Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2005/09

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Application of the Principal Stratification Approach to the Faenza Randomized Experiment on Breast Self-Examination

Alessandra Mattei, Fabrizia Mealli

Università degli Studi
di Firenze

Application of the Principal Stratification Approach to the Faenza Randomized Experiment on Breast Self-Examination

Alessandra Mattei
Dipartimento di Statistica - Università di Firenze
mattei@ds.unifi.it

and

Fabrizia Mealli
Dipartimento di Statistica - Università di Firenze
mealli@ds.unifi.it

Abstract

Many scientific problems require that treatment comparisons be adjusted for posttreatment variables, such as treatment noncompliance, missing outcomes following treatment noncompliance, and "truncation by death". In the last years, there has been substantial progress in the analysis of randomized experiment suffering from noncompliance and missing outcome data, and recent work has addressed the problem of truncation by death; but, we are not aware of any previous work that addresses all these complications jointly. We present an extended framework for the analysis of data from randomized experiments which suffer from treatment noncompliance, missing outcomes following treatment noncompliance, and "truncation by death". There are two key feature of this framework: we use the principal stratification (Frangakis and Rubin, 2002) approach for comparing treatments adjusting for posttreatment variables, and we adopt a Bayesian approach for inference and sensitivity analysis. This framework is illustrated in the context of a randomized trial of Breast Self-examination.

KEYWORDS: Causal inference, Noncompliance, Missing data, Truncation by death, Pattern mixture models, Principal Stratification, Rubin causal model.

1 Introduction

Many randomized studies involving human subjects suffer from complications due to missing data and noncompliance with the randomly assigned treatment. As it is well known, in such a case, a standard analysis, which drops subjects with missing outcomes and ignores compliance information, can lead to biased results, even when the goal is to estimate simple intention-to-treat effects (Frangakis and Rubin, 1999). In addition, many such experiments require that treatment comparisons are also adjusted for "truncation by death". Traditional approaches addressing this issue ignore the fact that the outcome after the truncation is neither "censored" nor "missing", but should be treated as being defined on an extended sample space (Rubin, 2000; Frangakis and Rubin, 2002; Zhang and Rubin, 2003).

In this paper, we develop an extended framework for the analysis of randomized experiments which require that treatment comparisons are adjusted for (*i*) noncompliance with the randomly assigned treatment, (*ii*) missing outcomes (dropout) following treatment noncompliance, and (*iii*) "truncation by death". In our work, we focus on describing and addressing these complications using a Bayesian approach with the framework of principal stratification (Frangakis and Rubin, 2002).

We apply these methods to a reanalysis of a data set on Breast Self-Examination (BSE) previously studied by Ferro et al. (1996) and Mealli et al. (2004). In the study two methods of teaching BSE, consisting of either mail information about BSE (standard treatment) or the attendance of a course (new treatment) involving theoretical and practical sessions, 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. In addition, quality of self exams is "truncated by death" in the sense that there is no hidden value of it masked by the truncating event: the quality outcome can only be observed for women who practice BSE, and it is not only unobserved but also undefined on the usual sample space for those who do not practice BSE.

We present the framework we use in section 3 and section 4, and discuss our model's structural assumptions in section 5. Section 6 describes our parametric model specification, and main results of the analysis are presented in section 7. We discuss model building and checking in section 8 and conclude in section 9.

2 The Faenza Randomized Study on Breast Self-Examination and its Data Complications

Breast Self-Examination (BSE) remains the most controversial of commonly recommended procedures for breast screening. The rationale behind extending BSE as screening test stems from the fact that breast cancer is frequently detected by women themselves without any other symptoms.

In this paper, we reanalyze data of a randomized experiment on Breast Self-Examination conducted between January 1988 and December 1990 at the Oncologic Center of the Faenza Health District in Italy. In this study, two BSE teaching methods were compared, a standard treatment of receiving mailed information only, and a new treatment of additional attendance in a self-exam course, held by specialized medical staff, which consisted of a one hour session, a group discussion and a fifteen-minute individual practical session. Both treatment levels were selected on the basis of their practical feasibility and their acceptability according to the cultural profile of the area.

Details of the design are described in Ferro et al. (1996). In the final sample there were 657 women; they completed a self-administered pretest questionnaire aimed at evaluating their knowledge of breast pathophysiology, presence of known risk factors for breast cancer, preventive beliefs, level of knowledge, practice and examination skills of BSE, and other individual characteristics. Then, respondents to the pretest questionnaire were randomly assigned to either a new, enhanced teaching treatment (330) or to a standard treatment group (327).

Actually, of the 330 women randomly assigned to the enhanced treatment, only 182 com-

plied 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 received only the standard treatment of the mailed information. Therefore, attendance at the BSE teaching course was not perfectly correlated with the assignment.

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, proving information on posttreatment BSE practice and on quality of self exams. This is likely partly due to the fact that the outcome data were collected at a later date than the covariate and assignment data.

When the outcomes are not observed for all units, analyses based only on complete observations could lead to biased estimates of effects of treatment, because missingness of outcomes that occurs after randomization is not guaranteed to be balanced between the randomized arms. For example, we observe that response is substantially lower among women assigned to receive the active treatment (62%) than among the other women (69%). Moreover, standard adjustments for outcome missingness ignore its potential interaction with the other complications and generally make implicit and unrealistic assumptions.

In the Faenza study, the question of interest was the effect of an enhanced training class on BSE practice and quality of self-exam execution. The quality outcome was assessed using a compilation of different practice indicators, which resulted in a variable that could take on integer values between 0 and 21. As suggested in other works (Ferro et al., 1996; Mealli et al., 2004), in our analysis we consider a binary quality outcome variable equal to H ("High") if the individual's quality indicator is greater than the overall sample median (in this case 17) and L ("Low") otherwise. Clearly, such outcome was defined only for those women practicing BSE before and after the educational interventions, so we need to find a way of adjusting for BSE practice status pre- and post-treatment. The solution to such a problem is often to assume the quality outcome variable for women who do not practice BSE 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, because they ignore the fact that the quality outcome for women who do not practice BSE is neither "censored" nor "missing"; simply it is not defined on the usual sample space, here, $\{L, H\}$. Following Zhang and Rubin (2003), it can be defined as $*$ on the extended space $\{L, H, *\}$ in order to properly "account for" BSE practice when addressing causal effects of the teaching program on BSE quality.

3 Principal Stratification and its Role for Causal Inference

In order to address the data complications in the Faenza study, first we introduce "potential outcomes" (see Rubin, 1979; Holland, 1986) for all the posttreatment variables. Potential outcomes for any given variable comprise the observable manifestation of this variable under each of the possible treatment assignments. In particular, if woman i in the study ($i = 1, \ldots, N$) is to be assigned to treatment z ($z = T$ for new treatment and $z = C$ for control), we denote the following: $D_i(z)$ for the indicator equal to P if the woman actually attends the training program, and p if she receives only mailed information on BSE; $S_i(z)$ for the BSE practice indicator equal to B if the woman practices BSE and b otherwise; and $Y_i(z)$ for the potential quality outcome, where $Y_i(z) = H$ if the woman practices BSE with "high" quality, that is, whether her quality indicator of posttreatment BSE practice exceeds the designed

threshold, and $Y_i(z) = L$ if the woman practices BSE with "low" quality, that is, whether her BSE quality is lower than the fixed threshold. Recall that, quality of BSE practice can only be observed for women who practice BSE ($S_i(z) = B$), and it is not only unobserved but also undefined on the usual sample space for women who do not practice the self exams. Formally, we assume that the quality outcome is defined on the extended space $\{L, H, *\}$, with $Y_i(z) \in \{L, H\}$ if $S_i(z) = B$, and $Y_i(z) = *$ if $S_i(z) = b$. Lastly, we denote with $R_i(z)$ the indicator equal to 1 if the woman i responds to the posttest questionnaire, and 0 otherwise.

The outcomes D, S, Y, and R are called potential outcomes because only one version of them can ever be observed, the version under the assigned treatment; the other versions, under the unassigned treatments cannot be observed. Each participant is randomly assigned to one treatment arm, therefore, if we indicate with Z_i^{obs} the observed treatment assignment, the observed data are

$$\left(Z_i^{\text{obs}}, D(Z_i^{\text{obs}}), R(Z_i^{\text{obs}}), S(Z_i^{\text{obs}}), Y(Z_i^{\text{obs}})\right) \qquad i = 1, \ldots, N,$$

which we will denote by $(Z_i^{\text{obs}}, D_i^{\text{obs}}, R_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}})$, $i = 1, \ldots, N$. In addition, three covariates are observed: X_{i1}^{obs}, a binary indicator of previous BSE practice, X_{i2}^{obs} a binary indicator of good knowledge of breast pathophysiology, and X_{i3}^{obs} age in years. Corresponding to each set of these individual-specific random variables is a boldface variable (vector or matrix) without subscript i, that refers to the set of these variables across all study participants. In particular, let $\mathbf{Z}^{\text{obs}} = \{Z_i^{\text{obs}}, i = 1, \ldots, N\}$, $\mathbf{D}^{\text{obs}} = \{D_i^{\text{obs}}, i = 1, \ldots, N\}$, $\mathbf{R}^{\text{obs}} = \{R_i^{\text{obs}}, i = 1, \ldots, N\}$, $\mathbf{S}^{\text{obs}} = \{S_i^{\text{obs}}, i = 1, \ldots, N\}$, and $\mathbf{Y}^{\text{obs}} = \{Y_i^{\text{obs}}, i = 1, \ldots, N\}$. Lastly, let $\mathbf{X}^{\text{obs}}$ be the $N \times 3$ matrix with ith row equal to $\mathbf{X}_i^{\text{obs}} = (X_{i1}^{\text{obs}}, X_{i2}^{\text{obs}}, X_{i3}^{\text{obs}})$.

Consider the potential outcomes $D_i(T)$, $S_i(C)$, and $S_i(T)$. The Faenza study 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. The potential outcome $D_i(T)$, that is, the treatment woman i would received if assigned to the active treatment, defines the compliance behavior of each subject. If $D_i(T) = P$, then woman i is a "complier"; among these individuals $D(Z^{\text{obs}} = T) = P$ (as observed), and by the structure of the experimental setting, had they instead been assigned to standard treatment, $D(Z^{\text{obs}} = C) = p$, by definition. Thus for these units $\mathrm{I}\{D_i^{\text{obs}} = P\} = \mathrm{I}\{Z_i^{\text{obs}} = T\}$, where $\mathrm{I}\{\cdot\}$ is the indicator function: they always comply with their treatment assignment. In contrast, if $D_i(T) = p$ 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 $D_i(z) = p$, for both $z = C$ and T. Each of the two strata of people - compliers and never-takers - defined by the compliance status can be further classified into four groups according to the joint potential values of the BSE practice variable under each of the treatments being compared: $(S_i(C), S_i(T))$. Thus, within each cell defined by a specific value of the pretreatment variables, the participants in the trial can be stratified into eight groups according to the joint value of the potential outcomes $(D_i(T), S_i(C), S_i(T))$:

PBB $= \{i : D_i(T) = P, S_i(C) = B, S_i(T) = B\}$: compliers who would practice BSE under both treatment arms, which comprise a proportion $\pi(\text{PBB})$ of all women;

PbB $= \{i : D_i(T) = P, S_i(C) = b, S_i(T) = B\}$: compliers who would not practice BSE under control but would practice BSE under treatment, which comprise a proportion $\pi(\text{PbB})$ of all women;

Table 1: Principal Stratification and associated pattern for potential outcomes.

Principal Stratum	$D_i(T)$	$S_i(C)$	$S_i(T)$	$Y_i(C)$	$Y_i(T)$
PBB	P	B	B	$\in \{L, H\}$	$\in \{L, H\}$
PbB	P	b	B	$*$	$\in \{L, H\}$
PBb	P	B	b	$\in \{L, H\}$	$*$
Pbb	P	b	b	$*$	$*$
pBB	p	B	B	$\in \{L, H\}$	$\in \{L, H\}$
pbB	p	b	B	$*$	$\in \{L, H\}$
pBb	p	B	b	$\in \{L, H\}$	$*$
pbb	p	b	b	$*$	$*$

PBb $= \{i : D_i(T) = P, S_i(C) = B, S_i(T) = b\}$: compliers who would practice BSE under control but would not practice BSE under treatment, which comprise a proportion π(PBb) of all women;

Pbb $= \{i : D_i(T) = P, S_i(C) = b, S_i(T) = b\}$: compliers who would practice BSE under neither treatment arms, which comprise a proportion π(Pbb) of all women;

pBB $= \{i : D_i(T) = p, S_i(C) = B, S_i(T) = B\}$: never-takers who would practice BSE under both treatment arms, which comprise a proportion π(pBB) of all women;

pbB $= \{i : D_i(T) = p, S_i(C) = b, S_i(T) = B\}$: never-takers who would not practice BSE under control but would practice BSE under treatment, which comprise a proportion π(pbB) of all women;

pBb $= \{i : D_i(T) = p, S_i(C) = B, S_i(T) = b\}$: never-takers who would practice BSE under control but would not practice BSE under treatment, which comprise a proportion π(pBb) of all women;

pbb $= \{i : D_i(T) = p, S_i(C) = b, S_i(T) = b\}$: never-takers who would practice BSE under neither treatment arms, which comprise a proportion π(pbb) of all women.

This partition of the women is a direct application of the idea of principal stratification (Frangakis and Rubin, 2002) using the framework of Rubin's Causal Model. For now, we suppose being already within cells defined by pretreatment covariates. Then, the pattern for potential outcomes associated with each basic principal stratum is shown in Table 1.

In each principal stratum, there are respondents and non-respondents. Specifically, in each of above strata, there will be women who respond under either assignment, who respond only if assigned to control but not if assigned to treatment, who respond if assigned to treatment but not if assigned to control, and who do not respond regardless of assignment. To address this problem, we propose a new missing data model, which allows us to properly take into account the potential interactions between missingness and the other data complications. Our missing data model bases on two key assumptions: the response exclusion restriction for compliers on the effect of assignment, and the ignorability of the missing data mechanism

with respect to the quality outcome Y within each principal strata defined by the vector $(D_i(T), S_i(C), S_i(T))$.

Let G_i represent the principal stratum indicator for subject i. The N-dimensional vector of G_i's will be denoted by $\mathbf{G}$. The principal stratum indicator G_i is not affected by treatment assignment Z_i^{obs}, so it only reflects characteristics of subject i, and can be regarded as a covariate, which is only partially observed in the sample (Angrist et al., 1996); by randomization, however, it is guaranteed to have the same distribution in both treatment arms.

It is common practice to adjust for important pretreatment variables in doing causal inference, and thus, we need to adjust for the potentially important covariate - the principal strata. As noted above, generally, we cannot observe the principal stratum to which a woman belongs; however, if memberships G_i were known, stratification of the subjects by G_i would adjust for the personal characteristics reflected in the post-treatment variable without introducing treatment selection bias. In other words, principal strata might represent potentially important unobservable characteristics of the subjects. If adjustment for principal strata is not made, since we assume that we are already within cells defined by pretreatment variables, an implicit assumption is that, given these pretreatment variables, the principal strata do not give additional information about the characteristics of the participants. Such assumption is often invalid, since we surely have reason to believe that each latent group has different awareness of the risk of breast cancer, and probably lands on different points on the scale of potential ability, even if it is observed to have pretreatment variables similar to the other groups.

The idea underlying principal stratification is to propose a framework for adjusting for posttreatment variables, which always generates causal effects because it always compares potential outcomes for a common set of people. Principal strata, G_i, have two important properties. First, they are not affected by assignment. Second, comparisons of potential outcomes under different assignment within principal strata, called principal effects, are well defined causal effects (Frangakis and Rubin, 2002). These properties make principal stratification a powerful framework for evaluation because it allows us (a) to define explicitly estimands that better represent the effect of treatment, and (b) to explore richer and explicit sets of assumptions that allow estimation of these effects under more plausible conditions than standard ones.

4 Observed Groups

Unfortunately, we cannot directly observe the principal strata for the participants. In our experimental setting, for women assigned to the active treatment we can observe the compliance status, though we cannot observe the value of the BSE practice indicator under the unassigned standard treatment. Specifically, in the treatment arm, we can observe the following groups:

$\text{OBS}(T, P, 1, B) = \left\{i : Z_i^{\text{obs}} = T, D_i^{\text{obs}} = P, R_i^{\text{obs}} = 1, S_i^{\text{obs}} = B\right\}$: compliers who are assigned to treatment, respond, and practice BSE;

$\text{OBS}(T, P, 1, b) = \left\{i : Z_i^{\text{obs}} = T, D_i^{\text{obs}} = P, R_i^{\text{obs}} = 1, S_i^{\text{obs}} = b\right\}$: compliers who are assigned to treatment, respond, but do not practice BSE;

$\text{OBS}(T, p, 1, B) = \left\{i : Z_i^{\text{obs}} = T, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 1, S_i^{\text{obs}} = B\right\}$: never-takers who are assigned to treatment, respond, and practice BSE;

$\text{OBS}(T, p, 1, b) = \left\{i : Z_i^{\text{obs}} = T, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 1, S_i^{\text{obs}} = b\right\}$: never-takers who are assigned to treatment, respond, but do not practice BSE;

$\text{OBS}(T, P, 0, ?) = \left\{i : Z_i^{\text{obs}} = T, D_i^{\text{obs}} = P, R_i^{\text{obs}} = 0, S_i^{\text{obs}} = ?\right\}$: compliers who are assigned to treatment and do not respond;

$\text{OBS}(T, p, 0, ?) = \left\{i : Z_i^{\text{obs}} = T, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 0, S_i^{\text{obs}} = ?\right\}$: never-takers who are assigned to treatment and do not respond.

In the control arm, not even the compliance status is directly observed; thus each observed group of women assigned the standard treatment, in general, comprises a mixture of compliers ($D_i(T) = P$) and never-takers ($D_i(T) = p$). Specifically, what we can observe in the control arm are the following three groups:

$\text{OBS}(C, p, 1, B) = \left\{i : Z_i^{\text{obs}} = C, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 1, S_i^{\text{obs}} = B\right\}$: women (mixture of compliers and never-takers) who are assigned to control, respond, and practice BSE;

$\text{OBS}(C, p, 1, b) = \left\{i : Z_i^{\text{obs}} = C, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 1, S_i^{\text{obs}} = B\right\}$: women (mixture of compliers and never-takers) who are assigned to control, respond, but do not practice BSE;

$\text{OBS}(C, p, 0, ?) = \left\{i : Z_i^{\text{obs}} = C, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 0, S_i^{\text{obs}} = ?\right\}$: women (mixture of compliers and never-takers) who are assigned to control, and do not respond.

Each woman is observed to fall into one of these groups, but also belongs to a latent (unobserved) principal stratum, G_i. If all eight principal strata exist, that is, if $\pi(g) > 0$, for each $g \in \{\text{PBB}, \text{PbB}, \text{PBb}, \text{Pbb}, \text{pBB}, \text{pbB}, \text{pBb}, \text{pbb}\}$, each observed group $\text{OBS}(Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}})$ would be a mixture of two or more principal strata. For example, $\text{OBS}(T, P, 1, B)$ would be a mixture of the PBB group and the PbB group, and $\text{OBS}(C, p, 1, B)$ would be a mixture of four principal strata: PBB, PBb, pBB, and pBb. The data pattern and the latent principal strata associated with each observed group are shown in Table 2.

Table 2 gives us an idea on why a standard Intention-To-Treat (ITT) analysis, which ignores compliance information and drops subjects with missing or truncated outcomes, is misleading for inferences about the causal effect on BSE quality. The reason is that the group of women who are observed practising BSE under treatment and the group of women who are observed practising BSE under control involve different combinations of basic principal strata, and thus are different groups of people. In order to learn about the causal effect of the treatment on BSE quality, we should focus on women belonging to the PBB principal stratum. In our study, which suffered from noncompliance, this approach leads to estimate the Complier Average Causal Effect (CACE) on BSE quality. We could be also interested in measuring the effect of the encouragement, that is, the effect of the randomized encouragement on the quality of self exam for all the women who practice BSE under both assignments (women who belong to the PBB group or to the pBB group).

Concerning the BSE practice outcome, we focus on three estimands: (1) the Intention-To-Treat (ITT) effect, that is, the effect of the randomized encouragement on all subjects; (2) the Complier Average Causal Effect (CACE), that is, the effects of the randomized encouragement on all subjects who would comply with their treatment assignment no matter which assignment they would be given (here, women who would have attended the enhanced BSE teaching program if they had been invited, and would not have had they not invited); and (3) the Never-taker Average Causal Effect (NACE), that is, the effect of the randomized encouragement on all subjects who never take the treatment no matter the assignment

Table 2: Group classification based on observed data OBS($Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}}$) and associated data pattern and possible latent principal strata (?=data missing).

Observed Group OBS($Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}}$)	Z_i^{obs}	D_i^{obs}	R_i^{obs}	S_i^{obs}	Latent Group G_i			
OBS($T, P, 1, B$)	T	P	1	B		PBB	PbB	
OBS($T, P, 1, b$)	T	P	1	b		PBb	Pbb	
OBS($T, P, 0, ?$)	T	P	0	?	PBB	PbB	PBb	Pbb
OBS($T, p, 1, B$)	T	p	1	B		pBB	pbB	
OBS($T, p, 1, b$)	T	p	1	b		pBb	pbb	
OBS($T, p, 0, ?$)	T	p	0	?	pBB	pbB	pBb	pbb
OBS($C, p, 1, B$)	C	p	1	B	PBB	PBb	pBB	pBb
OBS($C, p, 1, b$)	C	p	1	b	PbB	Pbb	pbB	pbb
OBS($C, p, 0, ?$)	C	p	0	?	PBB pBB	PbB pbB	PBb pBb	Pbb pbb

(here, women who would not have attended the training course if they had been invited to partecipate in it). As we will see, these estimands depend on the proportions $\pi(g)$, $g \in \{\text{PBB}, \text{PbB}, \text{PBb}, \text{Pbb}, \text{pBB}, \text{pbB}, \text{pBb}, \text{pbb}\}$. All these quantities will be defined more formally in section 7.

5 Structural Assumptions

First we state explicitly our assumptions about the data with regard to causal processes, the missing data mechanism, the compliance structure, and the BSE practice behavior. These assumptions are expressed without reference to a particular parametric distribution.

5.1 SUTVA

A standard assumption made in causal analysis is the Stable Unit Treatment Value Assumption (SUTVA), formalized with potential outcomes by Rubin (1978, 1980, 1990). SUTVA combines the no-interference assumption (Cox, 1958) that one unit's treatment assignment does not affect another unit's outcomes with the assumption that there are "no versions of treatments". For no-interference to hold, whether or not one woman was invited to attend the "hand-on" training course on BSE techniques should not affect another woman's outcomes such as her compliance behavior, her choice to practice BSE or her quality of self-exam execution. We expect our results to be robust to the types and degree of deviations from no interference that might be anticipated in this study. To satisfy the "no versions of treatments", we need to limit the definition of BSE training programs to those performed in our experiment. Generalizability of results to other methods for teaching breast self-exam

techniques would have to be judged separately.

5.2 Ignorability of Treatment Assignment

The study design of the Faenza randomized trial implies

Assumption 1. *(*Ignorability of Treatment Assignment*)*

$$\Pr\big(Z_i \mid D_i(T), S_i(C), S_i(T), R_i(C), R_i(T), Y_i(C), Y_i(T), \boldsymbol{X}_i^{\text{obs}}, \theta\big) = \Pr\big(Z_i \mid \boldsymbol{X}_i^{\text{obs}}, \theta\big) = \Pr\big(Z_i \mid \boldsymbol{X}_i^{\text{obs}}\big),$$

where θ is generic notation for the parameters governing the distribution of all the variables. There is no dependence on θ because there are no unknown parameters controlling the treatment assignment mechanism. Participants in the study were randomly assigned to either the new teaching treatment or to the standard treatment group, and the randomization probabilities within cells defined by pretreatment variables are known.

5.3 Monotonicity Assumptions

We impose two monotonicity assumptions which rule out the existence of two principal strata.

We distinguish two components of our monotonicity assumption: one for compliers and one for never-takers. In the first component we assume that there is no PBb group, namely, no woman who complies with her assignment, would practice BSE under control but would not practice BSE under treatment. This monotonicity assumption can be formally expressed as

Assumption 2. *(*Monotonicity for Compliers*)*

For all compliers ($D_i(T) = P$),

$$\mathrm{I}\{S_i(T) = B\} \geq \mathrm{I}\{S_i(C) = B\}.$$

In the second component of the monotonicity assumption we assume that there is no pbB group, namely, no woman who never complies with her assignment, would not practice BSE under control, but would practice BSE under treatment. Formally,

Assumption 3. *(*Monotonicity for Never-Takers*)*

For all never-takers ($D_i(T) = p$),

$$\mathrm{I}\{S_i(T) = B\} \leq \mathrm{I}\{S_i(C) = B\}.$$

These monotonicity assumptions formalize the notion that being invited to partecipate to the new BSE teaching program could improve BSE practice rate for compliers, but worsen it for never-takers. This idea arises from preliminary analyses, which suggest that in our study the intention-to-treat effects on BSE practice for compliers and never-takers could be both nonzero, and have opposite sign, with the sign for compliers being positive (see Mealli et al., 2004, and Mattei and Mealli, 2004).

Table 3: Group classification based on observed data OBS($Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}}$) and associated data pattern and possible latent principal strata under the monotonicity assumptions 2 and 3 (? = data missing).

Observed Group					Latent Group		
OBS($Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}}$)	Z_i^{obs}	D_i^{obs}	R_i^{obs}	S_i^{obs}	G_i		
OBS($T, P, 1, B$)	T	P	1	B	PBB	PbB	
OBS($T, P, 1, b$)	T	P	1	b		Pbb	
OBS($T, P, 0, ?$)	T	P	0	?	PBB	PbB	Pbb
OBS($T, p, 1, B$)	T	p	1	B		pBB	
OBS($T, p, 1, b$)	T	p	1	b	pBb	pbb	
OBS($T, p, 0, ?$)	T	p	0	?	pBB	pBb	pbb
OBS($C, p, 1, B$)	C	p	1	B	PBB	pBB	pBb
OBS($C, p, 1, b$)	C	p	1	b	PbB	Pbb	pbb
OBS($C, p, 0, ?$)	C	p	0	?	PBB pBB	PbB pBb	Pbb pbb

In principle, both the PBb groups and the pbB groups could exist, therefore assessment of the monotonicity assumptions is crucial for any sensible inference based on them. We will discuss this issue further in section 8.

As we can see in Table 3, under the two monotonicity assumptions 2 and 3, the data pattern and the latent group associated with each observed group gets easier: some principal strata, such as the Pbb group and the pBB group, can be directly observed. However, it should be noted that without any additional assumption, this simplification does not help to identify the proportions of women belonging to each principal stratum. In addition, despite the two monotonicity assumptions, most of observed groups are still a mixture of two or more principal strata, and so their observed distributions are a mixture of two or more distributions.

5.4 Exclusion Restrictions

In order to address complications because the principal strata are not directly observed, we impose two additional assumptions: two exclusion restrictions on the effect of assignment.

The former assumes that within subpopulations of never-takers who would practice BSE under both assignments, and with the same value of the pretreatment covariates, the distri-

butions of the two potential quality outcomes $Y_i(C)$ and $Y_i(T)$ are the same:

Assumption 4. *(*Stochastic Quality Outcome Exclusion Restriction for the pBB Group*)*

$$\begin{aligned}&\Pr\big(Y_i(T) = H \mid D_i(T) = p, S_i(C) = B, S_i(T) = B, \boldsymbol{X}_i\big)\\ &\quad = \Pr\big(Y_i(C) = H \mid D_i(T) = p, S_i(C) = B, S_i(T) = B, \boldsymbol{X}_i\big).\end{aligned}$$

In the pBB group there are women who never comply with their assignment and practice BSE under both treatment and control. Since for this type of units the treatment actually received and the BSE behavior would be the same no matter what their assignment, the intervention of assignment within this study is arguably of little relevance to this group. Consequently, assumption 4 asserts that for never-takers who practice BSE under both treatment and control there is no effect of assignment on their potential quality outcome $Y_i(z)$.

Since in our study complications due to noncompliance and missing outcomes are both present, compliance behavior and response behavior have to be jointly taken in account and modeled in some principled way. To address these issues, we introduce a new missing data model that is specially suited in the context of randomized designs where treatment comparisons should be adjusted for noncompliance and truncation by death. Our missing data model is based on two assumptions; one of these is a particular type of exclusion restriction which assumes that within subpopulatons of compliers with the same value of the pretreatment variables and the same vector $(S_i(C), S_i(T))$, the distributions of the two potential response indicators $R_i(C)$ and $R_i(T)$ are the same:

Assumption 5. *(*Stochastic Response Exclusion Restriction for Compliers*)*

$$\Pr\big(R_i(T) = 1 \mid D_i(T) = P, S_i(C), S_i(T), \boldsymbol{X}_i\big) = \Pr\big(R_i(C) = 1 \mid D_i(T) = P, S_i(C), S_i(T), \boldsymbol{X}_i\big)$$

This assumption implies that compliers ($D_i(T) = P$) have the same response behavior irrespective of the treatment arm they are assigned to, given the partially observed covariate $(S_i(C), S_i(T))$ and the pretreatment variables. As compliers are willing to follow the protocol in their assigned treatment, it seems plausible that they would not be affected in their response behavior by that assignment.

5.5 Latent Ignorability

The other key assumption of our missing data model imposes that potential outcomes on quality are independent of missingness given pretreatment variables conditional on the principal strata defined by the covariates $D_i(T)$ and $(S_i(C), S_i(T))$; formally

Assumption 6. *(*Latent Ignorability*) Potential quality outcomes and potential nonresponse indicators are independent within principal strata:*

(a) when assigned standard treatment

$$\begin{aligned}&\Pr\big(Y_i(C) \mid D_i(T), S_i(C) = B, S_i(T), R_i(C), \boldsymbol{X}_i\big)\\ &\quad = \Pr\big(Y_i(C) \mid D_i(T), S_i(C) = B, S_i(T), \boldsymbol{X}_i\big);\end{aligned}$$

(b) when assigned new treatment

$$\begin{aligned}&\Pr\big(Y_i(T) \mid D_i(T), S_i(C), S_i(T) = B, R_i(T), \boldsymbol{X}_i\big)\\ &\quad = \Pr\big(Y_i(T) \mid D_i(T), S_i(C), S_i(T) = B, \boldsymbol{X}_i\big).\end{aligned}$$

This assumption represents a form of Latent Ignorability (LI) (Frangakis and Rubin, 1999) in that it conditions on variables that are (at least partially) unobserved or latent, here, principal strata defined by the vector $(D_i(T), S_i(C), S_i(T))$. This assumption requires that potential BSE quality indicators and associated potential nonresponse indicators are independent within the PBB and pBB principal strata of the same pretreatment assignment levels. In addition, LI imposes that Y and R are independent within the pBb groups when assigned standard treatment, and within the PbB groups when assigned new treatment. Note that, because principal strata are not directly observed, the missing data process is in fact non-ignorable.

The intuition behind the LI assumption in our study is that, for a subgroup of people with the same values of covariates, and the same principal stratum, a flip of a coin could determine which of these individuals shows up for the posttest questionnaire. This is a more reasonable assumption than standard ignorability (Rubin, 1978; Little and Rubin, 1987) because it seems quite likely that each principal stratum would exhibit a different attendance behavior for posttest questionnaire, even conditional on the other background variables.

5.6 Additional Assumptions

Since never-takers would not attend the enhanced BSE teaching course no matter what their assignment, it seems that these women do not regard the risk of breast cancer as high enough. Therefore, it might reasonable to assume that the willingness to respond of these women is not related to their BSE practice behavior. Then, assumption 7 comes into play

Assumption 7.

$$\Pr\big(R_i(Z_i) = 1 \mid D_i(T) = p, S_i(C), S_i(T), \boldsymbol{X}_i\big) = \Pr\big(R_i(Z_i) = 1 \mid D_i(T) = p, \boldsymbol{X}_i\big).$$

Finally, we assume a stochastic dominance assumption which ranks the self-examination skills among compliers. This assumption is a version of the ranked ability assumption proposed by Zhang (2002) and Zhang and Rubin (2004) in the context of a randomized experiment with perfect compliance where the primary outcome is truncated by death.

In our stochastic dominance assumption we assume that when assigned to treatment, the proportion of women who practice BSE with high quality in the PBB principal stratum is no less that in the PbB principal stratum:

Assumption 8. *(*Stochastic Dominance of the PBB Group over the PbB Group*)*

$$\begin{aligned}&\Pr\big(Y_i(T)=H \mid D_i(T)=P, S_i(C)=B, S_i(T)=B, \boldsymbol{X}_i\big)\\ &\quad\geq \Pr\big(Y_i(T)=H \mid D_i(T)=P, S_i(C)=b, S_i(T)=B, \boldsymbol{X}_i\big).\end{aligned}$$

This assumption formalizes the notion that the PBB group is more capable or has higher motivation than the PbB group; the PBB group would practice BSE under either treatment arm, whereas the PbB group would practice the self exam under only the treatment arm. This seems to be plausible, because ability or motivation might tend to be positively correlated with BSE quality for women who comply with their assignment.[1]

6 Parametric Pattern Mixture Model

Generally speaking, constrained estimation of separate analyses within missing data patterns is the motivation behind pattern mixture modeling. Typically pattern mixture models partition the data with respect to the missingness of the variables. Here, we use a partial pattern mixture model approach, that is, we partition the data with respect to principal strata defined by the latent vector $(D(T), S(C), S(T))$, as well as the pretreatment variables X.

To capitalize on the structural assumptions, consider the factorization of the joint distribution for the potential outcomes and principal strata conditional on the covariates:

$$\begin{aligned}&\Pr\big(D_i(T), S_i(C), S_i(T), R_i(C), R_i(T), Y_i(C), Y_i(T) \mid \mathbf{X}_i^{\text{obs}}; \theta\big)\\ &\quad= \Pr\big(D_i(T), S_i(C), S_i(T) \mid \mathbf{X}_i^{\text{obs}}; \theta^G\big)\\ &\quad\times \Pr\big(R_i(C), R_i(T) \mid D_i(T), S_i(C), S_i(T), \mathbf{X}_i^{\text{obs}}; \theta^R\big)\\ &\quad\times \Pr\big(Y_i(C), Y_i(T) \mid D_i(T), S_i(C), S_i(T), \mathbf{X}_i^{\text{obs}}; \theta^Y\big),\end{aligned}$$

where the last product on the right follows by latent ignorability, and $\theta = (\theta^G, \theta^R, \theta^Y)'$. In our experimental setting, it is also useful to factorize the conditional principal stratum distribution as

$$\begin{aligned}&\Pr\big(D_i(T), S_i(C), S_i(T) \mid \mathbf{X}_i^{\text{obs}}; \theta^G\big)\\ &\quad= \Pr\big(D_i(T) \mid \mathbf{X}_i^{\text{obs}}; \theta^{D(T)}\big) \times \Pr\big(S_i(C), S_i(T) \mid D_i(T), \mathbf{X}_i^{\text{obs}}; \theta^S\big),\end{aligned}$$

where $\theta^G = (\theta^{D(T)}, \theta^S)'$; $\Pr(D_i(T) \mid \mathbf{X}_i^{\text{obs}}, \theta^{D(T)})$ is the compliance principal stratum distribution conditional on the covariates; and $\Pr(S_i(C), S_i(T) \mid D_i(T), \mathbf{X}_i^{\text{obs}}, \theta^S)$ is the joint distribution of the intermediate potential outcomes $S_i(C)$ and $S_i(T)$ given the compliance status $D_i(T)$ and the pretreatment variables $\mathbf{X}_i^{\text{obs}}$.

In the Faenza study the compliance covariate is dichotomous, therefore we assume that its distribution has a logistic regression form:

$$\pi_i^{D(T)} = \Pr\big(D_i(T) = P \mid \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \alpha\big) = \frac{\exp(\alpha_0 + \alpha_1' \mathbf{x}_i)}{1 + \exp(\alpha_0 + \alpha_1' \mathbf{x}_i)}.$$

[1] We did not impose this assumption a-priori, but we regarded it as possibly controversial, and we investigated its consequences in some detail. Because we found similar results from models with and without assumption 8, we decided to impose it in the final model (details on this sensitivity analysis are omitted).

In the general model, which does not impose the monotonicity assumptions 2 and 3, compliers and never-takers can be respectively classified into four groups according to the combination of the potential BSE practice indicators: BB - those who would practice BSE under both treatment arms; bB - those who would not practice BSE under control, but practice under treatment; Bb - those who would practice BSE under control but not under treatment; and bb - those who would practice BSE under neither treatment arm. The monotonicity assumptions 2 and 3 eliminate two groups: the Bb group among compliers and the bB group among never-takers. Then, given the compliance status $D(T)$, we can model the probabilities of belonging to one of the remaining three groups defined by the vector $(S(C), S(T))$ using a multinomial logit. Specifically, we assume that the joint distribution of the potential outcomes $S(C)$ and $S(T)$ conditional on $D(T) = P$, given the pretreatment variables $\mathbf{X}^{\text{obs}}$, has the following form:

$$\begin{aligned}
\pi_i^{\text{BB}}(P) &= \Pr\big(S_i(C) = B, S_i(T) = B \mid D_i(T) = P, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \gamma_P\big), \\
\pi_i^{\text{bB}}(P) &= \Pr\big(S_i(C) = b, S_i(T) = B \mid D_i(T) = P, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \gamma_P\big), \\
\pi_i^{\text{bb}}(P) &= \Pr\big(S_i(C) = b, S_i(T) = b \mid D_i(T) = P, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \gamma_P\big),
\end{aligned}$$

where $\gamma_P = (\gamma_P^{\text{BB}}, \gamma_P^{\text{bB}}, \gamma_P^{\text{bb}})$, and for $s_C s_T \in \{\text{BB}, \text{bB}, \text{bb}\}$ we have

$$\pi_i^{s_C s_T}(P) = \frac{\exp(\gamma_{0P}^{s_C s_T} + \gamma_{1P}^{s_C s_T'} \mathbf{x}_i)}{\exp(\gamma_{0P}^{BB} + \gamma_{1P}^{BB'} \mathbf{x}_i) + \exp(\gamma_{0P}^{bB} + \gamma_{1P}^{bB'} \mathbf{x}_i) + \exp(\gamma_{0P}^{bb} + \gamma_{1P}^{bb'} \mathbf{x}_i)}.$$

We normalize these probabilities by setting γ_P^{BB} equal to a vector of zeros. Similarly, for the conditional distribution $\Pr(S_i(C), S_i(T) \mid D_i(T) = p, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \theta^S)$ we use the following multinomial logit model:

$$\begin{aligned}
\pi_i^{\text{BB}}(p) &= \Pr\big(S_i(C) = B, S_i(T) = B \mid D_i(T) = p, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \gamma_p\big), \\
\pi_i^{\text{Bb}}(p) &= \Pr\big(S_i(C) = B, S_i(T) = b \mid D_i(T) = p, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \gamma_p\big), \\
\pi_i^{\text{bb}}(p) &= \Pr\big(S_i(C) = b, S_i(T) = b \mid D_i(T) = p, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \gamma_p\big),
\end{aligned}$$

where $\gamma_p = (\gamma_p^{\text{BB}}, \gamma_p^{\text{Bb}}, \gamma_p^{\text{bb}})$, and for $s_C s_T \in \{\text{BB}, \text{Bb}, \text{bb}\}$ we have

$$\pi_i^{s_C s_T}(p) = \frac{\exp(\gamma_{0p}^{s_C s_T} + \gamma_{1p}^{s_C s_T'} \mathbf{x}_i)}{\exp(\gamma_{0p}^{BB} + \gamma_{1p}^{BB'} \mathbf{x}_i) + \exp(\gamma_{0p}^{Bb} + \gamma_{1p}^{Bb'} \mathbf{x}_i) + \exp(\gamma_{0p}^{bb} + \gamma_{1p}^{bb'} \mathbf{x}_i)}.$$

As before, we take the BB group as the baseline group by setting $\gamma_p^{\text{BB}} = \mathbf{0}$.

The potential BSE quality indicators, Y, (when they exist) are dichotomous, therefore we assume that their distributions take the form of logistic regressions. Recall that quality of BSE practice is defined only for women who practice the self exams, namely, under the two monotonicity assumptions 2 and 3, for the PBB and the pBB principal strata no matter the assignment, for the PbB group when assigned to treatment, and for the pBb group when assigned to control. Thus, we have six quality distributions:

$$\begin{aligned}
\pi_{iz}^{Y}(\text{PBB}) &= \Pr\big(Y_i = H \mid Z_i = z_i, G_i = \text{PBB}, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \delta_z(\text{PBB})\big) \\
&= \frac{\exp(\delta_{0z}(\text{PBB}) + \delta_{1z}(\text{PBB})' \mathbf{x}_i)}{1 + \exp(\delta_{0z}(\text{PBB}) + \delta_{1z}(\text{PBB})' \mathbf{x}_i)}, \qquad z_i = C, T;
\end{aligned}$$

$$\begin{aligned}\pi^Y_{iz}(\text{pBB}) &= \Pr\big(Y_i = H \mid Z_i = z_i, G_i = \text{pBB}, \mathbf{X}^{\text{obs}}_i = \mathbf{x}_i; \delta_z(\text{pBB})\big) \\ &= \frac{\exp(\delta_{0z}(\text{pBB}) + \delta_{1z}(\text{pBB})'\mathbf{x}_i)}{1+\exp(\delta_{0z}(\text{pBB}) + \delta_{1z}(\text{pBB})'\mathbf{x}_i)}, \qquad z_i = C, T;\end{aligned}$$

$$\begin{aligned}\pi^Y_{iT}(\text{PbB}) &= \Pr\big(Y_i = H \mid Z_i = T, G_i = \text{PbB}, \mathbf{X}^{\text{obs}}_i = \mathbf{x}_i; \delta_T(\text{PbB})\big) \\ &= \frac{\exp(\delta_{0T}(\text{PbB}) + \delta_{1T}(\text{PbB})'\mathbf{x}_i)}{1+\exp(\delta_{0T}(\text{PbB}) + \delta_{1T}(\text{PbB})'\mathbf{x}_i)};\end{aligned}$$

and

$$\begin{aligned}\pi^Y_{iC}(\text{pBb}) &= \Pr\big(Y_i = H \mid Z_i = C, G_i = \text{pBb}, \mathbf{X}^{\text{obs}}_i = \mathbf{x}_i; \delta_C(\text{pBb})\big) \\ &= \frac{\exp(\delta_{0C}(\text{pBb}) + \delta_{1C}(\text{pBb})'\mathbf{x}_i)}{1+\exp(\delta_{0C}(\text{pBb}) + \delta_{1C}(\text{pBb})'\mathbf{x}_i)}.\end{aligned}$$

Here, $Y_i(C)$ and $Y_i(T)$, when both of them exist, are assumed conditionally independent, an assumption which has no effect on inference for our super-population parameters of interest (Rubin, 1978).

Finally, we also use a logit model for the potential response indicators R:

$$\pi^R_{iz}(g) = \Pr\big(R_i = 1 \mid Z_i = z_i, G_i = g, \mathbf{X}^{\text{obs}}_i = \mathbf{x}_i; \beta_z(g)\big) = \frac{\exp(\beta_{0z}(g) + \beta_{1z}(g)'\mathbf{x}_i)}{1+\exp(\beta_{0z}(g) + \beta_{1z}(g)'\mathbf{x}_i)},$$

for $z = C, T$ and $g \in \{\text{PBB}, \text{PbB}, \text{Pbb}, \text{pBB}, \text{pBb}, \text{pbb}\}$. Using the same justification as for the potential outcomes $Y(C)$ and $Y(T)$, we assume that $R_i(C)$ and $R_i(T)$ are conditionally independent.

For inference, we use the complete-data likelihood function, based on observing $\mathbf{Z}^{\text{obs}}$, $\mathbf{D}^{\text{obs}}$, $\mathbf{R}^{\text{obs}}$, $\mathbf{S}^{\text{obs}}$, $\mathbf{Y}^{\text{obs}}$, and $\mathbf{X}^{\text{obs}}$ as well as the vector of principal stratum indicators $\mathbf{G}$. Because we do not observe the principal stratum G_i of each unit, we cannot work directly with this complete-data likelihood function. In order to exploit it, we use the Data Augmentation (DA) algorithm (Tanner an Wong, 1987). In Appendix, we describe the numerical methods used to generate the inference reported in the next section.

In our application, the posterior distribution can be sensitive to the choice of prior distribution. For example, standard diffuse, improper prior distributions can lead to improper posterior distributions. We therefore use a proper prior distribution with a simple conjugate form. Our prior distribution corresponds to adding to the likelihood function 18 extra observations: there are 3 additional observations for each principal stratum $g \in$ (PBB, PbB, Pbb, pBB, pBb, pbb); for each principal stratum the 3 additional observations are split into $3/k(g)$ for each of the $k(g)$ combinations of the binary variables (Z_i, R_i, Y_i), where $k(g)$ varies across principal strata. Specifically, $k(g) = 6$ for $g = \text{PBB}, \text{pBB}$; $k(g) = 5$ for $g = \text{PbB}, \text{pBb}$; and $k(g) = 4$ for $g = \text{Pbb}, \text{pbb}$. These $3/k(g)$ observations are further split into $(3/k(g))/N$ artificial observation for each of the N observed values of the pretreatment variables, $\mathbf{X}^{\text{obs}}_i$. More formally, the prior distribution is proportional to

$$p(\theta) \propto \prod_{i=1}^{N} \times \prod_{g \in \{\text{PBB},\text{PbB},\text{Pbb},\text{pBB},\text{pBb},\text{pbb}\}} \times \prod_{z=C,T} \prod_{r=0,1} \prod_{y=L,H}$$
$$\left[\pi_i(g)\left(\pi^R_{iz}(g)\Big(\big(\pi^Y_{iz}(g)\big)^{\text{I}\{y=H\}}\big(1-\pi^Y_{iz}(g)\big)^{\text{I}\{y=L\}}\Big)^{\text{I}\{S(z)=B\}}\right)^r\Big(1-\pi^R_{iz}(g)\Big)^{(1-r)}\right]^{\frac{3}{k(g)N}}.$$

In the final model for the Faenza study data, we exclude age, so that we have two slope coefficients in each submodel. In addition, we impose prior equality of the slope coefficients in the response outcome regressions for compliers and never-takers: $\beta_{1C}(\text{PBB}) = \beta_{1T}(\text{PBB}) = \beta_{1C}(\text{PbB}) = \beta_{1T}(\text{PbB}) = \beta_{1C}(\text{Pbb}) = \beta_{1T}(\text{Pbb}) \equiv \beta_1(P)$, where $\beta_1(P) = (\beta_{X_1}(P), \beta_{X_2}(P))'$, and $\beta_{1C}(\text{pBB}) = \beta_{1T}(\text{pBB}) = \beta_{1C}(\text{pBb}) = \beta_{1T}(\text{pBb}) = \beta_{1C}(\text{pbb}) = \beta_{1T}(\text{pbb}) \equiv \beta_1(p)$, where $\beta_1(p) = (\beta_{X_1}(p), \beta_{X_2}(p))'$. Finally, we impose the requirement that the logistic parameters $\delta_{1z}(\text{PBB}) = \mathbf{0}$, $\delta_{1z}(\text{pBB}) = \mathbf{0}$ for $z = C, T$, and $\delta_{1T}(\text{PbB}) = \mathbf{0}$, and $\delta_{1C}(\text{pBb}) = \mathbf{0}$. Relaxing these restrictions would not complicate the computational methodology greatly, but given the relatively small sample size, would lead to imprecise estimates.

To demonstrate that our proper prior distribution does not lead to highly informative prior distribution for the estimands of interest, Table 4 presents summary statistics of the marginal prior distributions of the estimands of primary interest. The comparison of the summary statistics in Table 4 with the corresponding values in Tables 5 and 7, reported in the next sections, indicates that our prior distribution is relatively uninformative about quantities of interest.

Table 4: Summary statistics: prior distribution.

Estimand	Mean	s.d.
CACE on BSE practice	0.312	0.178
NACE on BSE practice	-0.339	0.186
ITT on BSE practice	0.005	0.167
CACE on BSE quality adjusted for BSE practice	-0.147	0.542
ITT on BSE quality adjusted for BSE practice	-0.086	0.334

7 Results

All results below were obtained from the same Bayesian analysis. We first focus on results for proportions of principal strata and for the CACE, NACE, and ITT estimands on BSE practice, which are functions of the proportions of principal strata. Then, we report results for the ITT and CACE estimands on quality of self-exam execution and for outcome response rates. Note that the reported estimands are not, in general, parameters of the models but functions of parameters and data.

7.1 Proportions in Principal Strata and Causal Effects on BSE Practice

Table 5 summarizes the posterior distribution of the estimands of the marginal probability of being a compliers, and of the conditional probability of being in a substratum defined by the vector $(S(C), S(T))$, given the compliance status $D(T)$. To draw from these distributions we use the steps: (1) draw θ, $D_i(T)$, and $(S_i(C), S_i(T))$ given $D_i(T)$, $i = 1, \ldots, N$, from the posterior distribution (see Appendix); (2) calculate $\Pr(D_i(T) = P \mid \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \theta)$, and

Table 5: Proportions of compliance principal strata and conditional proportions of BSE practice principal strata given compliance status - summary statistics of the posterior distributions.

Estimand	Mean	s.d.	Percentiles				
			2.5%	25%	50%	75%	97.5%
π^P	0.554	0.027	0.501	0.536	0.554	0.573	0.608
$\pi^{\mathrm{BB}}(P)$	0.667	0.072	0.530	0.618	0.669	0.717	0.810
$\pi^{\mathrm{bB}}(P)$	0.162	0.073	0.038	0.107	0.159	0.213	0.307
$\pi^{\mathrm{bb}}(P)$	0.171	0.061	0.072	0.122	0.164	0.214	0.295
$\pi^{\mathrm{BB}}(p)$	0.474	0.054	0.369	0.437	0.473	0.512	0.585
$\pi^{\mathrm{Bb}}(p)$	0.283	0.098	0.095	0.215	0.284	0.347	0.483
$\pi^{\mathrm{bb}}(p)$	0.243	0.096	0.050	0.175	0.243	0.309	0.425
ITT on BSE practice	-0.037	0.048	-0.125	-0.070	-0.039	-0.005	0.060

Table 6: Summary statistics of the posterior distributions of principal strata.

Estimand	Mean	s.d.	Percentiles				
			2.5%	25%	50%	75%	97.5%
π(PBB)	0.370	0.044	0.287	0.338	0.371	0.400	0.458
π(PbB)	0.090	0.040	0.022	0.059	0.088	0.118	0.171
π(Pbb)	0.095	0.034	0.040	0.067	0.090	0.119	0.166
π(pBB)	0.213	0.029	0.159	0.192	0.212	0.232	0.271
π(pBb)	0.126	0.045	0.041	0.095	0.126	0.157	0.218
π(pbb)	0.107	0.043	0.022	0.076	0.107	0.137	0.187

$\Pr(S_i(C), S_i(T) \mid D_i(T), \mathbf{X}_i^{\mathrm{obs}} = \mathbf{x}_i; \theta)$, for each subject based on the models in section 6; and (3) average the values of the first distribution in (2) over all the subjects to obtain $\Pr(D_i(T) = P \mid \theta)$, and the values of the second distribution in (2) over the subjects within subclasses defined by $D(T)$ to obtain $\Pr(S_i(C), S_i(T) \mid D_i(T); \theta)$. In addition, averaging the

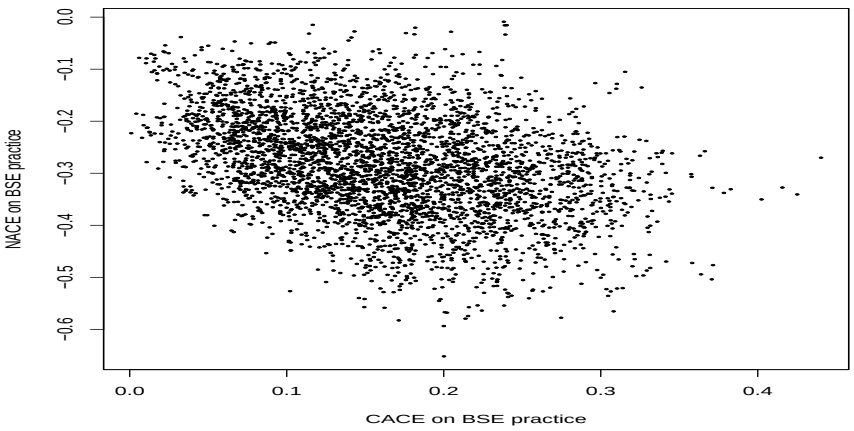

Figure 1: Simulation scatterplot of the joint posterior distribution of CACE and NACE on BSE practice.

product of the distributions in step (2) over all the subjects, we obtain a draw from the posterior distributions of principal strata, $\Pr(D(T), S(C), S(T) \mid \theta)$, which are summarized in Table 6.

The clearest pattern revealed by Tables 5 and 6 is that women who would practice BSE under both treatment arms are more likely to be compliers: 66.7% of compliers would practice the self exams under both assignments; in contrast only 47.4% of never-takers would practice BSE both if assigned to treatment and if assigned to control. Compliers are women who attend the enhanced BSE course if so assigned, and it appears plausible to believe that they are very aware of the risk of breast cancer. In contrast, never-takers are women who would not attend the BSE course in any case. If these women did not regard the risk of breast cancer as high enough to warrant BSE practice, they might consider BSE a boring task, and so practice the self exams only in one's spare time. These remarks agree with the idea underlying the monotonicity assumptions 2 and 3, which imply that the effect of invitation to partecipate to the BSE teaching course on BSE practice is positive for compliers, and negative for never-takers. Specifically, under our model, we have

$$\begin{aligned} \mathrm{E}\big(S_i(T) - S_i(C) \mid D_i(T) = P; \theta\big) &= \pi_i^{\mathrm{bB}}(P) \quad \text{by assumption 2} \\ \mathrm{E}\big(S_i(T) - S_i(C) \mid D_i(T) = p; \theta\big) &= -\pi_i^{\mathrm{Bb}}(p) \quad \text{by assumption 3,} \end{aligned}$$

where the expected values in the left sides are the CACE and the NACE estimands (on BSE practice) for individual i, respectively.

As shown in Table 5, our analysis suggests that there is a non-negligible and quite strong negative ITT effect on BSE practice for never-takers. Concerning the CACE estimand, we find that the Faenza teaching program would increase on average BSE practice of 16.2%, from 66.7% of compliers who received only a mailed informational leaflet to 82.9% of compliers who attended the training course, and this effect appears to be quite significant at the 5% level, according to a standard two-side t-test.

The marginal distributions of the subpopulation ITT effects on BSE practice show that the negative effect for never-takers is larger than the positive effect for compliers. Examining

their joint distribution in Figure 1, we see that these effects are negatively correlated. In addition, the presence of the strong negative NACE effect on BSE practice implies that the posterior distribution of the global ITT effect on S - summarized in the last row of Table 5 - has mass primarily ($> 75\%$) to the left of zero.

7.2 BSE Quality Results

Now, we address the following two questions:

1. What is the impact of being invited to participate to a "hands-on" teaching course on BSE techniques on examination skills, namely, the ITT estimand on BSE quality, Y?

2. What is the impact of attending a BSE training course on quality of self exams, namely, the CACE estimand on BSE quality, Y?

Recall that these causal estimands are defined on specific subpopulations of women: women (compliers or never-takers) who would practice BSE under both assignments and women belonging to the PBB group, respectively.

We cannot draw any meaningful inference about the causal effects on Y for women who would practice BSE under only one of the two treatment arms and for those who would not practice BSE in any case, because either $Y_i(C)$ or $Y_i(T)$ is not defined on the sample space $\{L, H\}$ for these types of subjects.

Table 7 shows summary statistics of the posterior distributions of the estimands of interest about BSE quality.

Table 7: Summary statistics of the posterior distributions of the quality causal estimands.

Estimand	Mean	s.d.	Percentiles				
			2.5%	25%	50%	75%	97.5%
CACE on BSE quality adjusted for BSE practice	0.174	0.101	-0.018	0.103	0.170	0.242	0.379
ITT on BSE quality adjusted for BSE practice	0.110	0.063	-0.011	0.065	0.108	0.152	0.237
$\mathrm{E}\big(Yi(C) \mid G_i = \mathrm{PBB}; \theta\big)$	0.701	0.073	0.560	0.653	0.701	0.749	0.842
$\mathrm{E}\big(Yi(T) \mid G_i = \mathrm{PBB}; \theta\big)$	0.875	0.079	0.742	0.815	0.861	0.934	1.000
$\mathrm{E}\big(Yi(T) \mid G_i = \mathrm{PbB}; \theta\big)$	0.278	0.203	0.011	0.110	0.236	0.421	0.710
$\mathrm{E}\big(Yi(C) \mid G_i = \mathrm{pBB}; \theta\big)$	0.448	0.091	0.272	0.386	0.448	0.510	0.627
$\mathrm{E}\big(Yi(T) \mid G_i = \mathrm{pBB}; \theta\big)$	0.448	0.091	0.272	0.386	0.448	0.510	0.627
$\mathrm{E}\big(Yi(C) \mid G_i = \mathrm{pBb}; \theta\big)$	0.407	0.186	0.057	0.275	0.412	0.534	0.771

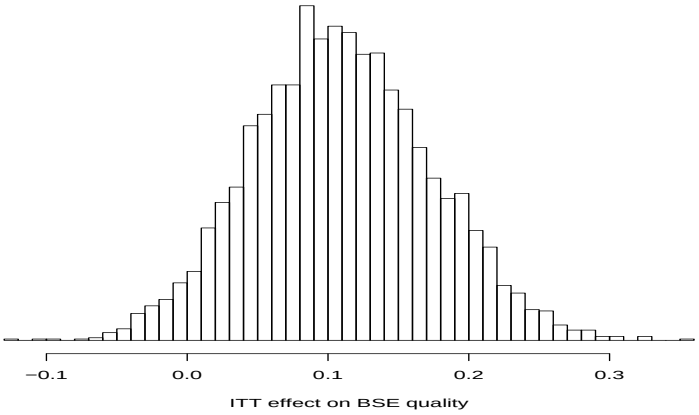

Figure 2: Simulation histogram of the posterior distribution of the intention-to-treat effect on BSE quality.

Effect of Offering the Teaching Program on BSE Quality

We examine the impact of being offered a teaching program on BSE quality among women who would practice BSE under both assignments: the ITT effect on posttreatment quality of self exams adjusted for the intermediate outcome "BSE practice", S. The corresponding estimand for individual i is defined as

$$\mathrm{E}\big(Y_i(T) - Y_i(C) \mid S_i(C) = B, S_i(T) = B; \theta\big).$$

The simulated posterior distribution of this ITT effect is summarized in the second row in Table 7. Figure 2 shows its histogram. To draw from this distribution, we use the same step (1) described in the previous section and then calculate the expected causal effect $\mathrm{E}\big(Y_i(T) - Y_i(C) \mid D_i(T), S_i(C) = B, S_i(T) = B, \mathbf{X}_i^{\mathrm{obs}} = \mathbf{x}_i; \theta\big)$ for each subject based on the BSE quality submodel (see section 6), and average these values over the subjects whose current draw of $(S_i(C), S_i(T))$ is (B, B).

The estimate of the average ITT effect on BSE quality is approximately equal to 11%, with a standard deviation of 0.063. The posterior probability that this ITT effect is positive, that is, that the invitation to the BSE teaching program improves the examination skills, is approximately 96.3%. Thus, there appears to be some evidence that the teaching course can improve the quality of self-exam execution, although a standard two-side t-test suggests this is not quite significant at the 5% level: the 95% posterior interval of our ITT effect on BSE quality covers zero.

Effect of BSE Teaching Course on BSE Quality

We now examine the effect of offering the BSE teaching program on BSE quality outcome, Y, by focusing only on the compliers who would practice BSE under both assignments - the CACE on BSE quality, which for individual i is defined as

$$\mathrm{E}\big(Y_i(T) - Y_i(C) \mid D_i(T) = P, S_i(C) = B, S_i(T) = B; \theta\big).$$

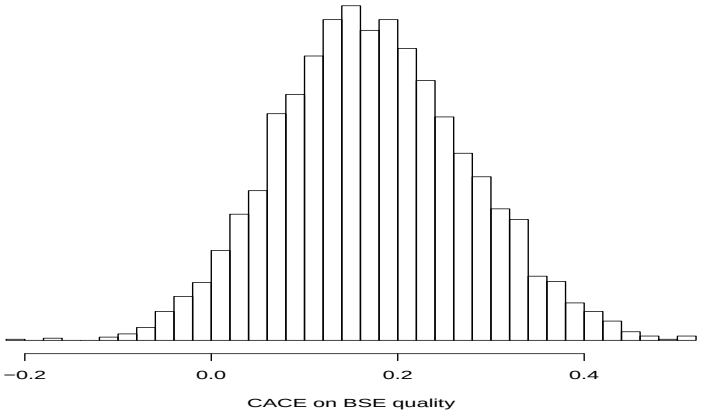

Figure 3: Simulation histogram of the posterior distribution of the complier average causal effect on BSE quality.

This analysis defines the treatment as attendance of the training course on BSE techniques. The simulated posterior distribution of the CACE on BSE quality is summarized in the first row in Table 7, and Figure 3 shows its histogram. A draw from this distribution is obtained using the same steps described above for the ITT estimand (on BSE quality) with the exception that now the averaging is restricted to the subjects who belong to the PBB principal stratum in the current draw.

The effect of attending the teaching program on BSE quality follows a similar pattern to that of the ITT effect on Y, but the posterior mean is slightly bigger than ITT. The posterior interval has also grown, reflecting that this effect is for only a specific subpopulation of all women who would practice BSE under both treatment arms, those who belong to the PBB group.

As we can see in Table 7, our analysis suggests that there is some evidence that the BSE training course has some beneficial effect on self examination skills, although it is not much significant. In fact, compliers who would practice BSE under both treatment arms tend to execute self exams with high quality both if assigned to treatment and if assigned to control. This gives reason for believing that women belonging to the PBB group are very sensitive to the risk of breast cancer, and so they tend to practice BSE correctly.

7.3 Missing Outcomes

As stated earlier, theoretically under our structural assumptions, an analysis based on ad-hoc approaches to missing data would be likely not appropriate for evaluating the causal estimands of interest because it is highly probable that principal strata have differential response (i.e., outcome missing data) behaviors. To evaluate this here, we simulated the posterior distributions of

$$\Pr\big(R_i(z) \mid G_i = g; \theta\big) \qquad z = C, T.$$

To draw from the distributions of these estimands, we use step (1) described in section 7.1 and then, for $z = C, T$, for each subject, calculate $\Pr\big(R_i(z) \mid G_i = g, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \theta\big)$. We

then average these values over subjects within subclasses defined by the principal stratum indicator G_i. Table 8 presents some summary statistics of these posterior distributions.

Table 8: Summary statistics of the posterior distributions of the response outcome probability

Estimand	Mean	s.d.	Percentiles				
			2.5%	25%	50%	75%	97.5%
$\Pr\big(R_i(\cdot)=1 \mid G_i=\text{PBB};\theta\big)$	0.928	0.069	0.765	0.885	0.949	0.987	1.000
$\Pr\big(R_i(\cdot)=1 \mid G_i=\text{PbB};\theta\big)$	0.579	0.267	0.055	0.385	0.577	0.804	0.995
$\Pr\big(R_i(\cdot)=1 \mid G_i=\text{Pbb};\theta\big)$	0.549	0.201	0.247	0.390	0.517	0.694	0.965
$\Pr\big(R_i(C)=1 \mid D_i(T)=p;\theta\big)$	0.575	0.070	0.440	0.529	0.575	0.623	0.713
$\Pr\big(R_i(T)=1 \mid D_i(T)=p;\theta\big)$	0.420	0.041	0.340	0.392	0.419	0.449	0.502

Our missing data model gives plausible figures for the response probabilities. The response rate among compliers, irrespective of their BSE practice behavior, is approximately 80.7%. This implies that never-takers have lower response rates than compliers per assigned level: 0.420 versus 0.807 for those assigned to the active treatment, 0.575 versus 0.807 for those assigned to 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 be 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. Concerning the compliers' response behavior, we find that compliers who would practice BSE under both assignments have the highest response rate. The PbB group and the Pbb group have similar response rates and they are significantly lower than the response rate of the PBB group. These results confirm that the latent covariates $D(T)$ and $(S_i(C), S_i(T))$ are important factors in response, alone as well as in interaction with assignment.

8 Model Building and Checking

This model was built through a process of fitting and checking a succession of models. The key features of our model are linked to the choice of the form of the likelihood function and its associated prior distribution. We emphasize the use of "weakly identified" models: "identified" in the sense of having a proper prior distribution, but "weakly" in the sense of not having unique maximum likelihood estimates.

8.1 Convergence Checks

Because posterior distributions were simulated from an MCMC algorithm (Appendix), it is important to assess its convergence. To do this, we ran four chains from some overdispersed initial distribution and compare their realizations. As initial distribution, we took a multivariate normal distribution derived from a simulation on a single chain, and inflated the variance matrix. Each chain was run for 25,000 iterations. At 5,000 iteration, and based on the four chains, we calculated the potential scale reduction (Gelman and Rubin, 1992) for each estimand. The results suggested good mixing of the chains and provided no evidence against convergence. Posterior inference is based on a single chain, which was run for 97,500 iterations after the burn-in stage, saving every 25th iteration. For the prior distribution, the chain was run for 45000 iteration after burn-in, saving every 10th iteration.

8.2 Model Checks

We evaluate the influence of the model presented in section 6 using posterior predictive checks. A posterior predictive check generally involves: (a) choosing a discrepancy measure, and (b) computing a posterior predictive p-value (PPPV) (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996). In our application, a posterior predictive p-value can be measured by

$$\Pr\Big(\Lambda\big(\text{data}^{\text{rep}}, \mathbf{G}^{\text{rep}}, \theta\big) \geq \Lambda\big(\text{data}, \mathbf{G}, \theta\big) \mid \text{data}\Big),$$

where $\Lambda(\cdot,\cdot,\cdot)$ is a discrepancy variable, $\mathbf{G}$ is the vector of the missing latent group indicators, and data^{rep} and $\mathbf{G}^{\text{rep}}$ are drawn from their joint posterior predictive distribution.

Posterior predictive checks in general, and PPPVs in particular, can be used for judging whether the model can adequately preserve features of the data reflected in the discrepancy measure, where the model here includes the prior distribution as well as the likelihood (Meng, 1994).

To evaluate the fit of our Bayesian model to the observed data, we use ten posterior predictive checks: six checks for BSE practice outcome, three for quality outcome and one to assess the monotonicity assumptions 2 and 3. To get a more efficient test of the model, we have fixed the number of women assigned to treatment and the number of women assigned to control in the replicated data to be the same as those in the observed data.

The first six posterior predictive discrepancy measures we used here are function of

$$\mathcal{A}^{\text{rep}}_{d,z}(S) = \big\{S^{\text{rep}}_i : \text{I}\{R^{\text{rep}}_i = 1\}\text{I}\{D^{\text{rep}}_i(T) = d\}\text{I}\{Z_i = z\} = 1\big\},$$

for the measures that are functions of data S^{rep}_i, R^{rep}_i, and $D^{\text{rep}}_i(T)$ from a replicated study; and,

$$\mathcal{A}^{\text{study}}_{d,z}(S) = \big\{S_i : \text{I}\{R_i = 1\}\text{I}\{D_i(T) = d\}\text{I}\{Z_i = z\} = 1\big\},$$

for the measures that are functions of our study's data. Following Barnard et al. (2002), the discrepancy measures, "rep" and "study", that we used for BSE practice outcome in each subpopulations defined by the compliance status $D(T)$ were (1) the absolute value of the difference between the BSE practice rate of $\mathcal{A}_{d,T}(S)$ and the BSE practice rate of $\mathcal{A}_{d,C}(S)$ ("signal"), (2) the standard error based on a simple two sample comparison for this difference ("noise"), and (3) the ratio of (1) and (2) ("signal to noise").

For the quality outcome, Y, we calculated the same discrepancy measures just defined, focusing on the subpopulation of compliers who practice BSE under both assignments. So the posterior predictive checks we chose for the quality outcome are functions of

$$\mathcal{A}_z^{\text{rep}}(Y) = \left\{Y_i^{\text{rep}} : \text{I}\{R_i^{\text{rep}} = 1\}\text{I}\{G_i^{\text{rep}} = \text{PBB}\}\text{I}\{Z_i = z\} = 1\right\},$$

for the measures that are functions of data Y_i^{rep}, R_i^{rep}, and G_i^{rep} from a replicated study; and,

$$\mathcal{A}_z^{\text{study}}(Y) = \left\{Y_i : \text{I}\{R_i = 1\}\text{I}\{G_i = \text{PBB}\}\text{I}\{Z_i = z\} = 1\right\},$$

for the measures that are functions of our study's data.

Although these measures are not treatment effects, we chose them here for assessing whether the model can preserve broad features of signal, noise, and signal to noise ratio in the involved distributions, which we think can be very influential in estimating the treatment effects of section 7. More preferable measures might have been the posterior mean and standard deviation for the actual estimands in section 7 for each replicated dataset but this required a prohibitive amount of computer memory due to the nested structure of that algorithm.

As stated previously, our inferential results depend heavily on the structural assumptions described in section 5. Among these, the two monotonicity assumptions 2 and 3 have a strong impact on the estimation of causal effects of interest, so assessment of them appears crucial.

To determine if the monotonicity assumptions 2 and 3 are supported by our data, we use the log-likelihood ratio discrepancy:

$$\text{L}^2(\text{data}, \mathbf{G}, \theta) = 2\sum_g n(g) \ln\left(\frac{n(g)}{\hat{n}(g)}\right),$$

where $g \in \{\text{PBB}, \text{PbB}, \text{Pbb}, \text{pBB}, \text{pBb}, \text{pbb}\}$, ln is the natural logarithm with $0 \ln 0 = 0$ by convention, $n(g)$ is the number of women in the principal stratum g, and $\hat{n}(g)$ represents the number of women estimated to belong to the g group under the model.

PPPVs for the discrepancy measures we chose were calculated as the percentage of draws in which the replicated discrepancy measures exceeded the value of the study's discrepancy measures. Extreme values, close to 0 or 1, of a PPPV would indicate a failure of the prior distribution and likelihood to replicate the corresponding measure of location, dispersion or their relative magnitude, and would indicate an undesirable influence of the model in estimation of our estimands. Results from these checks, displayed in Table 9 provide no special evidence for such influence of the model.

9 Concluding Remarks

In this paper, we defined the framework for principal stratification in randomized experiments to accomodate noncompliance, missing outcome data, and truncation by "death". We make explicit a set of structural assumptions, and we provide a parametric model that is appropriate for practical implementation of the framework in setting such as ours.

Results from our model in the Faenza BSE study suggest that the treatment has some beneficial effects on BSE practice. They also show a strong evidence that the encouragement has a negative effect on BSE practice for women who would not attend the teaching program

Table 9: Posterior predictive p-values.

	Signal	Noise	Signal to noise
BSE practice - compliers	0.629	0.896	0.576
BSE practice - never-takers	0.263	0.571	0.259
BSE quality - PBB group	0.548	0.816	0.481
PPPV based on the log-likelihood ratio discrepancy: 0.783			

regardless of the encouragement, the never-takes. We interpret this result as evidence that never-takers do not regard the risk of breast cancer as high enough to warrant the attendance of the course. Concerning BSE quality, our results do not indicate strong treatment effects.

Our analysis also reveals significant differences in missing data pattern across principal strata. We find that women who 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 partecipate in the treatment program. Among compliers, our results suggest that women, who are probably more sensible to the risk of breast cancer and have likely higher motivation (women belonging to the PBB group), are more willing to respond to the survey.

Appendix

Details of Calculations

Our approach to inference treats the latent principal strata $\mathbf{G} = (G_i, \ldots, G_N)$ as missing data and applies modern missing data technology for Bayesian models.

We construct a general state Markov chain that has the joint distribution of the model parameters θ and the missing latent group indicators $\mathbf{G}$ as its unique invariant equilibrium distribution. The Markov chain algorithm is a variant of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), which use the Data Augmentation (DA) method of Tanner and Wong (1987). The algorithm can be described as follows. Let $(\mathbf{G}^{(j)}, \theta^{(j)})$ denote the state of the chain at time j, where $\mathbf{G}^{(j)}$ depends on the current value of the matrix $[\mathbf{D(T)} \mid \mathbf{S(C)} \mid \mathbf{S(T)}]$. The state of the chain at time $j+1$ follows from applying the following steps.

First, we draw $\mathbf{D(T)}^{(j)}$ according to $\Pr(D(T) = P \mid \theta^{(j)}, \mathrm{W})$ where we use $\mathrm{W} = (\mathbf{Z}^{\mathrm{obs}}, \mathbf{D}^{\mathrm{obs}}, \mathbf{R}^{\mathrm{obs}}, \mathbf{S}^{\mathrm{obs}}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{X}^{\mathrm{obs}})$ to simplify the notation. This conditional distribution has a simple form. Conditional on θ and W, the $D_i(T)$ are independent of $D_j(T)$, Z_j^{obs}, D_j^{obs}, R_j^{obs}, S_j^{obs}, Y_j^{obs} for all $j \neq i$. Then,

$$\begin{aligned} \Pr\big(D_i(T) = P \mid Z_i^{\mathrm{obs}} = T, D_i^{\mathrm{obs}} = P, R_i^{\mathrm{obs}}, S_i^{\mathrm{obs}}, Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) &= 1; \\ \Pr\big(D_i(T) = P \mid Z_i^{\mathrm{obs}} = T, D_i^{\mathrm{obs}} = p, R_i^{\mathrm{obs}}, S_i^{\mathrm{obs}}, Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) &= 0. \end{aligned}$$

Finally, for observations with $Z_i^{\text{obs}} = C$, who have $D_i^{\text{obs}} = p$ by construction in our experimental setting, we have

$$\Pr\big(D_i(T) = P \mid Z_i^{\text{obs}} = C, D_i^{\text{obs}} = p, R_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big) \propto$$

$$\begin{aligned}
\pi_i^P \Bigg[& \Big(\pi_i^{\text{BB}}(P)\pi_{iC}^R(\text{PBB})\big(\pi_{iC}^Y(\text{PBB})\big)^{\text{I}\{Y_i^{\text{obs}}=H\}} \\
& \big(1-\pi_{iC}^Y(\text{PBB})\big)^{\text{I}\{Y_i^{\text{obs}}=L\}}\Big)^{\text{I}\{R_i^{\text{obs}}=1, S_i^{\text{obs}}=B\}} \\
& \Big(\pi_i^{\text{bB}}(P)\pi_{iC}^R(\text{PbB}) + \pi_i^{\text{bb}}(P)\pi_{iC}^R(\text{Pbb})\Big)^{\text{I}\{R_i^{\text{obs}}=1, S_i^{\text{obs}}=b\}} \\
& \Big(\pi_i^{\text{BB}}(P)\big(1-\pi_{iC}^R(\text{PBB})\big) + \pi_i^{\text{bB}}(P)\big(1-\pi_{iC}^R(\text{PbB})\big) + \\
& \pi_i^{\text{bb}}(P)\big(1-\pi_{iC}^R(\text{Pbb})\big)\Big)^{\text{I}\{R_i^{\text{obs}}=0\}}\Bigg].
\end{aligned}$$

At the second step, we draw $(\mathbf{S(C)}, \mathbf{S(T)})^{(j+1)}$ given $\mathbf{D(T)}^{(j+1)}$, the current state of $\mathbf{D(T)}$, according to $\Pr(S(C), S(T) \mid D(T)^{(j+1)}, \theta^{(j)}, \text{W})$. First focus on $D(T)^{(j+1)} = P$. The monotonicity assumption 2 implies that:

$$\Pr\big(S_i(C)=s_C, S_i(T)=s_T \,|\, D_i(T)^{(j+1)} = P, \text{OBS}(T,P,1,b), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big)=0,$$

for $(s_C, s_T) \in \{(B,B),(b,B)\}$, and

$$\Pr\big(S_i(C)=b, S_i(T)=b \,|\, D_i(T)^{(j+1)} = P, \text{OBS}(T,P,1,b), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big)=1.$$

In addition, by the monotonicity assumption 2

$$\Pr\big(S_i(C)=B, S_i(T)=B \,|\, D_i(T)^{(j+1)} = P, \text{OBS}(C,p,1,B), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big)=1$$

and

$$\Pr\big(S_i(C)=s_C, S_i(T)=s_T \,|\, D_i(T)^{(j+1)} = P, \text{OBS}(C,p,1,B), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big)=0$$

for $(s_C, s_T) \in \{(b,B),(b,b)\}$. Lastly, the assumption 2 implies that

$$\Pr\big(S_i(C)=B, S_i(T)=b \mid D_i(T)^{(j+1)} = P, \text{OBS}(Z_i^{\text{obs}}, D_i^{\text{obs}}, R_i^{\text{obs}}, S_i^{\text{obs}}), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big) = 0,$$

for each $\text{OBS}(Z_i^{\text{obs}}, D_i^{\text{obs}}, R_i^{\text{obs}}, S_i^{\text{obs}}) = \text{OBS}(T,P,1,B)$, $\text{OBS}(T,P,1,b)$, $\text{OBS}(T,P,0,?)$, $\text{OBS}(C,p,1,B)$, $\text{OBS}(C,p,1,b)$, $\text{OBS}(C,p,0,?)$. Obviously,

$$\Pr\big(S_i(C)=s_C, S_i(T)=s_T \mid D_i(T)^{(j+1)} = P, \text{OBS}(T,p,R_i^{\text{obs}}, S_i^{\text{obs}}), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big) = 0$$

for each $(s_C, s_T) \in \{(B,B),(b,B),(b,b)\}$, and $(R_i^{\text{obs}}, S_i^{\text{obs}}) \in \{(1,B),(1,b),(0,?)\}$.

It remains to consider the observed groups $\text{OBS}(T,P,1,B)$, $\text{OBS}(T,P,0,?)$, $\text{OBS}(C,p,1,b)$, and $\text{OBS}(C,p,1,?)$ for observations with $D_i(T)^{(j+1)} = P$. For the women who are observed fall into the group $\text{OBS}(T,P,1,B)$, we have

$$\begin{aligned}
\Pr\big(S_i(C)=B, S_i(T)=B \mid D_i(T)^{(j+1)} = P, \text{OBS}(T,P,1,B), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\big) \propto \\
\pi_i^{\text{BB}}(P)\pi_{iT}^R(\text{PBB})\big(\pi_{iT}^Y(\text{PBB})\big)^{\text{I}\{Y_i^{\text{obs}}=H\}}\big(1-\pi_{iT}^Y(\text{PBB})\big)^{\text{I}\{Y_i^{\text{obs}}=L\}};
\end{aligned}$$

$$\Pr\big(S_i(C)=b, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(T,P,1,B), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto$$
$$\pi_i^{\mathrm{bB}}(P)\pi_{iT}^{R}(\mathrm{PbB})\big(\pi_{iT}^{Y}(\mathrm{PbB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=H\}}\big(1-\pi_{iT}^{Y}(\mathrm{PbB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=L\}};$$

$$\Pr\big(S_i(C)=b, S_i(T)=b \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(T,P,1,B), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big)=0.$$

For women belonging to the $\mathrm{OBS}(T,P,0,?)$ group:

$$\Pr\big(S_i(C)=B, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(T,P,0,?), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{BB}}(P)\big(1-\pi_{iT}^{R}(\mathrm{PBB})\big);$$

$$\Pr\big(S_i(C)=b, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(T,P,0,?), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{bB}}(P)\big(1-\pi_{iT}^{R}(\mathrm{PbB})\big);$$

$$\Pr\big(S_i(C)=b, S_i(T)=b \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(T,P,0,?), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{bb}}(P)\big(1-\pi_{iT}^{R}(\mathrm{Pbb})\big).$$

For the women who are observed fall into the group $\mathrm{OBS}(C,p,1,b)$, we have

$$\Pr\big(S_i(C)=B, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(C,p,1,b), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big)=0$$

$$\Pr\big(S_i(C)=b, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(C,p,1,b), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{bB}}(P)\pi_{iC}^{R}(\mathrm{PbB});$$

$$\Pr\big(S_i(C)=b, S_i(T)=b \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(C,p,1,b), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{bb}}(P)\pi_{iC}^{R}(\mathrm{Pbb}).$$

Lastly, for the women who are observed fall into the group $\mathrm{OBS}(C,p,0,?)$, we have

$$\Pr\big(S_i(C)=B, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(C,p,0,?), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{BB}}(P)\big(1-\pi_{iC}^{R}(\mathrm{PBB})\big);$$

$$\Pr\big(S_i(C)=b, S_i(T)=B \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(C,p,0,?), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{bB}}(P)\big(1-\pi_{iC}^{R}(\mathrm{PbB})\big);$$

$$\Pr\big(S_i(C)=b, S_i(T)=b \mid D_i(T)^{(j+1)}=P, \mathrm{OBS}(C,p,0,?), Y_i^{\mathrm{obs}}, \mathbf{X}_i^{\mathrm{obs}}\big) \propto \pi_i^{\mathrm{bb}}(P)\big(1-\pi_{iC}^{R}(\mathrm{Pbb})\big).$$

The drawing of $(S_i(C), S_i(T))$ for women whose current draw of $D(T)^{(j+1)}$ is p, is done in a similar way, using the monotonicity assumption 3.

We then draw for the following subvectors of θ in sequence, conditional on all others: $\{\alpha\}$; $\{\gamma_P^{\mathrm{bB}}\}$; $\{\gamma_P^{\mathrm{bb}}\}$; $\{\gamma_p^{\mathrm{Bb}}\}$; $\{\gamma_p^{\mathrm{bb}}\}$; $\{\beta_{0\cdot}(\mathrm{PBB})\}$; $\{\beta_{0\cdot}(\mathrm{PbB})\}$; $\{\beta_{0\cdot}(\mathrm{Pbb})\}$; $\{\beta_{1\cdot}(P)\}$; $\{\beta_{0C}(p)\}$; $\{\beta_{0T}(p)\}$; $\{\beta_{1\cdot}(p)\}$; $\{\delta_{0C}(\mathrm{PBB})\}$; $\{\delta_{0T}(\mathrm{PBB})\}$; $\{\delta_{0T}(\mathrm{PbB})\}$; $\{\delta_{0\cdot}(\mathrm{pBB})\}$; and $\{\delta_{0C}(\mathrm{pBb})\}$. Recall that we impose some prior equalities of the slope coefficients and some other equalities are implied by the exclusion restrictions.

For the parameters $\delta_{0C}(\mathrm{PBB})$, $\delta_{0\cdot}(\mathrm{pBB})$, and $\delta_{0C}(\mathrm{pBb})$ we know the full conditional distributions (they are Beta distributions), so we can directly draw from them.[2] The parameter $\delta_{0T}(\mathrm{PbB})$ is drawn from a truncated Beta distributions. For the other subvectors of θ, in our specification, it is rather difficult to draw directly from the appropriate conditional distributions, however, it is straightforward to calculate the (complete-data) posterior density up to a normalizing constant at any parameter value, so we can use Metropolis-Hastings steps. For example, to draw α, we draw *candidate* values α^* from a density $\mathrm{g}(\alpha \mid \theta^{(j)})$. The candidate draw is accepted with probability

$$\tau = \min\left\{\frac{p(\alpha^*, \gamma^{(j)}, \beta^{(j)}, \delta^{(j)} \mid \mathrm{W}, \mathbf{G})}{p(\alpha^{(j)}, \gamma^{(j)}, \beta^{(j)}, \delta^{(j)} \mid \mathrm{W}, \mathbf{G})} \cdot \frac{\mathrm{g}(\alpha^{(j)} \mid \alpha^*, \gamma^{(j)}, \beta^{(j)}, \delta^{(j)})}{\mathrm{g}(\alpha^* \mid \alpha^{(j)}, \gamma^{(j)}, \beta^{(j)}, \delta^{(j)})}, 1\right\},$$

[2] If we did not impose assumption 8 - the stochastic dominance assumption of the PBB group over the PbB group - also the full conditional distributions of the parameters $\delta_{0T}(\mathrm{PBB})$ and $\delta_{0T}(\mathrm{PbB})$ would be Beta distributions.

where p is the posterior density, up to a normalizing constant, of the parameter vector. For the candidate density g, we use a vector of scaled t-Student random variables with five degrees of freedom, centered at $\alpha^{(j)}$. This has the convenient property that

$$g(\alpha^* \mid \alpha^{(j)}, \gamma, \beta, \delta) = g(\alpha^{(j)} \mid \alpha^*, \gamma, \beta, \delta),$$

simplifying the expression for τ slightly.

The scaling factors were chosen based on preliminary runs of the chain. It is desirable to strike a balance between rejecting too often and rejecting too infrequently, so that the resulting chain will cover the support of the target distribution relatively efficiently, not staying at the same point too much but also not taking steps that are too small.

References

Angrist J. D., Imbens, G. W., and Rubin D. B. (1996) Identification of causal effects using instrumental variables, (with Discussion). *Journal of the American Statistical Association*, **91**, 444–472.

Barnard J., Frangakis, C. E., Hill J. L. and Rubin, D. B. (2003) Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, **98**, 299–311.

Cox, D. R. (1958) *Planning of experiment*. New York, Wiley.

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, 4**, 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 treat-ment-non-compliance 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.

Gelman, A., and Rubin, D. B. (1992) Inference from iterative simulations using multiple sequences. *Statistical Science*, **7**, 457–511.

Gelman, A., Meng, X.-L., and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies (Disc: P760-807). *Statistica Sinica*, **6**, 733–760.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hirano, K., Imbens, G. W., Rubin D. B., and Zhou, X. H. (2000) Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1**, 69–88.

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

Imbens, G. W. and Rubin, D. B. (1997) Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, **25**, 305–327.

Little, R. J. A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R. J. A. (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, **52**, 98–111.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical analysis with missing data.* Wiley, New York.

Mattei, A., and Mealli, F. (2004) Assessing the Effect of a Teaching Program on Breast Self-Examination in a Randomized Trial with Noncompliance and Missing Data. *Proceedings of the 19th International Workshop on Statistical Modelling*, 455–459.

Mealli, F., and Rubin, D. B. (2002) Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services and Outcomes Research Methodology*, **3**, 225–232.

Mealli, F., Imbens, G. W., Ferro, S. and Biggeri, A. (2004) Analyzing randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*, **5** 207–222.

Meng, X.-L. (1994) Posterior predictive p-values. *Annals of Statistics*, **22**, 1142–1160.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

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

Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, **6**, 34–58.

Rubin, D. B. (1980) Comments on "Randomization analysis of experimental data; the Fisher randomization test". *Journal of the American Statistical Association*, **75**, 591–593.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151–1172.

Rubin, D. B. (2000) Comment on "Causal inference without counterfactual" by P. Dawid. *Journal of American Statistical Association*, **95**, 407–448.

Tanner, M., and Wong W. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of American Statistical Association*, **82**, 528–550.

Zhang, J. L. (2002) *Causal inference with principal stratification: some theory and applications* PH.D Thesis, Harvard University, Cambridge (MA).

Zhang, J. L. and Rubin, D. B. (2003) Estimation of causal effects via principal stratification when some outcomes are truncated by "death". Forcoming in *Journal of Educational and Behavioral Statistics*.