Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes

FABRIZIA MEALLI

Dipartimento di Statistica 'G. Parenti', Università di Firenze, Italy

GUIDO W. IMBENS

University of California at Berkeley, Department of Economics and Department of Agricultural and Resource Economics, Berkeley, CA 94720-3880, USA
imbens@econ.berkeley.edu

SALVATORE FERRO[†]

Istituto Oncologico Romagnolo, Faenza, Italy

ANNIBALE BIGGERI

Dipartimento di Statistica 'G. Parenti', Università di Firenze, Italy

SUMMARY

Recently, instrumental variables methods have been used to address non-compliance in randomized experiments. Complicating such analyses is often the presence of missing data. The standard model for missing data, missing at random (MAR), has some unattractive features in this context. In this paper we compare MAR-based estimates of the complier average causal effect (CACE) with an estimator based on an alternative, nonignorable model for the missing data process, developed by Frangakis and Rubin (1999). We also introduce a new missing data model that, like the Frangakis–Rubin model, is specially suited for models with instrumental variables, but makes different substantive assumptions. We analyze these issues in the context of a randomized trial of breast self-examination (BSE). In the study two methods of teaching BSE, consisting of either mailed information about BSE (the standard treatment) or the attendance of a course involving theoretical and practical sessions (the new treatment), were compared with the aim of assessing whether teaching programs could increase BSE practice and improve examination skills. The study was affected by the two sources of bias mentioned above: only 55% of women assigned to receive the new treatment complied with their assignment and 35% of the women did not respond to the post-test questionnaire. Comparing the causal estimand of the new treatment using the MAR, Frangakis–Rubin, and our new approach, the results suggest that for these data the MAR assumption appears least plausible, and that the new model appears most plausible among the three choices.

[†]Currently at Ospedale S. Orsola-Malpighi, Bologna.

1. INTRODUCTION

Estimating causal effects of interventions is often the focus of empirical studies in medicine and the social sciences. For the most part, randomized experiments are the only generally accepted tools for causal inference. Yet even randomized experiments may suffer from a number of complications that may compromise the study and require additional assumptions. Two such key complications are non-compliance (experimental units receiving treatment levels other than that assigned to them under randomization) and missing outcome information.

Considering the first complication, when compliance is imperfect but the outcome is observed for each subject, the biases associated with estimating the causal effect 'as-treated' (where subjects are compared by treatment received rather than by treatment assigned) or 'per protocol' (where only outcomes for subjects who comply with their assignment are analyzed) are well known (Robins and Greenland, 1994; Sheiner and Rubin, 1995; Barnard *et al.*, 1998). To avoid such potential biases in imperfect compliance cases researchers typically focus on the global intention-to-treat (ITT) effect (comparing all units by their assignment rather than by the treatment actually received). More recently researchers have also focused on the ITT effect for the subpopulation of compliers (units who always comply with their treatment assignment) (Bloom, 1984; Sommer and Zeger, 1991; Imbens and Angrist, 1994; Angrist *et al.*, 1996; Imbens and Rubin, 1997b; Baker, 1998, 2000; Little and Yau, 1998). Such analyses require that researchers be able to identify compliers by exploiting appropriate instrumental variables exclusion restrictions.

When compliance is imperfect and the outcomes are not observed for all units, an ITT analysis based only on complete observations can lead to biased causal estimates, depending on the missing data process. Only if the data are missing completely at random is such a complete data ITT analysis justified (Little and Rubin, 2002). The missing completely at random assumption, which is fairly strong and in many cases implausible, has several testable implications and can often be rejected. Another, potentially more plausible assumption is the missing at random (MAR) model proposed by Rubin (1976) and discussed in (Little and Rubin, 2002). A third alternative missing data model, which is more specifically designed for instrumental variables settings and explicitly allows for a nonignorable missing data process, is that proposed by Frangakis and Rubin (1999). In this paper we propose another missing data model that, like the Frangakis and Rubin approach, is specifically designed for the class of randomized experiments with imperfect compliance.[†] The critical assumption in the Frangakis–Rubin model requires that those subjects who are unwilling to take the new treatment when assigned to it would have the same response behavior irrespective of whether they were assigned to the new treatment or to the standard one. Note that here we are discussing an individual's *response* behavior (in this case the likelihood that they will respond to the post-treatment questionnaire), not their outcome behavior, although the latter is only observable if the individual chooses to respond. Our proposed model replaces this assumption with the alternative restriction that those subjects who *always* comply with their assignment (whether to the new or standard treatment) are not affected in their response behavior by their assignment. Neither of these assumptions is directly testable; however under certain circumstances our newly introduced alternative may be more plausible than that required by Frangakis and Rubin.

In this paper we discuss the relative merits of our proposed missing data model, in particular comparing this new model with MAR and Frangakis–Rubin. To do this we apply these various models to data from a randomized trial of breast self-examination (BSE) conducted in Faenza, Italy that suffered

---

[†]Both the Frangakis–Rubin and the model proposed in the current paper differ from the class of nonignorable missing data models discussed by Scharfstein *et al.* (1999). The Frangakis–Rubin model and the current model impose some additional restrictions on the MAR while relaxing others in a way that makes them estimable, whereas in the Scharfstein–Rotnitzky–Robins setup the nonignorable model does not impose additional restrictions and hence can only be estimated if some of the parameters are fixed *a priori*.

from substantial noncompliance and item nonresponse. Given these data, the goal of this present paper is to estimate the ITT effect for compliers on BSE practice using each of these three models, and to compare the results to assess the appropriateness of the various missing data approaches. Section 2 describes the Faenza randomized trial on BSE, introduces notation, and uses this notation to list some summary statistics from the Faenza study. In Section 3 we introduce our assumptions regarding compliance behavior and compare them to alternative (ignorable and nonignorable) missing data mechanisms. Sections 4 and 5 present a parametric model specification and parallel estimation results, respectively. Section 6 concludes.

2. THE FAENZA RANDOMIZED TRIAL ON BREAST SELF-EXAMINATION

Breast self-examination remains the most controversial of commonly recommended procedures for breast cancer screening. The rationale behind expanding BSE use as a screening mechanism stems from the fact that breast cancer is frequently detected by women themselves without any other symptoms. Although BSE is simple, noninvasive, and inexpensive, its effectiveness is heavily debated in spite of more than 30 years of research (Baxter, 2001; Spurgeon, 2001; Miller and Baines, 2001). Despite these controversies, many field trials have been undertaken to evaluate the effectiveness of methods for teaching BSE techniques, particularly in developing countries. These studies usually compare a BSE class to alternative forms of health education, for instance physician message or informational leaflets; quality of self-exam execution and BSE practice are the two outcomes most often considered (Kalichman *et al.*, 2000; Ortega-Altamirano *et al.*, 2000; Strickland *et al.*, 1997; Mishra *et al.*, 1998; Giles *et al.*, 2001).

In this paper we will consider one such study in which two BSE teaching methods were compared: a 'standard' treatment of receiving mailed information only, and an 'enhanced' treatment of additional attendance in a self-exam course. The study took place between January 1988 and December 1990 at the Oncologic Center of the Faenza Health District in Italy (see previous analysis by Ferro *et al.*, 1996). Both treatment levels were selected on the basis of their practical feasibility and their acceptability according to the cultural profile of the area.

In this experiment a random sample of 825 women, aged 20 to 64 years, was drawn from the demographic files of the city of Faenza. The sample was stratified by age and excluded women with a current breast pathology, a history of breast cancer, a mental or physical disorder, or a terminal illness. Of the 825 women selected, 168 declined participation. The remaining 657 women completed a self-administered pretest questionnaire aimed at evaluating their knowledge of breast pathophysiology, risk factors for breast cancer, preventive beliefs, practice of BSE, and other individual characteristics. This is the population of interest for the purposes of our study.

Responders to the pretest BSE questionnaire were randomly assigned to either a new, enhanced, teaching treatment (330) or to a standard treatment group (327). The standard treatment consisted of receiving information about BSE in the mail in the form of a leaflet containing theoretical as well as graphical material describing how to perform BSE correctly. In contrast, women assigned to the new enhanced treatment group received both the mailed information, and in addition were invited to the Faenza Oncologic Center to receive a 'hands-on' training course on BSE techniques. The course was held by specialized medical staff and consisted of a one hour theoretical session, a group discussion and a fifteen-minute individual practice session. Of the 330 women randomly assigned to the enhanced treatment, only 182 complied with their assignment, i.e. attended the course. Thus only 55% of the women assigned to the enhanced treatment complied with their assignment; the remainder receiving only the standard treatment of the mailed information. One year later the knowledge level of each woman was assessed by the same procedure used at the start of the study, namely by a self-administered questionnaire. Of the 657 women included in the study only 429 (65%) of the total population completed this questionnaire, providing information on post-treatment BSE practice. This is likely partly due to the fact that the outcome data

were collected at a later date than the covariate and assignment data.

In this study the question of interest was the effect of an enhanced training class on BSE practices. However, due to noncompliance in treatment assignment and the lack of outcome data for a substantial proportion of the initial study group, this study offers us a good opportunity to compare intention-to-treat causal estimators based on various missing data models.

2.1 *Notation*

In order to address the noncompliance and missing data problems discussed above, let us first introduce some notation. The study presented above is a two-arm randomized experiment that compares a new, enhanced teaching treatment to a standard treatment; access to the new training course is only available to those in the enhanced treatment group, and compliance is all-or-nothing. For each individual i who participates in the study, let Z_i represent their treatment assignment: $Z_i = 1$ for those assigned to the enhanced treatment, $Z_i = 0$ for those assigned the standard treatment. In addition, let $D_i(z)$ be an indicator for the treatment received, given assignment z, and let $D_i = D_i(Z_i)$ be the actual treatment received. Thus if $D_i = 1$, individual i participated in the training course; $D_i(0) = 0$ by definition, as those assigned to the standard treatment had no access to the training course. Similarly, define $Y_i(z)$ as the potential outcome, given assignment to treatment level z, and let $Y_i = Y_i(Z_i)$ be the actual outcome observed. Lastly, let $R_i(z)$ represent the potential response indicator (1 if a subject responds to the post-test questionnaire, 0 for non-responders), given treatment z, and let $R_i = R_i(Z_i)$ represent the actual response indicator. In addition, a vector of pretreatment variables, $\mathbf{X}_i$, is observed per subject. Hence, the observed data are:

$$\{Z_i, D_i, R_i, \mathbf{X}_i, (Y_i : R_i = 1), i = 1, \ldots, N\}.$$

Following Frangakis and Rubin (1999), let U_i represent the treatment woman i would receive if assigned to the active treatment ($U_i = D_i(1)$). If $U_i = 1$ then woman i is a 'complier'; among these individuals $D(1) = 1$ (as observed), and by the structure of the experimental setting, had they instead been assigned to the standard treatment, $D(0) = 0$, by definition. Thus for these units $D_i = Z_i$: they always comply with their treatment assignment. In contrast, if $U_i = 0$ this individual is a 'never-taker'; by the structure of the experiment she could not select into it if assigned to the standard treatment. Thus among this subset $D_i(z) = 0$, for both $z = 0$ and 1. As explained in Figure 1 for this experimental setting, this compliance status U_i can be viewed as a covariate which is observed only for women with $Z_i = 1$ (Angrist *et al.*, 1996); by randomization, however, it is guaranteed to have the same distribution in both treatment arms. This latent compliance covariate plays a crucial role in the causal estimation models discussed in this paper. Since 'never-takers' are never observed exposed to the new treatment, it is only for 'compliers' that we can hope to learn anything about the effect of the new treatment. Even for 'compliers', however, inferring causal effects of the treatment is controversial (Angrist *et al.*, 1996).

2.2 *Summary statistics*

Table 1 presents some summary statistics[†] for the sample of 657 women included in the Faenza study, grouped by assignment Z_i and treatment received D_i. The outcome Y_i listed in Table 1 is a binary indicator of whether BSE was practiced a year after the treatment was received. Later we will discuss a second outcome indicator of whether the quality of an individual's post-treatment BSE practice exceeds a designated threshold. This quality measure is in fact a compilation of different practice indicators,

---

[†]For a complete descriptive data analysis, including a full description of the variables obtained from the questionnaire see Ferro *et al.* (1996)

Z_i	binary assignment (observed)
D_i	binary receipt of enhanced treatment (observed)
Y_i	binary outcome: BSE practice (observed if respondent)
R_i	binary response indicator for outcome (observed)
U_i	binary indicator for compliance type (complier/nevertaker, observed if assigned to enhanced treatment)
X_i	vector of pretreatment variables (observed): prior BSE practice knowledge of pathophysiology

Fig. 1. Notation

resulting in a variable that can take on integer values between 0 and 21 (see Ferro *et al.*, 1996 for details on this variable). As suggested in other works (Ferro *et al.*, 1996; Miller and Baines, 2001), in this analysis we consider a binary quality outcome variable Q_i equal to 1 if an individual's quality indicator is greater than the overall study sample median (in this case 17) and 0 otherwise.

The observed baseline covariates are X_{i1}, a binary indicator of previous BSE practice, X_{i2}, a binary indicator of good knowledge of breast pathophysiology and X_{i3}, age. As can be seen in columns (2) and (3) of Table 1, due to randomization pretreatment variables are well balanced in the two subsamples defined by treatment assignment. Randomization does not, however, imply that the pretreatment variables will be balanced in the subsamples defined by the actual treatment *received* (columns (5) and (6)). Note that $D_i = 1$ implies $Z_i = 1$, thus column (5) ($Z_i = 1, D_i = 1$) includes all observations with $D_i = 1$. Column (6) similarly includes all observations who did *not* attend the course, including all women assigned to the standard treatment and as well as those assigned to the enhanced treatment who chose not to comply. Knowledge of breast pathophysiology (X_{i2}) prior to the program, for example, is well balanced in the two groups defined by treatment assignment (i.e. for $Z_i = 1$ and $Z_i = 0$), as expected. However, the knowledge is signicantly higher for those women who attended the course ($D_i = 1$) than for those who did not ($D_i = 0$), suggesting that simply comparing outcomes by treatment *received*, as one would do with a naive 'per-protocol' or 'as-treated' analysis, is invalid (e.g. Sheiner and Rubin, 1995), and in this case likely to overstate the effect of the program.

Concerning the response behavior (R_i), we observe that within the group assigned to receive the active treatment, response rates significantly differ between those who complied with their assigned treatment ($D_i = 1$) and those who did not ($D_i = 0$). This suggests that the compliance covariate, U_i, may be related to the willingness to respond, R_i, of the subjects. In addition, Table 2 shows that the levels of BSE practice, as well as the complete-case ITT estimates that can easily be derived, differ a lot depending on the values of the pretreatment variables. This stresses the importance of balancing observed and unobserved covariates in the analysis of causal effects.

As Barnard *et al.* (1998) point out ' ... reasons for missing outcomes can be different for compliers and never-takers, and also, can be affected by treatment assignment, creating even more disparity between the types of people being compared.' Frangakis and Rubin (1999) have shown that in such cases the complete-case (or respondent-based) ITT estimator (with an estimate of -0.021 (s.e. 0.040) in the present study) is generally biased for the ITT effect.

Table 1. *Faenza BSE study—summary statistics*

	Means					
	1	2	3	4	5	6
	All	$Z_i = 0$	$Z_i = 1$	$Z_i = 1$ $D_i = 0$	$Z_i = 1$ $D_i = 1$	$D_i = 0$
N	657	327	330	148	182	475
Assignment (Z_i)	0.502	0	1	1	1	0.312
Course attendance (D_i)	0.277	0	0.551	0	1	0
Response (R_i)	0.653	0.688	0.618	0.399	0.797	0.598
BSE practice (Y_i)*	0.785	0.796	0.774	0.475	0.897	0.729
BSE quality (Q_i)*	0.493	0.402	0.595	0.250	0.669	0.382
Pretreatment variables:						
BSE practice (X_{i1}) **	0.585	0.591	0.579	0.551	0.601	0.579
Knowledge of breast pathophysiology (X_{i2})	0.554	0.560	0.548	0.439	0.637	0.522
Age (X_{i3})	41.4	41.5	41.3	41.7	41.0	41.6
Complete Data ITT on Y		$-0.021 (0.040)$				

*Computed on respondents only. **Available for 615 women.

Table 2. *Means of outcome Y for respondents, by value of pretreatment variables, treatment assigned and treatment received*

	Means				
	$Z_i = 0$	$Z_i = 1$	$Z_i = 1$ $D_i = 0$	$Z_i = 1$ $D_i = 1$	$D_i = 0$
N	327	330	148	182	475
Pretreatment variables:					
BSE practice yes ($X_{i1} = 1$)	0.906	0.891	0.759	0.939	0.881
BSE practice no ($X_{i1} = 0$)	0.589	0.611	0.178	0.824	0.475
Knowledge of breast pathophysiology yes ($X_{i2} = 1$)	0.837	0.875	0.714	0.921	0.815
Knowledge of breast pathophysiology no ($X_{i2} = 0$)	0.740	0.600	0.258	0.840	0.622

3. ASSUMPTIONS REGARDING COMPLIANCE AND RESPONSE BEHAVIOR

3.1 *Noncompliance*

The summary statistics displayed in Tables 1 and 2 suggest that an ITT analysis of the Faenza experimental results, or other naive alternatives, such as the 'as-treated' or 'per-protocol' analysis, will potentially provide very misleading results, since the treatment received is correlated with important pretreatment variables. The 'as-treated' analysis would compare individuals by the actual treatment received (columns (5) and (6)), while the 'per-protocol' analysis would compare women who did receive the treatment (column (5)) with the standard treatment group (column (2)). However, in both cases pretreatment covariates are not well balanced between the two groups; in particular the women who receive the active

treatment in fact practice BSE more often and have a better knowledge of breast pathophysiology even before the treatment is administered, even though these variables are well balanced in the two assignment groups.

In order to address the noncompliance problem we consider the following assumption.

ASSUMPTION 1 (EXCLUSION RESTRICTION FOR NEVER-TAKERS)

$$Y_i(Z_i) \perp Z_i | \mathbf{X}_i, U_i = 0.$$

This assumption implies that $\Pr(Y_i(1)|\mathbf{X}_i, U_i = 0) = \Pr(Y_i(0)|\mathbf{X}_i, U_i = 0)$, so that for subpopulations of never-takers with the same covariate values, the distributions of the two potential outcomes are the same. This is a type of instrumental variables assumption because it rules out a direct effect of the assignment on the outcome for this specific subpopulation. In the absence of non-response, this exclusion restriction would allow us to identify the ITT effect for the subpopulation defined by the compliance status covariate U_i, namely the ITT effect for compliers:

$$ITT_C = E[Y(1) - Y(0)|U = 1]$$

without any further assumptions[†] (Angrist *et al.*, 1996; Imbens and Rubin, 1997b).

Assumption 1 has some testable restrictions in the form of inequalities stemming from the implication that those assigned to the standard treatment are a mixture of never-takers and compliers with the outcome distribution for never-takers identical to that for never-takers assigned to the new treatment (Imbens and Rubin, 1997a; Balke and Pearl, 1997). In order to estimate models that relax this never-taker exclusion restriction, however, it is typically useful to make additional assumptions, such as imposing some parametric form on the likelihood function, or using informative prior distributions within a Bayesian approach (Hirano *et al.*, 2000). We will return to this issue later.

In our experimental setting the ITT_C effect is the only ITT effect that potentially addresses the causal effect of the receipt of the enhanced treatment, because it compares outcomes under the new treatment with those under the standard one. At least in this study, and especially under the exclusion restriction requiring that for never-takers there is no direct effect of the assignment, it seems plausible to attribute the effect of assignment for the compliers to the effect of the receipt of the treatment. This ITT effect is sometimes referred to as the local average treatment effect (LATE, Imbens and Angrist, 1994) or the complier average causal effect, CACE, as we will do in the remainder of the paper.

3.2 *Nonresponse*

We now review the two principal models that have been proposed in the literature to address the missing data problem observed in the Faenza study and in many like it. The first model assumes that the outcome variable Y is MAR (Rubin, 1976): the probability of observing Y is the same for all subjects with the same value of the observed covariates, treatment assigned and treatment received. In our case the i.i.d. MAR assumption can thus be stated as follows.

ASSUMPTION 2 (MISSING AT RANDOM)

$$Y_i \perp R_i | Z_i, \mathbf{X}_i, D_i.$$

[†]Intuitively, we can easily obtain an estimate of $E(Y(1)|U = 1)$ as the sample mean of Y for subjects assigned to the new treatment who complied with this assignment. We can also obtain an estimate of the proportion of compliers π^U, as the sample proportion of those assigned to the new treatment who complied, and an estimate of $E(Y(0)|U = 0)$, because, from the exclusion restriction, this is equal to $E(Y(1)|U = 0)$, which can be easily estimated by the sample mean of Y for subjects assigned to the new treatment who chose not to receive it. Finally, an estimate of $E(Y(0)|U = 1)$ can be obtained by exploiting the following equality: $E(Y(0)) = E(Y(0)|U = 1)\pi^U + E(Y(0)|U = 0)(1 - \pi^U)$, where $E(Y(0))$ is estimated by the sample mean of Y for subjects assigned to the standard treatment.

This assumption implies that $\Pr(R_i|Y_i, Z_i, \mathbf{X}_i, D_i) = \Pr(R_i|Z_i, \mathbf{X}_i, D_i)$. If the parameters of the missing data mechanism are distinct from those of the outcome data distribution, the missing data mechanism is said to be ignorable (Rubin, 1976; Little and Rubin, 2002). In terms of the response behavior of never-takers and compliers, this assumption implies that these two groups may have different response behavior in the new treatment arm (since their D_i would differ), but the same behavior in the standard one (since their D_i would be the same). Under this assumption it is the treatment received—which is a deterministic function of Z_i and U_i—rather than the true compliance covariate U_i, that determines the response behavior. Such a situation might arise if the intervention itself leads to dissatisfaction and reduces the willingness to respond. (This particular argument does not work with the data as the willingness to respond is higher for those who receive the enhanced program than for those who do not, but similar arguments could work.) Unlike the exclusion restriction of Assumption 1, the MAR assumption has no testable implications allowing us to assess its plausibility in a specific experimental setting. Given these two assumptions we can estimate the causal effect of enhanced BSE training without additional restrictions.

A special case of MAR arises when

$$R_i \perp Z_i, \mathbf{X}_i, D_i$$

and the data are said to be missing completely at random. Under this assumption limiting the analysis to only those complete observations would result in no bias.[†] In comparison to the more general MAR case, however, this model does have testable restrictions, as it implies that the distribution of covariates $\mathbf{X}_i$ is the same in the complete and incomplete data subsamples.

Before discussing the next model, it is useful to consider what the observed data can show about the response behavior of the two groups of individuals (never-takers and compliers), using a method-of-moment estimation reasoning. Avoiding, for simplicity, the conditioning on the pretreatment covariates $\mathbf{X}_i$, denote $\Pr(R_i = 1|Z_i = z, U_i = u)$ by π_{zu}, which are the four response probabilities for never-takers and compliers in the standard and new treatment arms. From the data, only π_{11} and π_{10} can be estimated directly (as $\hat{\pi}_{11} = \sum R_i Z_i D_i / \sum Z_i D_i$ and $\hat{\pi}_{10} = \sum R_i Z_i (1 - D_i) / \sum Z_i (1 - D_i)$) using individuals assigned to the active treatment. For the standard treatment group only the mixture $\Pr(U_i = 1)\pi_{01} + (1 - \Pr(U_i = 1))\pi_{00}$ can be estimated (as $\sum R_i (1 - Z_i) / \sum (1 - Z_i)$), since we cannot identify the 'true' compliance status for subjects in this group. If one knew (or could estimate) either π_{01} or π_{00} *a priori*, this information could be used to estimate both π_{01} and π_{00} separately. But given information only on the mixture, one has to rely on assumptions—that are not testable without auxiliary information—in order to disentangle the information to get separate estimates of π_{01} and π_{00}. Some such assumptions are implicit or explicit elements of the following models proposed in the literature to address the missing outcome data problem.

An alternative to MAR, proposed by Frangakis and Rubin (1999), makes use of the compliance covariate U_i.

ASSUMPTION 3 (LATENT IGNORABILITY)

$$Y_i \perp R_i | Z_i, \mathbf{X}_i, U_i.$$

Under this assumption potential outcomes and potential nonresponse indicators are independent within subpopulations of the same compliance covariate and pretreatment/assignment levels. This assumption implies that

$$\Pr(Y_i, R_i|Z_i, \mathbf{X}_i, U_i) = \Pr(Y_i|Z_i, \mathbf{X}_i, U_i) \cdot \Pr(R_i|Z_i, \mathbf{X}_i, U_i)$$

[†] As the analysis is always conditional on Z, the complete data analysis would not be biased also under the weaker assumption $R_i(0) \perp \mathbf{X}_i, D_i(0)$ and $R_i(1) \perp \mathbf{X}_i, D_i(1)$, which implies that $\Pr(R_i|Y_i, Z_i, \mathbf{X}_i, D_i) = \Pr(R_i|Z_i)$.

so that if U_i were observed for all subjects and the parameters of the missing data process are distinct from those of the outcome distribution the missing data process would be ignorable. But because the true compliance covariate is unobserved for those individuals assigned to the standard treatment, the missing data process is in fact nonignorable.

On its own, therefore, this assumption is not sufficient to identify the ITT effect for compliers. To address the complications due to the fact the U_i is only partially observed, different assumptions can be exploited. Frangakis and Rubin propose the following.

ASSUMPTION 4 (RESPONSE EXCLUSION RESTRICTION FOR NEVER-TAKERS)

$$R_i(Z_i) \perp Z_i | \mathbf{X}_i, U_i = 0.$$

Assumption 4 implies that never-takers have the same response behavior irrespective of their treatment assignment, thus that $\Pr(R_i(1)|\mathbf{X}_i, U_i = 0) = \Pr(R_i(0)|\mathbf{X}_i, U_i = 0)$, i.e. $\pi_{00} = \pi_{10}$. Assumption 4 combined with Assumption 1 (that the distribution of the outcome variable for never-takers is also independent of their assignment level) is the stochastic version of the compound exclusion restriction of Frangakis and Rubin (1999). The combination of these assumptions with Assumption 3 (the latent ignorability of the missing data process) will be referred to as the FR model. As explained in Frangakis and Rubin (1999), under this set of assumptions, all the quantities on which the ITT effect for compliers depends have a sample counterpart, and using these counterparts to estimate the complier ITT effects leads to the Frangakis–Rubin estimator.

While the exclusion restriction on the outcome variable (Assumption 1) seems plausible in many circumstances, the response exclusion restriction for never-takers appears more questionable. This is especially the case when we lack a comparable set-up in the standard (control) treatment arm, i.e. no blind placebo-like setting that allows those assigned to the standard treatment to display their complier status along with their response distribution. Never-takers who were assigned to the new treatment and declined participation might in fact easily lower their subsequent response probability. In comparison to those never-takers receiving the standard treatment, their explicit refusal to comply with their assigned (active) treatment may plausibly induce them to refuse to respond in the post-test questionnaire as well.

An alternative to the Frangakis–Rubin response exclusion restriction for never-takers is to assume that compliers do not change their response behavior with assignment.

ASSUMPTION 5 (RESPONSE EXCLUSION RESTRICTION FOR COMPLIERS)

$$R_i(Z_i) \perp Z_i | \mathbf{X}_i, U_i = 1.$$

This assumption implies that compliers have the same response behavior irrespective of the treatment arm they are assigned to. As compliers are willing to follow the protocol in their assigned treatment, it seems more plausible that they would not be affected in their response behavior by that assignment. Under this alternative assumption the missing data process is again not ignorable. We will refer to the combination of this assumption (as a replacement for Assumption 4) with the exclusion restriction for never-takers and the latent ignorability assumption (Assumptions 1 and 3) as the modified FR (MFR) model. Similar to the FR model, it can be easily shown under this set of assumptions that all the quantities on which the ITT effect for compliers depends have a sample counterpart, and using these sample analogues a simple estimator can be derived.

The three sets of Assumptions, 1 and 2 (MAR), 1, 3 and 4 (FR) and 1, 3 and 5 (MFR), have no testable implication beyond inequality restrictions of the type discussed in Imbens and Rubin (1997a) and Balke and Pearl (1997). Yet unless one of these sets of assumptions is made, in a setting containing significant non-compliance and missing outcome data, the model would have no unique maximum likelihood estimates. This same problem is encountered when relaxing Assumption 1 in randomized experiments

that suffer only from non-compliance. However, the presence of observed pretreatment variables may help investigating violations of the various exclusion restrictions inherent in each of these three models (Imbens and Rubin, 1997a).

4. MODEL SPECIFICATION

Method-of-moment estimators are useful to understand where information comes from within the observed data and what assumptions help us identifying estimands of interest. In the presence of covariates, however, method-of-moment estimators are not easily implemented. Here we therefore prefer to use likelihood based estimators, which have been shown to improve upon conventional IV estimators (Imbens and Rubin, 1997a,b; see also Hirano *et al.*, 2000). In particular, we model three things: the conditional distribution of the compliance variable U given the pretreatment variables X, the conditional distribution of the potential outcome Y given X and U, and the conditional distribution of the potential response indicator R, also given X and U. As all the variables of interest are dichotomous, we assume that their distributions have a logistic regression form:

$$\pi_i^U = \Pr(U_i = 1 | \mathbf{X}_i = \mathbf{x}; \alpha) = \pi_i^U = \frac{\exp(\alpha_0 + \alpha_1' \mathbf{x})}{1 + \exp(\alpha_0 + \alpha_1' \mathbf{x})} \qquad (1)$$

$$\Pr(R_i = 1 | \mathbf{X}_i = \mathbf{x}, Z_i = z, U_i = u; \beta) = \pi_{i,z,u}^R = \frac{\exp(\beta_{zu0} + \beta_{zu1}' \mathbf{x})}{1 + \exp(\beta_{zu0} + \beta_{zu1}' \mathbf{x})} \qquad (2)$$

$$\Pr(Y_i = 1 | \mathbf{X}_i = \mathbf{x}, Z_i = z, U_i = u; \gamma) = f_{i,z,u}(1) = \frac{\exp(\gamma_{zu0} + \gamma_{zu1}' \mathbf{x})}{1 + \exp(\gamma_{zu0} + \gamma_{zu1}' \mathbf{x})}. \qquad (3)$$

In the analysis of the Faenza data we impose prior equality of the slope coefficients in the outcome distribution for compliers: $\gamma_{011}' = \gamma_{111}'$.

Under Assumptions 1 and 3 (the exclusion restriction on the outcome variable for never-takers and latent ignorability, respectively), the actual (observed) likelihood function is

$$
\begin{aligned}
L(\theta | \mathbf{Z}, \mathbf{X}, \mathbf{D}, \mathbf{R}, \mathbf{Y}) = \\
\prod_{i:Z_i=1, D_i=1, R_i=1} \pi_i^U \pi_{i11}^R f_{i11}(Y_i) \prod_{i:Z_i=1, D_i=1, R_i=0} \pi_i^U (1 - \pi_{i11}^R) \\
\prod_{i:Z_i=1, D_i=0, R_i=1} (1 - \pi_i^U) \pi_{i10}^R f_{i10}(Y_i) \prod_{i:Z_i=1, D_i=0, R_i=0} (1 - \pi_i^U)(1 - \pi_{i10}^R) \\
\prod_{i:Z_i=0, D_i=0, R_i=1} \left(\pi_i^U \pi_{i01}^R f_{i01}(Y_i) + (1 - \pi_i^U) \pi_{i00}^R f_{i10}(Y_i) \right) \\
\prod_{i:Z_i=0, D_i=0, R_i=0} \left(\pi_i^U (1 - \pi_{i01}^R) + (1 - \pi_i^U)(1 - \pi_{i00}^R) \right).
\end{aligned}
\qquad (4)
$$

The first two factors in the likelihood represent the contribution of the compliers assigned to the treatment, including both respondents and nonrespondents. The second two factors represent the contribution for never-takers assigned to the treatment, including respondents and nonrespondents. The last two factors represent the contribution to the likelihood function for those assigned to the standard treatment. This includes both compliers and never-takers and the likelihood contributions therefore consist of averages over the distribution of compliance types. The assumptions inherent in the MAR, FR and MFR models

can be imposed by using the following restrictions respectively:[†] $\pi_{i00}^R = \pi_{i01}^R$ (MAR), $\pi_{i00}^R = \pi_{i10}^R$ (FR), and $\pi_{i01}^R = \pi_{i11}^R$ (MFR).

In the complete case analysis, because observations with missing outcomes are excluded and response probabilities are assumed to be constant, the likelihood function simplifies to

$$L(\theta|\mathbf{Z}, \mathbf{X}, \mathbf{D}, \mathbf{Y}) =$$
$$\prod_{i:Z_i=1, D_i=1} \pi_i^U f_{i11}(Y_i) \prod_{i:Z_i=1, D_i=0} (1 - \pi_i^U) f_{i10}(Y_i)$$
$$\prod_{i:Z_i=0, D_i=0} \left(\pi_i^U f_{i01}(Y_i) + (1 - \pi_i^U) f_{i10}(Y_i) \right). \tag{5}$$

Using the missing data structure of the likelihood function, maximum likelihood estimates can be obtained using the EM algorithm (Dempster *et al.*, 1977) or via standard maximization routines. In this application the Newton–Raphson algorithm was implemented and standard errors were computed using the Delta method.

5. The effect of the enhanced BSE program on BSE quality

The discussion so far has focused on estimating the effect of the enhanced BSE training program on the use of self-exams. The estimation of causal effects of the enhanced BSE training program on the *quality* of self-exams is more problematic because quality can only be observed for women who practice BSE ($Y_i = 1$); it is not only unobserved but also undefined when $Y_i = 0$. The response to such a problem is often to assume the quality outcome variable as missing or censored, or assigning it a value of zero. Although often done, however, these approaches do not lead to properly defined causal estimands (see Rubin 2000; Frangakis and Rubin 2002 for more discussion on this).

In principle, a causal estimand of interest would be the effect of the treatment on the quality of self-exam for those women who would practice BSE under both assignments ($Y_i(z) = 1$ for $z = 0$ and 1). In a randomized experiment with non-compliance, such a causal estimand would be the effect of the treatment for compliers who would practice BSE under both treatments. The estimation of this causal effect would involve additional assumptions that are not further pursued in the current paper.

Here we will consider only the binary quality outcome Q which assumes the value 1 if the quality indicator is greater than its overall median value (in this sample 17) and 0 otherwise. As with the conditional distributions discussed above, we specify a conditional logit model for this secondary outcome, given pretreatment variables and compliance covariate, and conditional on practicing post-treatment BSE ($Y = 1$):

$$\Pr(Q_i = 1|\mathbf{X}_i = \mathbf{x}, Z_i = z, U_i = u, Y_i = 1) = \frac{\exp(\delta_{zu0} + \delta_{zu1}' \mathbf{x})}{1 + \exp(\delta_{zu0} + \delta_{zu1}' \mathbf{x})},$$

under the additional exclusion restriction for never-takers that the quality of BSE is independent of treatment assignment, given that these women would conduct self-exams under either assignment:

ASSUMPTION 6

$$Q_i(Z_i) \perp Z_i | \mathbf{X}_i, U_i = 0, Y_i = 1.$$

[†]Note that, under MAR, the likelihood factorizes so that parameters of the outcome and compliance status distributions can be estimated independently of the parameters of the response probabilities, i.e. the missing mechanism can be ignored.

This will allow us to obtain an estimate of

$$E[Q(1) - Q(0)|U = 1, Y = 1]. \qquad (6)$$

However, as explained above, this cannot be interpreted as a causal effect because it is conditional on the value of the outcome Y (i.e. of a post-treatment variable).

An additional two assumptions allow us to interpret the conditional mean difference in (6) as a causal effect.

ASSUMPTION 7 (MONOTONICITY OF THE OUTCOME FOR COMPLIERS)

$$Y_i(1) \geqslant Y_i(0)|U_i = 1.$$

ASSUMPTION 8 (QUALITY INDEPENDENCE OF THE PRIMARY OUTCOME UNDER STANDARD TREATMENT FOR COMPLIERS)

$$Q_i(1) \perp Y_i(0)|U_i = 1, Y_i(1) = 1.$$

Assumption 7 effectively states that those who practice post-treatment BSE under the standard treatment would also have done so under the enhanced treatment; thus observing that an individual in the standard treatment conducts post-treatment self-exams implies that she would also have done so if assigned to the enhanced treatment, if she is a complier. Assumption 8 says that the quality of BSE undertaken by those assigned to the enhanced treatment group who practice BSE ($Y(1) = 1$) is the same, irrespective of whether or not these same women would have conducted BSEs under the standard treatment. Assumption 7 cannot be verified directly at the individual level, although we can see some indirect evidence of it from estimating the CACE. Assumption 8 cannot be tested, but if it does not hold, the expected sign of the bias should be towards an underestimation of the real effect on BSE quality for compliers practicing self-exams under both treatment arms. This is because the compliers who practice BSE under the new treatment but who would not do so under the standard one, should plausibly practice BSE with a lower average quality than those who would practice BSE regardless of the assigned treatment. In the application of these assumptions on the Faenza study data we impose the requirement that the logistic parameter $\delta_{zu1} = \mathbf{0}$ in order to reduce the number of parameters[†] in each model.

6. RESULTS

Given the setup discussed above, we first estimate the model using no pretreatment variables. Table 3 shows the estimates of the effect on BSE practice and quality for compliers and the estimates of the four response probabilities under the three different models.[‡] First consider the results under the MAR assumptions, which show a surprising negative effect of the active treatment on BSE practice. Although it is conceivable that the course had little or no effect, it is more difficult to understand how, among a population of volunteers, the effect of the training course was in fact to cause significant decreases in post-treatment BSE compared to those receiving just a mailed informational leaflet. Using the pretreatment variables in Table 4 the alternative models show more plausible, small and non-significant effects of the course on BSE practice. Somewhat surprisingly the complete-case analysis also gives a plausible effect

[†]Given the relatively small sample size, relaxing this restriction, as well as the one on the slope coefficients in the logit for the outcome of compliers, would increase the computational burden and lead to imprecise estimates; in the final models age has been excluded.

[‡]Note that some of the estimated probabilities are identical under the different models, due to the structure of the likelihood function. For example, the proportion of compliers is the same under both the FR and MFR models, but not under the MAR model.

Table 3. *Effects for compliers under various missing data assumptions, without pretreatment variables. Exclusion restriction for never-takers is always maintained. (standard errors in parentheses)*

	Complete data	MAR $\pi_{i00}^R = \pi_{i01}^R$	FR $\pi_{i00}^R = \pi_{i10}^R$	MFR $\pi_{i01}^R = \pi_{i11}^R$
CACE	−0.030	−0.103	−0.012	−0.081
on BSE practice	(0.056)	(0.025)	(0.054)	(0.067)
Compliers' effect on BSE quality	0.235	0.206	0.239	0.225
conditional on BSE practice	(0.063)	(0.077)	(0.062)	(0.067)
π_{01}^R		0.69	0.92	0.80
		(0.026)	(0.062)	(0.030)
π_{11}^R		0.80	0.80	0.80
		(0.030)	(0.030)	(0.030)
π_{00}^R		0.69	0.40	0.55
		(0.026)	(0.040)	(0.066)
π_{10}^R		0.40	0.40	0.40
		(0.040)	(0.040)	(0.040)
π^U	0.71	0.56	0.55	0.55
	(0.032)	(0.026)	(0.027)	(0.027)

π_{zu}^R is the response probability for a woman assigned to treatment z with compliance status u.

on BSE practice of the active treatment (insignificant, but again less than zero), although, in general, one should not rely on such an analysis. Note, for example, that the estimated proportion of compliers under the complete-case analysis differs appreciably from those estimated without excluding nonrespondents.

Consider the estimated response probabilities in Table 3. The FR model (Assumptions 1, 3 and 4) gives figures for compliers' response probabilities that are not very plausible: compliers have a lower response rate if assigned to the new treatment ($\pi_{11} = 0.80$) than if assigned to the standard one ($\pi_{01} = 0.92$). This would only be plausible if we imagine that women assigned to the new treatment arm became frustrated with the extra burden of compliance (attending the course), and hence were more likely to choose to not respond to the post-treatment questionnaire. This seems unlikely given the short duration of the intervention and their volunteer status. In contrast, the MFR model (Assumptions 1, 3 and 5) gives more plausible figures for the response probabilities: per assigned treatment level, never-takers have lower response rates than compliers (0.40 vs. 0.80 for those assigned to the active treatment, 0.55 vs. 0.80 for the standard). In addition, never-takers have a lower response rate if assigned to the new treatment arm than if assigned to the standard treatment. This would agree with the hypothesis that once never-takers show that they are unwilling to follow the assignment protocol, they are less inclined to respond to the survey.

In Table 4 we now condition on the pretreatment variables. In most cases the estimates are now slightly more precise, although the pretreatment variables seem to have little effect on the significance levels of the causal estimands. Adding pretreatment variables in the analysis changes the sign to positive under latent ignorability (the FR and MFR models), although the CACE effect on BSE practice remains insignificant for both. The compliers' effect on BSE quality remains positive and significant.

Table 4. *Effects for compliers under various missing data assumptions, using pretreatment variables. The exclusion restriction for never-takers is always maintained. (standard errors in parentheses)*

	Complete data	MAR	FR	MFR
CACE				
on BSE practice	−0.001	−0.111	0.011	0.024
	(0.053)	(0.026)	(0.049)	(0.054)
Compliers' effect				
on BSE quality	0.239	0.226	0.244	0.227
conditional on BSE practice				
	(0.066)	(0.076)	(0.063)	(0.073)

7. CONCLUDING REMARKS

In this paper, using data from a randomized experiment suffering from substantial non-compliance and missing outcome data, we compare MAR-based estimates of the CACE with estimators based on alternative models for the missing data process, including one developed by Frangakis and Rubin (1999) and a modification of the Frangakis–Rubin model introduced here. Both of these last two models are specifically designed for instrumental variables settings. We illustrate these methods by re-analyzing data of a randomized trial of BSE that compared a standard method of teaching BSE via a mailed information leaflet, to an enhanced treatment of attendance in a course, with the aim of assessing whether teaching programs could increase BSE practice and improve examination skills.

Since the key assumptions of the three models discussed are all untestable, we judge the plausibility of these models by the estimates they produce. The MAR assumptions lead to an implausible significant negative effects of the training course on BSE practice; although one can easily imagine a positive or zero effect of the treatment, it is difficult to understand why a program designed to encourage BSE would have a negative effect. The two alternative models lead to more plausible, insignificant effects. However, of these two, the modified Frangakis–Rubin model leads to a more plausible *a priori* pattern of response rates; those that are potentially unwilling to comply with their assignment are also less likely to respond to the survey, and in particular they are less willing to respond if they have actually declined to participate in the treatment program. Thus we conclude that the MFR model appears to be particularly appropriate in the context of randomized encouragement designs where one-sided non-compliance is an issue, and double blinding is not feasible.

REFERENCES

ANGRIST, J. D., IMBENS, G. W. AND RUBIN, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–472.

BAKER, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association* **93**, 929–934.

BAKER, S. G. (2000). Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance. *Journal of the American Statistical Association* **95**, 43–50.

BALKE, A. AND PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92**, 1171–1176.

BARNARD, J,, DU, J., HILL, J. L. AND RUBIN, D. B. (1998). A broader template for analyzing broken randomized experiments. *Sociological Methods and Research* **27**, 285–317.

BAXTER, N. (2001). Preventive health care, 2001 update: should women be routinely taught breast self-examination to screen for breast cancer? *Canadian Medical Association Journal* **164**, 1837–1846.

BLOOM, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **82**, 225–246.

DEMPSTER, A. P., LAIRD, N. AND RUBIN, D. B. (1977). Maximun likelihood estimation from incomplete data using the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

FERRO, S., CAROLI, A., NANNI, O., BIGGERI, A. AND GAMBI, A. (1996). A randomized trail on breast self-examination in Faenza (Northern Italy). *Tumori* **82**, 329–334.

FRANGAKIS, C. E. AND RUBIN, D. B. (1999). Addressing complications of intention-to treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 365–379.

FRANGAKIS, C. E. AND RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

GILES, J. T., KENNEDY, D. T., DUNN, E. C., WALLACE, W. L., MEADOWS, S. L. AND CAFIERO, A. C. (2001). Results of a community pharmacy-based breast cancer risk-assessment and education program. *Pharmacotherapy* **212**, 243–253.

HIRANO, K., IMBENS, G. W., RUBIN, D. B. AND ZHOU, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.

HOLLAND, P. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–970.

IMBENS, G. W. AND ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–476.

IMBENS, G. W. AND RUBIN, D. B. (1997a). Estimating outcome distributions for compliers in instrumental variable. *Review of Economic Studies* **65**, 555–574.

IMBENS, G. W. AND RUBIN, D. B. (1997b). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25**, 305–327.

KALICHMAN, S. C., WILLIAMS, E. AND NACHIMSON, D. (2000). Randomized community trial of a brest self-examination skills-building intervention for inner-city African–American women. *Journal of the American Medical Women Association* **551**, 47–50.

LITTLE, R. AND RUBIN, D. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

LITTLE, R. AND YAU, L. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no–shows using Rubin's causal model. *Psychological Methods* **3**, 147–159.

MILLER, A. B. AND BAINES, C. (2001). Self examination contributes to reduction in mortality. *British Medical Journal Letter* **322**, 792.

MISHRA, S. I., CHAVEZ, L. R., MAGANA, J. R., NAVA, P., BURCIAGA VALDEZ, R. AND HUBBEL, F. A. (1998). Improving breast cancer control among Latinas: evaluation of a thery-based educational program. *Health Education Behaviour* **255**, 653–670.

Ortega-Altamirano, D., Lopez-Carrillo, L. and Lopez-Cervantes, M. (2000). Strategies for teaching self-examination of the breast to women in reproductive age. *Salud Publica Mexicana* **421**, 17–25.

Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association* **89**, 737–749.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (2000). Comment on 'Causal Inference without Conterfactuals' by P. Dawid. *Journal of the American Statistical Association* **95**, 407–448.

Scharfstein, D., Rotnitzky, A. and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Sheiner, L. B. and Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapy* **57**, 6–10.

Sommer, A. and Zeger, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45–52.

Spurgeon, D. (2001). Breast self-examination may do more harm than good. *British Medical Journal Letter* **323**, 11.

Strickland, C. J., Feigl, P., Upchurch, C., King, D. K., Pierce, H. I., Grevstad, P. K., Bearden, J. D., Dawson, M., Loewen, W. C. and Meyskens, Jr, F. L. (1997). Improving breast self-examination compliance: a Southwest Oncology Group randomized trial of three interventions. *Preventive Medicine* **263**, 320–332.