c, d\}} \omega_t ITT_t$  where  $ITT_t = E[Y_i(1) - Y_i(0)|C_i = t]$  is the average ITT effect of  $Z$  on  $Y$  for the subpopulation of compliance type  $t$ . Under monotonicity, defiers do not exist, and the noncompliers (never-takers and always-takers),

by definition of this group, do not carry information about the comparison between treatments, with respect to finding the causal effect of treatment on outcome (although they do provide information on the effect of assignment to treatment on outcome). Thus we focus on the subpopulation of compliers and define the complier average causal effect (CACE) to be  $ITT_c$ , or  $CACE = E[Y_i(1) - Y_i(0)|C_i = c]$ , which is the average treatment effect among the subpopulation of compliers, where under randomization,  $CACE = \eta_{1c} - \eta_{0c}$ .

#### 4. Framework for Multiple Imputation in the Setting of Noncompliance and Nonresponse

##### 4.1 Introduction to Multiple Imputation

Letting  $\theta = (\eta_{1n}, \eta_{1c}, \eta_{1a}, \eta_{0n}, \eta_{0c}, \eta_{0a}, \gamma_{1n}, \gamma_{1c}, \gamma_{1a}, \gamma_{0n}, \gamma_{0c}, \gamma_{0a}, \omega_n, \omega_c, \omega_a)$  be the parameter vector, we partition the complete data  $M$  into the observed data,  $M_{\text{obs}} = (Y_{\text{obs}}, Z, D, R)$ , and the missing data,  $M_{\text{mis}} = (Y_{\text{mis}}, C_{\text{mis}})$ . Inference is then based on the observed-data posterior density

$$P(\theta|M_{\text{obs}}) = \int p(\theta|M_{\text{obs}}, M_{\text{mis}})p(M_{\text{mis}}|M_{\text{obs}})dM_{\text{mis}}. \quad (1)$$

Here we see that the observed-data posterior density can be obtained by averaging the complete-data posterior distribution over the predictive distribution  $p(M_{\text{mis}}|M_{\text{obs}})$ ; we therefore draw the missing values from  $p(M_{\text{mis}}|M_{\text{obs}})$  to complete the data set, and then we draw  $\theta$  from its completed-data posterior distribution  $p(\theta|M_{\text{obs}}, M_{\text{mis}})$ . In multiple imputation, (1) is approximated by analyzing the data sets separately and then combining the results (discussed in Section 4.3). The theoretical motivation for multiple imputation is Bayesian although the estimators have been shown to have good frequentist properties (Rubin and Schenker, 1986; Schenker and Welsh, 1988; Rubin, 1996; Schafer, 1997; see discussion section).

##### 4.2 Computational Setup for Multiple Imputation

The idea of multiple imputation is to draw the missing data and parameters from the joint distribution  $P(M_{\text{mis}}, \theta|M_{\text{obs}}) \equiv P(Y_{\text{mis}}, C_{\text{mis}}, \theta|Y_{\text{obs}}, Z, D, R)$  by recursively iterating between the missing data and the parameters. We do this by a sequence of conditional distributions, where we model the conditional distribution of the compliance type  $C_i$ , and the conditional distribution of potential outcomes and potential response indicators given compliance type. We use data augmentation (Tanner and Wong, 1987) to draw from  $P(Y_{\text{mis}}, C_{\text{mis}}, \theta|Y_{\text{obs}}, Z, D, R) = P(Y_{\text{mis}}|C_{\text{mis}}, Y_{\text{obs}}, Z, D, R, \theta)P(C_{\text{mis}}|Y_{\text{obs}}, Z, D, R, \theta)P(\theta|Y_{\text{obs}}, Z, D, R)$ .

The first stage involves draws from  $P(Y_{\text{mis},i}|Y_{\text{obs},i}, C_i, Z_i, D_i, R_i, \theta)$  where, given  $(C_i, Z_i, D_i, R_i)$ , the  $Y_i$  are independent indicators dependent on the data only through  $Z_i, D_i$ , and  $C_i$  (under the SUTVA and latent ignorability). The second stage of data augmentation involves an analysis of the complete-data posterior distribution of  $\theta$ . Let  $\delta(z, d)$  indicate the subset of subjects with observed values ( $Z_i = z, D_i = d$ ) for  $z = 0, 1$  and  $d = 0, 1$  where  $\delta(z, \cdot) = \delta(z, 0) \cup \delta(z, 1)$  and let  $\zeta(t)$  indicate the subset of subjects with compliance type  $t$  for  $t = \{n, c, a\}$ . Then the conditional posterior distribution

of  $\theta$  given  $\mathbf{C}, \mathbf{Z}, \mathbf{D}, \mathbf{Y}$ , and  $\mathbf{R}$  has the simple structure:

$$p(\theta|\mathbf{C}, \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{R}) \propto p(\theta) \prod_{z \in \{0,1\}} \prod_{t \in \{n,c,a\}} \left( \prod_{i \in \{\zeta(t) \cap \delta(z, \cdot)\}} \omega_t f_{zt,i} \gamma_{zt}^{R_i} (1 - \gamma_{zt})^{1-R_i} \right).$$

To take advantage of this structure, we assume prior joint independence of all parameters so that:

$$p(\omega_n, \omega_c, \omega_a|\mathbf{C}, \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{R}) \propto p(\omega_n, \omega_c, \omega_a) \omega_n^{N_n} \omega_c^{N_c} \omega_a^{N_a}$$

$$p(\eta_{zt}|\mathbf{C}, \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{R}) \propto p(\eta_{zt}) \prod_{i \in \{\zeta(t) \cap \delta(z, \cdot)\}} \eta_{zt}^{Y_i} (1 - \eta_{zt})^{1-Y_i}$$

$$p(\gamma_{zt}|\mathbf{C}, \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{R}) \propto p(\gamma_{zt}) \prod_{i \in \{\zeta(t) \cap \delta(z, \cdot)\}} \gamma_{zt}^{R_i} (1 - \gamma_{zt})^{1-R_i}.$$

The final step involves draws from  $P(C_{\text{mis}}|Y_{\text{obs}}, Y_{\text{mis}}, Z, D, R, \theta)$  where subjects in  $\delta(0, 0)$  are a mixture of never-takers and compliers; subjects in  $\delta(1, 1)$  are a mixture of always-takers and compliers; and subjects in  $\delta(1, 0)$  and  $\delta(0, 1)$  are never-takers and always-takers, respectively (from monotonicity). Therefore  $P(C_i = n|Y_i, Z_i = 1, D_i = 0, R_i, \theta) = P(C_i = a|Y_i, Z_i = 0, D_i = 1, R_i, \theta) = 1$  and from Bayes' theorem,

$$\begin{aligned} P(C_i = t|Y_i, Z_i = 0, D_i = 0, R_i, \theta) &= \frac{\eta_{0t}^{Y_i} (1 - \eta_{0t})^{(1-Y_i)} \gamma_{0t}^{R_i} (1 - \gamma_{0t})^{(1-R_i)} \omega_t}{\sum_{t \in \{n,c\}} \eta_{0t}^{Y_i} (1 - \eta_{0t})^{(1-Y_i)} \gamma_{0t}^{R_i} (1 - \gamma_{0t})^{(1-R_i)} \omega_t} \\ &\text{for } t \in \{n, c\}, \end{aligned}$$

$$\begin{aligned} P(C_i = t|Y_i, Z_i = 1, D_i = 1, R_i, \theta) &= \frac{\eta_{1t}^{Y_i} (1 - \eta_{1t})^{(1-Y_i)} \gamma_{1t}^{R_i} (1 - \gamma_{1t})^{(1-R_i)} \omega_t}{\sum_{t \in \{a,c\}} \eta_{1t}^{Y_i} (1 - \eta_{1t})^{(1-Y_i)} \gamma_{1t}^{R_i} (1 - \gamma_{1t})^{(1-R_i)} \omega_t} \\ &\text{for } t \in \{n, c\}. \end{aligned}$$

##### 4.3 Analysis of the Multiply Imputed Data Sets

Rubin (1987) developed multiple imputation combining rules for interval estimation and hypothesis testing which account for imputation uncertainty. The complete-data CACE estimate  $\hat{Q}$  would be the observed difference in treatment means among the subpopulation of compliers with associated variance  $U = s_1^2/n_1 + s_2^2/n_2$  where  $s_1$  and  $s_2$  are the sample standard deviations and  $n_1$  and  $n_2$  are the sample sizes within treatment groups among compliers. Using this complete data method on the  $m$  completed data sets, we obtain the estimates  $(\hat{Q}^1, \hat{Q}^2, \dots, \hat{Q}^m)$  and associated variances  $(U^1, U^2, \dots, U^m)$  of the CACE.

We combine the estimates as follows (using the rules in Rubin, 1987): the overall estimate is  $\hat{Q} = 1/m \sum Q^i$  for  $i = 1, \dots, m$  and variance  $T_m = \bar{U}_m + \{1/(m+1)\} B_m$  where  $\bar{U}_m = 1/m \sum U^i$  is the complete-data variance estimate and  $\{1/(m+1)\} B_m$  is additional variance due to imputing the missing data where  $B_m = \{1/(m-1)\} \sum (\hat{Q}^i - \hat{Q})^2$ . Inference is then based on the  $t$ -distribution approximation

$T^{-1/2}(Q - \bar{Q}) \sim t_v$  with degrees of freedom  $v = (m - 1)[1 + \bar{U}_m / \{(1 + m^{-1})B_m\}]^2$ .

If the fraction of missing information about a scalar estimand is  $\lambda$ , the relative efficiency of a point estimate (on the variance scale) based on  $m$  imputations to one based on an infinite number of imputations is approximately  $(1 + \lambda/m)^{-1}$ , which can be estimated by  $\hat{\lambda} = (1 + 1/m)B_m/T_m$  (Rubin, 1978). For reasonable amounts of missing data, 3 to 5 imputations are adequate since the asymptotic efficiency of the repeated-imputation finite- $m$  estimate, relative to the infinite  $m$  estimate, is close to one in this case (Schafer, 1997). But since, under monotonicity, compliance type is missing for many subjects (and missing for all subjects without the assumption of monotonicity), more than the typical 3–5 imputations may be needed; for our setting, if the percent of missing information was even as high as 90%, then an estimate based on  $m = 10$  imputations would tend to have a standard error only  $\sqrt{(1 + 0.9/10)} = 1.04$  times as large as the estimate with  $m = \infty$  imputations; therefore 10 imputations is adequate for our setting.

## 5. Additional Assumptions

In dealing with missing outcomes, it is important to condition on important covariates, such as compliance type, before assuming nonresponse is independent of the outcome; we assume the *latent ignorability* assumption which states that, within each latent principal compliance type, potential outcomes and associated potential response indicators are independent, or

$$P[R_i(0)|Y_i(0), C_i] = P[R_i(0)|C_i]$$

$$P[R_i(1)|Y_i(1), C_i] = P[R_i(1)|C_i].$$

Unlike MI and Bayesian methods, ML and moment methods require further assumptions for identifiability of model parameters. Although the following assumption is not necessary for the MI analysis, it is often plausible and helps to facilitate inference for the CACE (by reducing variability in the estimate), and will also allow us to compare the MI methods to the likelihood and moment methods. In addition to the SUTVA, monotonicity, and latent ignorability, we chose to assume the *compound exclusion restriction for never-takers and always-takers*. This states that among the subpopulations of never-takers and always-takers, treatment assignment does not affect outcomes or response behaviors, or

$$P[Y_i(1), R_i(1)|C_i = n] = P[Y_i(0), R_i(0)|C_i = n]$$

$$P[Y_i(1), R_i(1)|C_i = a] = P[Y_i(0), R_i(0)|C_i = a].$$

## 6. Simulation Study with a Binary Outcomes

For the MI estimator, in addition to assuming the SUTVA and monotonicity, we consider the CACE parameter derived both under the model assuming the compound exclusion restriction (with the estimator denoted  $MI_1$ ) as well as under the model that imposes no compound exclusion restriction (with the estimator denoted  $MI_2$ ). Hirano et al. (2000) refers to the model which imposes no exclusion restriction as “weakly identifiable” in the sense that there is no unique ML estimator although the posterior distribution is proper. For the

purpose of comparison, we also calculate the moment estimator (MOM) and the EM estimator (MLE) for the CACE (Zhou and Li, 2005; Taylor and Zhou, 2008) derived under the SUTVA, monotonicity, and the compound exclusion restrictions.

Subjects are randomized to the control or new treatment arm with  $P(Z_i = 1) = 0.5$  where principal compliance type  $C_i$  is generated independently as a multinomial random variable with  $(\omega_n, \omega_c, \omega_a) = (0.3, 0.4, 0.3)$ .  $Y_i$  and  $R_i$  are generated from a binomial distribution with a mean conditional upon treatment assignment  $Z_i$  and principal compliance type  $C_i$ . We fix average outcomes  $\eta_{zt} = 0.5$  and average response probabilities  $\gamma_{zt} = 0.5$  where  $f_{zt} = \eta_{zt}^{Y_i} (1 - \eta_{zt})^{(1 - Y_i)}$  for  $z = 0, 1$  and  $t \in \{n, c, a\}$ . All posterior distributions are easy to draw from for conventional prior distributions because they only involve the Beta and Dirichlet distributions. We used two separate sets of noninformative priors: the Beta(1,1) and Dirichlet(1) priors for the outcome and compliance type distributions, respectively, as well as the Beta(0.5,0.5) and Dirichlet(0.5) priors corresponding to Jeffrey’s priors for these distributions. Since both sets of prior distributions gave similar results we only present the results using the Beta(1,1) and Dirichlet(1) priors. We use the ML estimates as starting values for the parameters in the DA chain (and since the ML estimates are derived under the compound exclusion restrictions, for  $MI_2$ , we let the starting values  $\eta_{0n} = \eta_{1n}, \eta_{0a} = \eta_{1a}, \gamma_{0n} = \gamma_{1n}$ , and  $\gamma_{0n} = \gamma_{1n}$ . One thousand data sets were created per scenario with 10,000 iterations used for  $MI_1$  and 50,000 iterations used for  $MI_2$  in the data augmentation procedure with  $m = 10$  imputations. Here imputations were obtained by subsampling a single chain, taking every  $k$ th iterate where the value  $k$  was determined by looking at time series and autocorrelation plots of the CACE parameter, as suggested by Schafer (1997); in our scenarios  $k$  ranged between 400 and 5000.

In this section we examine some finite sample properties of the estimators, first under hypothetical conditions that follow the assumptions of latent ignorability and the compound exclusion restriction, and then under certain deviations from these assumptions. Table 1 reports the mean squared error (MSE), bias, coverage rates for nominal 95% confidence intervals, and average confidence interval length for the CACE.

### 6.1 Simulation Results

The first three columns of Table 1 correspond to the behavior of the estimators when the compound exclusion restriction and latent ignorability hold in the data. For the smallest sample size of  $n = 100$ , both MI estimators ( $MI_1$  and  $MI_2$ ) perform the best in terms of MSE and interval length where  $MI_1$  has the smallest MSE and tightest confidence interval, with all estimators having similar bias and coverage rates. Here the MOM estimator performs the worst in terms of having a much larger MSE and much wider confidence intervals than the rest of the estimators. (Note that even under correct model assumptions, coverages for the MOM estimator are higher than the nominal 95 since the variance estimator for MOM involves a denominator of counts that could potentially be very small or even zero in which case the variance estimator would be quite large or possibly infinite.) For smaller sample sizes, the MI estimators tend to have better properties because they

**Table 1**  
Simulation results for the CACE with  $\{(\omega_n, \omega_c, \omega_a) = (0.3, 0.4, 0.3)\}$

	CER and LI Hold $(\eta_{0n}, \gamma_{0n}) = (0.5, 0.5)$			CER Violated $(\eta_{0n}, \gamma_{0n}) = (0.3, 0.6)$			LI Violated $r = \frac{1}{2}$		
	100	500	1000	100	500	1000	100	500	1000
MSE									
MOM	0.673	0.077	0.035	0.587	0.090	0.057	5.023	0.087	0.045
MLE	0.351	0.029	0.013	0.327	0.048	0.037	0.426	0.036	0.022
M <sub>1</sub>	0.121	0.036	0.016	0.132	0.055	0.040	0.128	0.042	0.025
M <sub>2</sub>	0.149	0.117	0.114	0.154	0.117	0.114	0.151	0.116	0.112
Bias									
MOM	0.004	-0.003	0.003	0.200	0.152	0.158	-0.071	-0.087	-0.091
MLE	0.005	-0.003	0.003	0.166	0.152	0.157	-0.084	-0.085	-0.091
M <sub>1</sub>	-0.001	-0.004	0.004	0.116	0.154	0.159	-0.053	-0.085	-0.092
M <sub>2</sub>	-0.004	-0.003	0.000	0.074	0.080	0.078	-0.044	-0.040	-0.035
Interval length									
MOM	4.690	0.850	0.579	4.091	0.775	0.536	4.192	0.871	0.593
MLE	1.920	0.650	0.450	1.770	0.620	0.430	2.000	0.660	0.450
M <sub>1</sub>	1.450	0.800	0.540	1.430	0.740	0.510	1.480	0.790	0.540
M <sub>2</sub>	1.640	1.500	1.480	1.630	1.450	1.440	1.650	1.480	1.460
% Coverage									
MOM	99.40	99.40	98.70	99.50	92.20	82.10	99.50	98.00	96.90
MLE	98.30	95.80	95.50	96.00	84.30	68.60	97.90	94.10	87.90
M <sub>1</sub>	97.70	95.70	95.90	95.60	86.60	75.60	97.20	93.40	89.60
M <sub>2</sub>	99.80	100	100	99.70	100	100	99.50	100	100

are in effect approximating the observed-data posterior by a finite mixture of densities rather than one single density as is the case for ML methods. In addition, the theory underlying multiple imputation is Bayesian, which is known to provide useful inference in smaller samples. It is also true that even noninformative priors can convey information about the parameters of interest which, in small samples, may increase the information. For larger sample sizes ( $n \geq 500$ ), bias and coverage were similar among the MOM, MLE, and MI<sub>1</sub> estimators. MSE for MOM was more than double that of MLE and MI<sub>1</sub> estimators; and MSE, confidence interval length, and coverage were much higher for the MI<sub>2</sub> estimator, which can be attributed to the increase in variability from estimating the four additional parameters in the model that imposes no compound exclusion restriction.

We violate the compound exclusion restriction in the data by letting  $\eta_{0n} = 0.3$  and  $\gamma_{0n} = 0.6$  keeping  $\eta_{1n} = 0.5$  and  $\gamma_{1n} = 0.5$  where results are reported in the second three columns of Table 1. Again for the smallest samples size ( $n = 100$ ), the MI estimators perform the best where MI<sub>1</sub> has the smallest MSE, bias, and average confidence interval length. For larger sample sizes ( $n \geq 500$ ), MI<sub>2</sub> has the smallest bias among the estimators although its MSE, average confidence interval length, and coverage probabilities (close to or equal to 100%), are the highest among the estimators, which can be attributed to the increase in variability from estimating the four additional parameters in the model that imposes no compound exclusion restriction. For the larger sample sizes, the MLE estimator has the smallest MSE and interval length although it has the smallest coverage (85.4% for  $n = 500$  and 69.8% for  $n = 1000$ ) of all the estimators.

Letting  $p_k = P(R_i = 1 | Z_i = 0, C_i = c, Y_i = k)$  for  $k = 0, 1$ , the odds ratio  $r = p_1(1 - p_1)/p_0(1 - p_1)$  is a simple measure of deviation from the latent ignorability assumption (where

$r$  and  $\gamma_{0c}$  determine the values of  $p_0$  and  $p_1$ ); we violate this assumption in the data by letting  $r = 1/2$  where results are reported in the last three columns of Table 1. For the smallest sample size ( $n = 100$ ) the MI estimators have the smallest MSE and tightest confidence intervals where the MOM estimator has the smallest bias although it has the largest MSE and widest confidence interval. For larger samples ( $n \geq 500$ ) the MLE and MI<sub>1</sub> estimators had similar MSE, bias, and coverage probabilities, where the MI<sub>2</sub> had the smallest bias but the largest interval length and coverage (close to or equal to 100%) of all the estimators.

In summary, for smaller sample sizes MI<sub>1</sub> tended to perform better than the MLE and moment estimator, where MI<sub>2</sub>

**Table 2**  
Influenza vaccine data

$R = 1, Z = 0$	$Y = 0$	$Y = 1$	Total
D = 0	573	49	622
D = 1	143	16	159
Total	716	65	781
$R = 1, Z = 1$	$Y = 0$	$Y = 1$	Total
D = 0	499	47	546
D = 1	256	20	276
Total	755	67	822
$R = 0, Y = .$	$D = 0$	$D = 1$	Total
Z = 0	492	17	509
Z = 1	497	9	506
Total	989	26	1015

Note:  $Y=1$  if there was a hospitalization, 0 if there was no hospitalization, and . if missing;  $R=1$  if outcome  $Y$  was observed and 0 otherwise;  $Z=1$  if randomized to the intervention group and 0 if randomized to the control;  $D=1$  if the patient received the flu vaccine and 0 otherwise.

**Table 3**  
Summary statistics for the influenza vaccination study

	MOM		MLE		MI <sub>1</sub>		MI <sub>2</sub>	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
$ITT_c \equiv CACE$	0.008	(0.141)	-0.007	(0.112)	-0.037	(0.121)	-0.288	(0.378)
$ITT_n$	0	0	0	0	0	0	0.020	(0.025)
$ITT_a$	0	0	0	0	0	0	-0.035	(0.050)
$E[Y_i(1) C_i = c]$	0.034	(0.059)	0.031	(0.054)	0.075	(0.053)	0.098	(0.110)
$E[Y_i(0) C_i = c]$	0.026	(0.128)	0.038	(0.098)	0.111	(0.111)	0.386	(0.320)
$E[Y_i(1) C_i = n]$	0.086	(0.012)	0.086	(0.012)	0.085	(0.013)	0.085	(0.012)
$E[Y_i(0) C_i = n]$	0.086	(0.012)	0.086	(0.012)	0.085	(0.013)	0.064	(0.019)
$E[Y_i(1) C_i = a]$	0.101	(0.024)	0.101	(0.024)	0.086	(0.020)	0.065	(0.044)
$E[Y_i(0) C_i = a]$	0.101	(0.024)	0.101	(0.024)	0.086	(0.020)	0.099	(0.024)
$E[R_i(1) C_i = c]$	1.073	(0.181)	1.000	(0.048)	0.988	(0.015)	0.943	(0.040)
$E[R_i(0) C_i = c]$	1.070	(0.552)	0.885	(0.220)	0.747	(0.227)	0.348	(0.289)
$E[R_i(1) C_i = n]$	0.523	(0.015)	0.523	(0.015)	0.530	(0.016)	0.523	(0.015)
$E[R_i(0) C_i = n]$	0.523	(0.015)	0.523	(0.015)	0.530	(0.016)	0.580	(0.030)
$E[R_i(1) C_i = a]$	0.903	(0.022)	0.926	(0.018)	0.928	(0.017)	0.976	(0.024)
$E[R_i(0) C_i = a]$	0.903	(0.022)	0.926	(0.018)	0.928	(0.017)	0.903	(0.022)
$Pr(C_i = c)$	0.069	(0.023)	0.084	(0.015)	0.101	(0.079)	0.092	(0.066)
$Pr(C_i = n)$	0.797	(0.020)	0.783	(0.011)	0.767	(0.061)	0.774	(0.047)
$Pr(C_i = a)$	0.134	(0.012)	0.134	(0.009)	0.132	(0.020)	0.134	(0.020)

had smaller bias than  $M_1$  when the exclusion restriction was violated. For larger sample sizes the MLE and  $M_1$  estimators tended to perform the best.

## 7. Influenza Vaccination Study

We return to the influenza vaccination study that motivates these methods (where the data are provided in Table 2). Since the study did not maintain records on the clustering of patients by doctor, we ignore this for the purposes of illustration. Estimation of the CACE under the model that imposes no exclusion restrictions may more realistic in this setting. For example, the subpopulation of always-takers, who received a flu shot regardless of their physician's treatment assignment, will more likely be at a higher risk of the flu. Here the exclusion restriction on outcomes for this particular subgroup may be violated if, for example, the letter prompts the physician to take other measures beyond the influenza vaccine like suggesting that the patient avoid unnecessary exposure to certain things or providing the vaccine earlier than they would have had they not received the reminder.

The MLE, MOM,  $M_1$ , and  $M_2$  estimators were used to estimate the effect of the influenza vaccine on flu-related hospitalizations. For the  $M_1$  and  $M_2$  estimators,  $m = 10$  imputations were drawn from the 100,000 iterations of the data augmentation algorithm, subsampling every 10,000th observation. The Jeffrey's priors were used for the outcome and compliance type distributions: Beta(0.5,0.5) and Dirichlet(0.5). Table 3 reports the mean and standard error (S.E.) for the parameters of the model.

### 7.1 Effectiveness of the Influenza Vaccine

The estimated CACE (which is the ITT effect among the subpopulation of compliers) is a percent reduction in flu-related hospitalizations which is close to zero for the MOM and ML estimators, 3.7% for the  $M_1$  estimator (estimated un-

der the model assuming the compound exclusion restriction), and 28.8% for  $M_2$  (estimated under no compound exclusion restriction). Interestingly enough this indicates that the compound exclusion restriction plays a significant role in estimating the CACE in this particular setting although the high standard errors for  $M_2$  indicate that there is no statistically significant reduction in flu-related hospitalizations, due to the flu vaccine, among compliers. When the compound exclusion restriction is not imposed, the ITT effect among the never-takers and always-takers is 2.0% and -3.5%, respectively, indicating no significant effect of *assignment* to treatment on flu-related hospitalizations among these subpopulations (e.g., the compound exclusion restrictions for never-takers and always-takers may hold in this setting). In summary, the data provide no evidence that the flu vaccine significantly reduces flu-related hospitalizations.

The estimated fraction of missing information,  $\hat{\lambda}$ , for  $M_1$  and  $M_2$  for all estimands reported in Table 3 ranged from 0.25 to 0.95 which, although very high, was expected due to the large amount of missing data: hospitalization outcomes were missing (where missingness was nonignorable and dependent upon latent compliance type) for 39% of the subjects, and compliance type was missing for 53% of the subjects.

## 8. Discussion

Here we derived multiple imputation estimators for the CACE in a randomized clinical trial with crossover noncompliance and outcome nonresponse, showing that for smaller sample sizes, multiple imputation may be more favorable as compared to ML or moment methods, for estimating the CACE. For situations where the exclusion restriction may be unrealistic, an appropriate alternative to the likelihood and moment methods is the multiple imputation estimator that imposes no exclusion restrictions, which has less bias but higher variability as a result of having to estimate additional parameters.

We acknowledge that the SUTVA may not hold in a setting that involves an infectious disease such as the influenza virus (Halloran and Struchiner, 1995); however if the study population is a small enough sample from a larger susceptible population and there is little contact between trial participants, which is a reasonable assumption in this study, then the SUTVA should approximately hold.

Extensions of these methods to include covariates is straightforward, where conditioning on covariates would make the latent ignorability assumption more plausible. Covariate information could also help decrease the uncertainty in predicting compliance principal strata, thereby reducing variability in the remaining estimates (Frangakis et al., 2002), as well as decreasing the fraction of missing information in the multiple imputation inference. With regards to the multiple imputation procedures used in this article, although there is controversy on the combining rules of Rubin (1987) under potentially misspecified imputation models (Meng, 1994; Fay, 1996; Rao, 1996; Rubin, 1996; Robins and Wang, 2000), Little (2006) comments that the combining rules are based on Bayesian principles, whereas the criticisms focus on frequentist issues like unbiased estimation of sampling variance, and hence the debate is more on the level of the underlying philosophy of inference. In this article we chose to use Rubin's combining rules, although it would be interesting to see how inferences change using other methods, such as those developed by Robins and Wang (2000). Although we focus on the MI estimator fully imposing or not imposing the compound exclusion restriction, it is possible to relax the compound exclusion restriction partly, so that it holds for always-takers only or never-takers only, for the purposes of a more thorough sensitivity analysis, as in Hirano et al. (2000). The development of methods that account for noncompliance and outcome nonresponse and which use fewer assumptions is an important feature of the multiple imputation methods in this setting and an important area for further research.

## REFERENCES

- Barnard, J., Du, J., Hill, J., and Rubin, D. B. (1988). A broader template for analyzing broken randomized experiments. *Sociological Methods and Research* **27**, 285–317.
- Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.
- Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **8**, 365–379.
- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.
- Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002). Clustered encouragement design with individual noncompliance: Bayesian inference and application to advance directive forms. *Biostatistics* **3**, 147–164.
- Halloran M. E. and Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6**(2), 142–151.
- Heitjan, D. F. (1999). Ignorability and bias in clinical trials. *Statistics in Medicine* **18**, 2421–2434.
- Hirano, K., Rubin, D. B., and Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**(1), 69–88.
- Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960.
- Imbens, G. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25**(1), 305–327.
- Levy, D. E., O'Malley, A. J., and Normand, S.-LT. (2004). Covariate adjustment in clinical trials with non-ignorable missing data and non-compliance. *Statistics in Medicine* **23**, 2319–2339.
- Little R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician* **60**, 213–233.
- McDonald, C. J., Hui, S. L., and Tierney, W. M. (1992). Effect of computer reminders for influenza vaccination on Morbidity During Influenza Epidemics. *M.D. Computing* **9**, 304–312.
- Meng, X.-L. (1994). Multiple imputation inferences with unconditional sources of input. *Statistical Science* **9**, 538–573.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (Translated in 1990.) *Statistical Science* **5**, 465–480.
- O'Malley, A. J. and Normand, S. L. (2005). Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics* **61**, 325–334.
- Rao, J. N. K. (1996). On variance estimation with imputed data. *JASA* **91**, 499–506.
- Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.
- Robins, J. M., Greenland, S., and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* **94**, 687–700.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Education and Psychology* **66**, 688–701.
- Rubin, D. B. (1978). Bayesian inference for causal effects. *Annals of Statistics* **6**, 34–58.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.
- Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research* **33**, 545–571.



- Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16**, 1550–1566.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *JASA* **82**, 528–550.
- Taylor, L. and Zhou, X. H. (2008). Relaxing latent ignorability in the ITT analysis of randomized studies with missing data and noncompliance. UW Biostatistics Working Paper Series. Working Paper 257. To appear in *Statistica Sinica*.
- Zhou, X. H. and Li, S. (2005). ITT analysis of randomized encouragement design studies with missing data. *Statistics in Medicine* **23**, 1991–2003.

*Received March 2007. Revised January 2008.*

*Accepted January 2008.*

