UNIVERSITÀ DEGLI STUDI DI FIRENZE

DIPARTIMENTO DI STATISTICA "G. PARENTI"

# ESTIMATING CAUSAL EFFECTS IN EXPERIMENTAL AND OBSERVATIONAL STUDIES SUFFERING FROM MISSING DATA

TESI DI DOTTORATO IN STATISTICA APPLICATA – XVII CICLO

Alessandra Mattei

*Director of graduate studies: Fabrizia Mealli*
*Supervisor: Fabrizia Mealli*

December 2004

# Preface

The research questions that motivate most studies in statistics-based sciences are causal in nature. Yet until very recently, the dominant methodology has been based almost exclusively on statistical analysis which, traditionally, has excluded causal vocabulary both from its mathematical language and from its educational program.

The aim of standard analysis, typified by regression and other estimation techniques, is to infer parameters of a distribution from samples drawn of that population. With the help of such parameters, one can infer associations among variables, estimates the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer aspects of the data generating process. With the help of such aspects, one can deduce not only the likelihood of events under static conditions, but also the dynamics of events under changing conditions. This capability include predicting the effects of interventions, (e.g., treatments or policy decisions) and spontaneous changes, (e.g., epidemics or natural disasters), identifying cause of reported events, and assessing responsibility and attribution.

This distinction implies that causal and statistical concepts do not mix. Statistic deals static conditions, while causal analysis deals with changing conditions. There is nothing in distribution function that would tell us how that distribution would differ if external conditions were to change because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. Even the theory of stochastic processes, which provides probabilistic characterization of certain dynamic phenomena, assumes a fixed density function over time-indexed variables. There is nothing in such a function to tell us how it would be altered if external conditions were to change. The additional information needed for making such predictions is provided by causal assumptions. The role of this information is to identify those aspects of the world that remain invariant when external conditions change, say due to treatments or policy decisions.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: one cannot substantiate causal claims from associations alone, even at the population level - behind every causal conclusion there must lie some causal assumption that is not testable in observational studies.

One difficulty that arises in talking about causation is the variety of questions that are subsumed under the heading. Some authors focus on the ultimate meaningfulness of the notion of causation. Others are concerned with deducing the causes of a given effect. Still others are interested in understanding the details of causal mechanism. The emphasis here will be on measuring the effects of causes because this seems to be a place where statistics, which is concerned with measurement, has contributions to make. The purpose is to construct a model that is complex enough to allow us to formalize basic intuitions concerning cause and effect.

A statistical framework for causal inference that has received increasing attention in recent years is the one based on potential outcomes. It is rooted in the statistical work on randomized experiments by Fisher (1918, 1925) and Neyman (1923), as extended by Rubin (1974, 1976, 1977, 1978, 1990) and subsequently by others to apply to nonrandomized studies and other forms of inference. This perspective was called "Rubin's Causal Model" (RCM, Holland, 1986) by because of it viewed causal inference as a problem of missing data with explicit mathematical modeling of the assignment mechanism as a process for revealing the observed data. The RCM allows the direct handling of complications, such as noncompliance with assigned treatment (wich bridges experiments and the econometric instrumental variables methods, Angrist et al., 1996). In the late 1980s and 1990s, many economists have accepted and adopted this framework as well (Bjorklund and Muffit, 1987; Heckman, 1989; Manski, 1990; Manski et al., 1992, Angrist and Imbens, 1995) because of the clarity it brings in questions of causality.

Assume that there are just two levels of treatment, denoted by $T$, the active treatment, and $C$, the control. The starting essential feature of the approach is to define a causal effect as the comparison of the potential outcomes on the same unit measured at the same time: $Y(C)$, the value of the outcome variable $Y$ if the unit is exposed to treatment $C$, and $Y(T)$, the value of $Y$ if exposed to $T$. Only one of these two potential outcomes can be observed, the one corresponding to the treatment that unit received, yet causal effects are defined by their comparison, e.g., $Y(T) - Y(C)$. Thus, causal inference becomes a problem of inference with missing data. Without potential outcomes, causal inference is exceedingly difficult and often misleading.

The assignment mechanism is then a stochastic rule for assigning treatments to units and thereby for revealing $Y(C)$ or $Y(T)$ for each unit. The assignment mechanism can depend on other measurements; if these other measurements are observed values, then the assignment mechanism is ignorable; if the given observed values involve missing values, possibly even missing $Y$'s, then it is nonignorable. All forms of statistical inference for causal effects, whether Bayesian or frequentist, require the positing of an assignment mechanism.

## *Outline of the Thesis*

In the view we accept of causality, inference for causal effects is a missing-data problem, because for any individual unit, we observe the value of the potential outcome under only one of the possible treatments, namely the treatment actually assigned, and the potential outcome under the other treatment is missing. The statistical literature classifies this type of missing-data mechanism as an *intentional* missing-data mechanism - intentional in the sense that it depends on the assignment mechanism and can be generally considered explicitly. Clearly, analysis of treatment response poses much more than a generic missing-data problem. One reason is that observations of realized outcomes, when combined with suitable assumptions, can provide information about counterfactual ones. Another is that practical problems of treatment choice motivate much research on treatment response and thereby determine what population parameters are of interest.

Unfortunately, inference on causal effects, by definition subject to an *intentional* missing-data mechanism, is often complicated by *unintentional* missing-data processes.

In general, when we analyze data subject to missing values, it can be useful to distinguish the missing-data pattern, which describes which values are observed in the data matrix and which values are missing, and the missing-data mechanism (or mechanisms), which concerns the relation between missingness and the values of variables in the data matrix. Missing-data mechanisms are crucial since the properties of the missing-data methods depend very strongly on the nature of the dependence in these mechanisms. The crucial role of the mechanism in the analysis of data with missing values was largely ignored until the concept was formalized in the theory of Rubin (1976), through the simply devise of treating the missing-data indicators as a random variables and assigning them a distribution.

The estimation of causal effects in presence of missing data is the thread running through this thesis. However, chapters are designed to be independently readable. This implies the unavoidable pitfall that some definitions and/or concepts can be repeated in different parts of the work.

Each chapter opens with an introduction section, where we explain the problem, the motivation of the study and the approach we use, and describe the structure of the chapter itself. Conclusions, results and potential extensions are discussed at the end of each chapter. Here, we will just outline the topics that will be covered in the work and the results we have obtained.

In the first chapter, we provide an overview of the statistical approach to the estimation of causal effects based on the concept of potential outcomes, often referred to as "Rubin's Causal Model" (RCM, Holland, 1986). Our main objective in this chapter is to set out the basic concepts of statistical inference in experiments and observational studies, so there are no original results nor methodological elaborations.

We introduce the key notions underlying the Rubin Causal model potential outcomes framework, with a particular emphasis on the central role of the assignment mechanism in inferring causal effects. Special sections are dedicate to randomized assignments and to the concept of propensity score matching as a link between observational studies and randomized experiments. Then, we move to examine how to analyze studies suffering from missing background and/or outcome data, with a particular focus on the assumptions implicit or explicit in the existing approaches to the problem.

In the second chapter we present an extended framework for the analysis of data from randomized experiments which suffer from complications due to treatment noncompliance, missing outcomes following treatment noncompliance, and "truncation by death". The original motivation for our work was 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 noncompliance with the randomly assigned treatment and missing outcome data. In addition, the quality outcome is "truncated by death" in the sense that quality can only be observed for women who practice BSE, and it is not only unobserved but also undefined for women who do not practice self exams.

In recent years, there has been substantial progress in the analysis of randomized experiments suffering from noncompliance and missing data. These complications can represent a threat to valid estimates of experimental effects. Concerning the problem of data truncated by death, recent work has shown that traditional approaches, which address this issue ignoring the fact that the outcome after truncation is neither censored nor missing, but should be treated as being defined on an extended sample space, do not lead to properly defined causal estimands (Rubin, 2000; Frangakis and Rubin, 2002).

This paper develops a model that accommodates all these complications, which is based on the general framework of "principal stratification" (Frangakis and Rubin, 2002), and thus relies on more plausible assumptions than standard methodology. Our analysis revealed a positive, even if not highly significant, effect on quality of self exams for women who always comply with their assignment and would practice BSE under both treatment arms. All the analyses in this paper were implemented using R-project.

In the last paper we analyze the impact of childbearing events on individuals' wellbeing in Indonesia, using a sample of women drawn from the Indonesia Family Life Survey (IFLS). We consider the impact of having children on wellbeing as a quasi-experimental problem. The main issue in this approach is that subjects who experience childbearing events might somewhat be self-selected. Researcher has no control over treatment assignment. As a result, large differences can exist between the treatment and control groups on observed covariates, which can lead to badly biased estimates effects. Propensity score methods are an increasingly popular method for balancing the distribution of the covariates in the two groups to reduce this bias. To estimate propensity scores, which are the conditional probabilities of being treated given a vector of observed covariates, we must model the distribution of the treatment indicator given these observed covariates. Much work has been done in the case where the covariate are fully observed. However, in our study, some covariates values are missing. In such a case, which commonly arises in practice, it is not clear how the propensity score should be estimated. Any technique will have to either make a strong assumption regarding ignorability of the assignment mechanism or will have to make an assumption about the missing data mechanism.

In the paper three approaches for estimating propensity scores with incomplete data are presented: a complete-case analysis, a multiple imputation approach, and a pattern-mixture model based approach, with a discussion of the assumptions implicit in each of them. For each approach, we use the resulting propensity scores to construct comparison groups to the group of treated subjects, and then we estimate the causal effect of interest, and compare the results (we use STATA 8.0 to implement our analysis). All the three approaches show some evidence that childbearing events has a negative impact on individuals' wellbeing, which gives more weight to this conclusion.

# Bibliography

Angrist J. D., Imbens, G. W. (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, **90**, 431–442.

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.

Bjorklund, A., and Moffitt, R. (1987) Estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics*, **69**, 42–49.

Fisher, R. A. (1918) The causes of human variability. *Eugenics Review*, **10**, 213–220.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*, 1st Edition. Oliver and Boyd, Edinburgh.

Frangakis, C. E., and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

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

Manski, C. F. (1990) Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, **80**, 319–323.

Manski, C. F., Sandefur, G. D., McLanahan, S., and Powers, D. (1992) Alternative estimates of the effects of family structure during adolescence on high school graduation. *Journal of the American Statistical Association*, **87**, 25–37.

Neyman, J. (1923) On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in *Statistical Science*, **5**, 465–480, 1990.

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. (1977) Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1–26.

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

Rubin, D. B. (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**, 472–480.

Rubin, D. B. (1996) Statistical Inference for Causal Effects in Epidemiological Studies Via Potential Outcomes. *Proceedings of the XL Scientific Meeting of the Italian Statistical Society*, Florence, Italy, 419–430.

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

# Contents

# Acknowledgements

There are lots of people I would like to thank for a huge variety of reasons.

Firstly, I am very grateful to my Supervisor, Professor Fabrizia Mealli, for her many suggestions and constant support during this research. With her commonsense, knowledge, and perceptiveness, she has been an excellent guidance. Throughout my thesis-writing period, she provided encouragement, sound advice, good teaching, and lots of good ideas.

A preliminary paper on the randomized trial of Breast Self-Examination (BSE) was presented as poster presentation at the 19th International Workshop on Statistical Modeling in Florence. Special thanks go to Hirano Keisuke for his help in computational matters.

The paper on BSE was completed when I was visiting the Department of Statistics at Harvard University where I spent a wonderful period. I would like to thank all the staff for providing me a warm hospitality and a very stimulating research environment. Professor Donald Rubin deserves special thanks for his insightful suggestions in the implementation of the principal stratification approach.

Concerning the IFLS paper, I would like to thank Letizia Mencarini, for her suggestions in demographic area and Jungho Kim for providing me valuable support in the analysis of the IFLS data. A special thank goes to Arnstein Aassve for his insightful suggestions and discussions and for allowing me to exploit and modify its STATA source codes. I would also like to thank Professor Steve Pudney for the opportunity he gave me to spend a very stimulating research week at the Institute for Social and Economic Research at University of Essex.

Last but not least, I want to thank Marco and my family, in particular my Mum and my Sister, who always encouraged and supported me in the hardest times.

# Statistical Inference for Causal Effects in Randomized and Nonrandomized Experiments: A Gentle Introduction

## Abstract

This paper provides a brief overview of the statistical approach to the estimation of causal effects based on the concept of potential outcomes, now often referred to as Rubin's Causal Model (Holland, 1986). We focus on the treatment effect estimation in randomized and nonrandomized studies suffering form missing data. Placing the problem within the framework of the Rubin Causal Model makes the assumptions explicit by illustrating the interaction between the treatment assignment and the missing data mechanism, and the potential interaction between response behavior and other possible complications such as noncompliance in randomized experiments.

KEYWORDS: Causal inference, Potential Outcomes, Assignment Mechanism, Randomization, Propensity Score, Matching, Principal Stratification, Missing data, Rubin Causal Model, Structural Equation Models, Direct Acyclic Graphs.

## 1 Introduction

Decision in medicine, public health, and social policy depend critically on appropriate evaluation of competing treatments and policies. The extraction of information about such comparison, which we can broadly view as causal inference, has a growing area of statistical research in recent years. A statistical framework for causal inference that has received especially increasing attention is the one based on potential outcomes, originally introduced by Neyman (1923) for randomized experiments and randomization-based inference and generalized and extended by Rubin (1974, 1977, 1978) for nonrandomized studies and alternative forms of inference. Fundamentally, in this framework, often termed Rubin's Causal Model (Holland, 1986), a unit is considered at a particular place and time; treatments are interventions each of which can be potentially applied to each unit; potential outcomes are all the outcomes that would be observed when each of the treatments would be applied to each of the units. Then a causal comparison between, say, two treatments is a comparison of the potential outcomes of the same group of units under the two treatment conditions.

A major difference between the potential outcomes and the other frameworks for causal inference (e.g., simultaneous equations; Goldberger, 1972; Heckman, 1978) is that, in the former, the definition of causal effects is separated from any probability models about the way in which units are assigned to treatments, namely the assignment mechanism (Rubin, 1978), and this separation is regarded broadly (though not uniformly; cf. Dawid, 2000) as useful. This clarifying role of potential outcomes has been important in research, including, e.g., the earlier works on the concept of ignorable assignment (Rubin, 1974, 1977, 1978), propensity scores (Rosenbaum and Rubin, 1983a), the concept of sequential ignorability and associated methods (Rubin, 1978; Robins, 1986), and others. More recently, methods also became available to address treatment noncompliance using potential outcomes, starting mainly with work by Baker and Lindeman (1994), Imben and Rubin (1994), Robins and Greenland (1994), Angrist, Imbens, and Rubin (1996), and are currently receiving even more attention (e.g., Frangakis and Rubin, 1999; Hirano et al., 2000).

This paper provides a brief overview of the statistical approach to the estimation of causal effects based on the concept of potential outcomes, with a particular emphasis on the use of this framework in studies suffering from partially missing data.

Section 2 sets out the basic framework for causal inference based on the Rubin Causal Model (RCM). We discuss three key notions underlaying our approach. The first is that of potential outcomes corresponding to the various levels of a treatment. Each of these outcomes would have been observed had the treatment level been the corresponding one, even though after the treatment is received only one of them can be observed. Second, we discuss the necessity, when drawing causal inference, of observing multiple units, and the related stability assumption to exploit the presence of multiple units. Finally, we discuss the central role of the assignment mechanism, which is crucial for inferring causal effects. We then specify the benefits of randomization in estimating the causal effects of treatments and describe how we can use randomized experiments as a template for the analysis of observational studies. Specifically, we focus on the propensity score methodologies and the underlying assumptions. Section 3 introduces the concept of principal effects based on principal stratification, a general framework for comparing treatments where the estimands are adjusted for posttreatment variables and yet are always causal effects. In sections 4 and 5 we briefly review the existing approaches and their assumptions to analyze randomized and nonrandomized studies suffering from missing data. Finally, section 6 gives an idea of the vivid debate on causation existing in the academic community, and section 7 concludes.

## 2 Rubin Causal Model

The research questions that motivate most studies in statistics-based sciences are causal in nature. For instance, potentially interesting questions concern the efficacy of a given drug in a given population, the role of education in employment and earning, the fraction of crimes that could have been avoided by an education policy. Not surprisingly, the central target of such studies is the elucidation of causal-effect relationships among variables of interest, for example, treatments, exposures, preconditions, and outcomes. Yet until very recently, the dominant methodology has been based almost exclusively on statistical analysis which, traditionally, has excluded causal vocabulary both from its mathematical language and from its mainstream educational program.

For causal inference, there are several primitives, concepts that are basic and on which we must build. The fundamental notion underlying our view of causality is tied to an action or treatment applied to a unit. A unit is a physical object, firm, or person, or collection of persons such as a classroom or a market, at a particular point in time. A treatment is an action that can be applied or withheld from that unit. Associated with each unit and each treatment there are two potential outcomes, the values of an outcome variable $Y$ when the treatment is applied and when it is withheld. The objective is to learn about the causal effect of the application of the treatment relative to its being withheld on the outcome.

Formally, let $Z$ indicate which treatment the unit received: $Z = T$ the active treatment, $Z = C$ the control treatment. In order to avoid unnecessary complications, we assume that the treatment indicator is dichotomous, however, the framework is immediately applicable to multilevel treatments. Also let $Y(T)$ be the value of the potential outcome if the unit received the active version, and $Y(C)$ the value if the unit received the control version. The causal effect of the active treatment relative to its control version is defined as a comparison of $Y(T)$ and $Y(C)$. Basically, the absolute difference between treatments, measuring $Y(T) - Y(C)$, as well as the relative difference, measuring $Y(T)/Y(C)$, can be compared. Here, the focus will be on absolute difference. However, this does not affect the generality of the results.

There are two important aspects of this definition of causal effects. First, the definition does not depend on which outcome is actually measured. Second, the causal effect is the comparison of outcomes at the same moment of time, where the time of the application of the treatment precedes that of the outcome.

The fundamental problem of causal inference is that, for any individual unit, we observe the value of the potential outcome under only one of the possible treatments, namely the treatment actually assigned, and the potential outcome under the other treatment is missing. Potential outcomes $(Y(C), Y(T))$ and assignment $Z$ jointly determine the values of the observed and missing outcomes through the two relations

$$\begin{aligned} Y^{\text{obs}} &= Y(Z) = && \mathrm{I}\{Z = T\} \cdot Y(T) + (1 - \mathrm{I}\{Z = T\}) \cdot Y(C) \\ Y^{\text{mis}} &= Y(Z^c) = && (1 - \mathrm{I}\{Z = T\}) \cdot Y(T) + \mathrm{I}\{Z = T\} \cdot Y(C), \end{aligned}$$

where $Z^c = C$ if $Z = T$, and $Z^c = T$ if $Z = C$, and $\mathrm{I}\{\cdot\}$ is the indicator function. Thus, under this straightforward perspective, inference for causal effects is a missing-data problem. This implies that we cannot learn from just the single observed outcome what the causal effect of the treatment is, because the causal effect involves the comparison of both potential outcomes. To learn about causal effects we must rely on multiple units. More specifically, we must observed units exposed to different treatments.

One option is to observe the same physical object at different points in time, that is, we might have the same unit measured on both treatments in two trials (a repeated measure design), but since there may exist carryover effects (e.g., the effect of the first treatment wears off slowly) or general time trend we cannot be certain that the unit's responses would be identical at both times. Therefore, a unit at a different time is, in general, a different unit. As an alternative to observing the same physical unit repeatedly, one might observe different physical units at similar times. Hence, assume that there are $N$ units for which we want to assess the causal effect.

By itself, the presence of multiple units does not solve the problem of causal inference. For instance, suppose we have two units. Now in general we have at least four potential outcomes for each unit; they are the values of the outcome variable when the treatment is applied or when it is withheld for both of the units. Specifically, for unit 1 we would have the potential outcomes $Y_1(C, C)$, $Y_1(C, T)$, $Y_1(T, C)$, and $Y_1(T, T)$, where, for example, $Y_1(T, C)$ is the outcome for unit 1 if unit 1 received active and unit 2 received control; and analogously for unit 2. In fact, there can be even more potential outcomes depending on the number of versions of the treatment.

Therefore, replication does not help unless we can restrict the explosion of potential outcomes. The most straightforward assumption is the “Stable Unit Treatment Value Assumption” (SUTVA - Rubin, 1980a, 1990) under which the potential outcomes for the $i$th unit just depend on the treatment that the $i$th unit received. That is, there is “no interference between units” and there are “no versions of treatments”. Then, all potential outcomes for the $N$ units can be represented by an array with $N$ rows and two column, each unit $i$ being a row with two potential outcomes, $Y_i(C)$ and $Y_i(T)$. Throughout this paper we make the Stable Unit Treatment Assumption. Recall that there is no assumption-free causal inference, and nothing is wrong with this. It is the quality of the assumptions that matters, not their existence.

Under SUTVA, an obvious definition of the causal effect of the $T$ versus $C$ treatment for the $N$ trials in the study is the average causal effect for the $N$ trials:

$$\tau = \mathrm{E}\big(Y(T) - Y(C)\big) = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(T) - Y_i(C)\big).$$

Even though other definitions can be interesting, we assume the average causal effect is the desired causal effect and proceed to the problem of its estimation given the obvious constraint that we can never actually measure both $Y_i(T)$ and $Y_i(C)$ for any unit.

In addition to the vector indicator of treatments, $\mathbf{Z} = \{Z_i : i = 1, \ldots, N\}$, the vector of potential outcomes when exposed to the active treatment, $\mathbf{Y}(T) = \{Y_i(T) : i = 1, \ldots, N\}$, and the vector of potential outcomes when not exposed $\mathbf{Y}(C) = \{Y_i(C) : i = 1, \ldots, N\}$, we suppose to have a $N \times K$ matrix of covariates, $\mathbf{X}$, with $i$th row equal to $\mathbf{X}_i = (X_{1i}, \ldots, X_{iK})$, a $K$-vector of background variables which encodes characteristics of unit $i$. As we will see, covariates, which are unaffected by treatment, play a particular important role in causal inference, above all in studies where the units exposed to the active treatment can differ on their distribution of covariates in important ways from the units not exposed.

## 2.1 The Assignment Mechanism

In the potential outcomes framework for causal inference, a key role is played by the assignment mechanism, that is, the mechanism that determines which units get which treatment. Formally, we define the assignment mechanism as a function assigning probabilities to all possible $N$-vectors of binary assignment $\mathbf{Z}$ given the $N$-vectors of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(T)$ and the $N \times K$ matrix of covariates $\mathbf{X}$:

**Definition 2.1.** *Given a population of $N$ units, the assignment mechanism is a row-exchangeable function* $\Pr(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(C), \boldsymbol{Y}(T))$ *taking on values in* $\{0,1\}^N$ *satisfying*

$$\sum_{\boldsymbol{Z}} \Pr(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{Y}(C), \boldsymbol{Y}(T)) = 1,$$

*for all* $\boldsymbol{X}$*,* $\boldsymbol{Y}(0)$*, and* $\boldsymbol{Y}(T)$*.*

An example of an assignment mechanism is a randomized experiment. It is an assignment mechanism such that $(i)$ it is ignorable, which means that it does not depend on the missing outcomes; $(ii)$ it is probabilistic, that is, $0 < \Pr(Z_i = T \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) < 1$ for all $i$, and for all $\mathbf{X}$, $\mathbf{Y}(0)$, and $\mathbf{Y}(T)$, where $\Pr(Z_i = T \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = \sum_{\mathbf{Z}: Z_i = T} \Pr(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T))$ is the unit assignment probability for unit $i$;[1] and $(iii)$ it is a known function of its arguments. Here, we will mainly be concerned with a special case of randomized experiments, classical randomized experiments, which in addition to the conditions required for randomized experiments assume local independence. This assumption requires the assignment mechanism to be separable in the unit assignment probabilities, at least conditional on $(\mathbf{Z}, \mathbf{X})$. Moreover, it requires the unit assignment probability for unit $i$ to be a function of outcomes and covariates for unit $i$ only, free of the values of outcomes

[1] Note that the row exchangeability of the assignment mechanism implies that $\Pr(Z_i = T \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = \Pr(Z_i = T \mid \mathbf{X}_\sigma, \mathbf{Y}_\sigma(C), \mathbf{Y}_\sigma(T))$ for each permutation $\sigma$ of the row indices $1, \ldots, N$.

and covariates for other units other than through the dependence of the joint assignment probabilities on these outcomes. Formally, an assignment mechanism is locally independent if

$$\begin{aligned}\Pr\big(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)\big) &= \\ &g\big(\mathbf{Z}, \mathbf{X}\big) \prod_{i=1}^{N} \Big(\Pr(Z_i = T \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T))\Big)^{\mathrm{I}\{Z_i=T\}} \\ &\qquad \times \Big(1 - \Pr(Z_i = T \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T))\Big)^{(1-\mathrm{I}\{Z_i=T\})}\end{aligned}$$

and

$$\Pr\big(Z_i = T \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)\big) = \Pr\big(Z_i = T \mid \mathbf{X}_i, Y_i(C), Y_i(T)\big) \quad \text{for all } i,$$

where $g(\mathbf{Z}, \mathbf{X})$ must be such that $\sum_{\mathbf{Z}} \Pr(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = 1$. An example of such assignment mechanisms is a Completely randomized experiment where $M$ out of $N$ units are randomly chosen to receive the treatment. For this assignment mechanism

$$\Pr(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = \begin{cases} \left(\frac{M}{N}\right)^{\mathrm{I}\{Z_i=T\}} \left(\frac{N-M}{N}\right)^{\mathrm{I}\{Z_i=C\}} & \text{if } \sum \mathrm{I}\{Z_i = T\} = M, \\ 0 & \text{otherwise.} \end{cases}$$

Most often $M = N/2$, so that half the units receive the active treatment and half receive the control treatment.

Being ignorable and locally independent, a classical randomized experiment is also unconfounded, that is, it does not depend on the potential outcomes:

$$\Pr(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(C), \mathbf{Y}(T)) = \Pr(\mathbf{Z} \mid \mathbf{X}).$$

With an unconfounded assignment mechanism, at each set of values of $\mathbf{X}_i$ that has a distinct probability of $Z_i = T$, there is effectively a randomized experiment.

The assignment mechanism is fundamental to causal inference because it tells us how we got to see what we saw. Causal inference is basically a missing data problem because at least half of the potential outcomes are not observed, and so missing. Without understanding the process that creates missing data, we have no hope of inferring anything about them. Without a model for how treatments are assigned to individuals, formal causal inference, at least using probabilistic statements, is impossible. This does not mean that we need to know the assignment mechanism, but rather that without positing one, we cannot make any statistical claims about causal effects, such as unbiasedness of estimates, confidence coverage of intervals for effects, significance levels of tests, or coverage of Bayesian posterior interval.

## 2.2 Estimating Causal Effects of Treatments in Randomized Experiments

Randomization is an assignment mechanism that allows particulary straightforward estimation of causal effects. Therefore, simple randomized experiments form the basis for inference for causal effects in more complicated situations, such as when the assignment probabilities depend on covariates or when there is noncompliance with the assignment mechanism. In addition, an unconfounded assignment mechanism, which essentially is a set of randomized experiments, forms the basis for the analysis of an observational nonrandomized study by using the randomized experiment as a template.

The central question of this section concerns the benefits of randomization in determining the causal effect of the active versus control treatment on an outcome $Y$. Therefore, suppose that a randomized experiment with $N$ trials has been performed to estimate the typical causal effect of the active versus control treatment on $Y$ for some population of units. For simplicity, suppose that no pretreatment covariates are recorded.

Randomization can never assure us that we are correctly estimating the causal effect of a treatment versus another for the $N$ trials under study, but it provides important benefits besides the intuitive ones that follow from making all systematic source of bias into random ones. Formally, randomization provides a mechanism to derive probabilistic properties of estimates without making other assumptions. We will consider two such properties that are important: (1) the average difference between the treatment and control group is an unbiased estimate of $\tau$, the typical causal effect for the $N$ units in the study, defined in section 2; and (2) precise probabilistic statements can be made indicating how unusual the observed average difference between the treatment and control group would be under specific hypothesized causal effects.

Let $\hat{\tau}$ be the observed average difference between the treatment and control group:

$$\hat{\tau} = \overline{Y}_T^{\text{obs}} - \overline{Y}_C^{\text{obs}} = \frac{\sum_{i=1}^N \mathrm{I}\{Z_i = T\} Y_i^{\text{obs}}}{\sum_{i=1}^N \mathrm{I}\{Z_i = T\}} - \frac{\sum_{i=1}^N \mathrm{I}\{Z_i = C\} Y_i^{\text{obs}}}{\sum_{i=1}^N \mathrm{I}\{Z_i = C\}}.$$

This is an unbiased estimator for the typical causal effect $\tau$ over the randomization set.

To show this, first we define the randomization set to be the set of $r$ allocations that were equally likely to be observed given the randomization plan. For instance, in a completely randomized experiment with $M < N$ units assigned to treatment, the randomization set is the collection of $r = \binom{N}{M}$ equally likely possible allocations.

For each of the $r$ possible allocations in the randomization set, there is a corresponding average difference $\hat{\tau}$ that would be calculated had that allocation been chosen. If the expectation of these $r$ possible differences equals $\tau$, the average difference $\hat{\tau}$ is called unbiased over the randomization set for estimating $\tau$. We now show that given randomly assigned treatments, the average difference $\hat{\tau}$ is an unbiased estimate of $\tau$, the typical causal effect for the $N$ units.

By definition of random assignment each unit has a known probability of receiving the active treatment, here assumed constant and equal to $p$. Hence, the contribution of the $i$th unit ($i = 1, \ldots, N$) to the average difference $\hat{\tau}$ in $p$ of the $r$ allocations in the randomization set is $Y_i(T)/(Np)$ and in the other $(1-p)$ is $-Y_i(C)/((1-p)N)$. The expected contribution of the $i$th unit to the average difference $\hat{\tau}$ is therefore

$$p\frac{Y_i(T)}{pN} + (1-p)\frac{-Y_i(C)}{(1-p)N}.$$

Summing over all $N$ units we have the expectation of the average difference $\hat{\tau}$ over the $r$ allocations in the randomization set is

$$\frac{1}{N}\sum_{i=1}^{N}\big(Y_i(T)-Y_i(C)\big),$$

which is the typical causal effect for the $N$ units in the trial, $\tau$.

The unbiasedness of the $\hat{\tau}$ estimator for $\tau$, that follows from the random assignment of treatments, is a desirably property because it indicates that on average we tend to estimate the correct quantity, however it hardly solves the problem of estimating the typical causal effect. As yet we have no indication whether to believe $\hat{\tau}$ is close to $\tau$ nor to any ability to adjust for important information we may possess.

Consider now the other formal advantage of randomization. We show that randomization provides a mechanism for making probabilistic statements indicating how unusual the observed difference $\hat{\tau}$ would be under specific hypotheses.

Suppose that the researcher hypothesizes exactly what the individual causal effects are for each of the $N$ units and these hypothesized values are $\tilde{\tau}_i$, $i = 1, \ldots, N$. The hypothesized typical causal effect for the $N$ units is thus

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\tau}_i.$$

Having the $\tilde{\tau}_i$ and the observed $Y_i(T)$, $i \in \{i : Z_i = T\}$ and $Y_i(C)$, $i \in \{i : Z_i = C\}$, we can easily calculate hypothesized values, say $\tilde{Y}_i(C)$ and $\tilde{Y}_i(T)$ for all the $N$ units, and using these, we can calculate an hypothesized average difference between the treatment and control group for each of the $r$ allocations of the $N$ units in the randomization set.

Since the average of the $r$ average differences between the treatment and control group is the hypothesized typical causal effect, $\tilde{\tau}$, and the $r$ allocations are equally likely, we can make the following probabilistic statement:

> Under the hypothesis that the causal effects are given by the $\tilde{\tau}_i$, $i = 1, \ldots, N$, the probability that we would observe an average difference between the treatment and control group that is as far or farther from $\tilde{\tau}$

> than one we have observed is $h/r$, where $h$ is the number of allocations in the randomization set that yield average differences between the treatment and control group that are as far or farther from $\tilde{\tau}$ than $\hat{\tau}$ (Rubin, 1974).

If this probability - called the *significance level* for the hypothesized $\tilde{\tau}_i$ - is very small, we either must admit that the observed value is unusual in the sense that it is in the tail of the distribution of the equally likely differences, or we must reject the plausibility of the hypothesized $\tilde{\tau}_i$ .

The ability to make precise probabilistic statements about the observed $\hat{\tau}$ under various hypotheses without additional assumptions is a tremendous benefit of randomization especially since $\hat{\tau}$ tends to estimate $\tau$. However, one must realize that these simple probabilistic statements refer only to the $N$ trials used in the study and do not reflect additional information that we may also have measured.

In order to make an intelligent adjustment for extra information, we cannot be guided only by the concept of unbiasedness over the randomization set. We need some model for the effect of prior variable in order to use their value in intelligent manner. The point of this statement is that when trying to estimate the typical causal effect in the $N$ trial experiment, handling additional variables may not be trivial without a well-developed causal model that will properly adjust for those prior variables that causally affect $Y$ and ignore other variables that do not causally affect $Y$ even if they are highly correlated with the observed values of $Y$. Without such a model, the researcher must be prepared to ignore some variables he feels cannot affect $Y$ and use a somewhat arbitrary model to adjust for those variables that he fells are important.

A researcher should also believe the results of an experiment are applicable to a population of units besides the $N$ in the experiment. Even though the trials in an experiment are often not very representative of the trials of interest, researchers must be willing to make this assumption - called “subjective random sampling assumption” (Rubin, 1974) - in order to believe their results are useful.

## 2.3 Estimating Causal Effects of Treatments in Observational Studies Using Propensity Score Methods

The same two issues discussed at the end of previous section as arising when presenting results of an experiment also arise when presenting the results of a nonrandomized study as being relevant. However, the first issue, the effect of variables not explicitly controlled is usually more serious in nonrandomized than in randomized studies, while the second, the applicability of the results to a population of interest is often more serious in randomized than in nonrandomized studies.

In randomized experiments, the results in the two treatment groups may often be directly compared because their units are likely to be similar, whereas in nonrandomized experiments, such direct comparisons may be misleading because the units exposed to one treatment generally differ systematically from the units exposed to the other treatment. Specifically, whereas in experimental situations one can obtain a control and treatment group which are homogeneous with respect to the observable characteristics, $\mathbf{X}$, this is not possible in nonexperimental studies since it is likely that the decision to be assigned to a treatment is in this case not independent from the observable as well as unobservable characteristics.[2] A possible way to address this complication in nonexperimental studies is to consider the randomized experiment as a template for the analysis of an observational (i.e., nonrandomized) study. Having the template of a randomized experiment means having to think about the underlying randomized experiment that could have been done, where in the randomized experiment underlying an observational study, the probabilities of assignment to treatments are not equal, but are rather functions of the covariates, and so the template is actually an unconfounded assignment mechanism.

[2] With random assignment, homogeneity of the control and treatment group with respect to the unobservable characteristics is also guaranteed if the size of the groups is sufficiently large.

To do this we make the *strong ignorability* or *unconfoundedness* assumption. Generally, we shall say treatment assignment is strongly ignorable given a vector of covariates $\mathbf{W}$ if

$$(Y(C), Y(T)) \perp Z \mid \mathbf{W} \quad \text{and} \quad 0 < \Pr(Z = T \mid \mathbf{W}) < 1, \tag{2.1}$$

for all $\mathbf{W}$. For brevity, when treatment assignment is strongly ignorable given the observed covariates $\mathbf{X}$, that is, when (2.1) holds with $\mathbf{W} = \mathbf{X}$, we shall say simply that treatment assignment is strongly ignorable.

The strong ignorability assumption asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates. In other words, within subpopulations defined by values of the covariates, we have random assignment. This assumption rules out the role of the unobservable variables. The issue of unobserved covariates should be addressed using models for sensitivity analysis (e.g., Rosenbaum and Rubin, 1983b) or using non parametric bounds for treatment effects (Manski, 1990; Manski et al., 1992).

Clearly, the strong ignorability assumption may be controversial. It requires that all variables that affect both outcomes and the likelihood of receiving the treatment are observed. Although this is not testable, it clearly is a very strong assumption, and one that need not generally be applicable. We view it as a useful starting point for two reason. One is that in some studies, as the Connors et al. (1996) study of right heart catherization, researchers have carefully investigated which variables are most likely to confound any comparison between treated and control units and made attempts to observe all such variables. Even if these attempts are not completely successful the assumption that all relevant variables are observed may be a reasonable approximation, especially if much information about pretreatment outcomes is available. Second, any alternative assumption that does not rely on unconfoundedness while allowing for consistent estimation of the average treatment effects must make alternative untestable assumptions, such as the instrumental variable technique (e.g., Angrist, 1990; Angrist and Krueger, 1991). Whereas the unconfoundedness assumption implies that the best matches units that differ only in their treatment status, but otherwise are identical, alternative assumptions implicitly match units that differ in the pretreatment characteristics. Often such assumptions are even more difficult to justify. The unconfoundedness assumption therefore may be a natural starting point after comparing average outcomes for treated and control units to adjust for observable pretreatment differences.

The unconfoundedness assumption validates the comparison of treated and control units with the same value of covariates. The treatment effect for the subpopulation with $\mathbf{X} = \mathbf{x}$ can be written as:

$$\begin{aligned}
\tau(\mathbf{X}) &= \mathrm{E}\big(Y(T) - Y(C) \mid \mathbf{X} = \mathbf{x}\big) \\
&= \mathrm{E}\big(Y(T) \mid Z = T, \mathbf{X} = \mathbf{x}\big) - \mathrm{E}\big(Y(C) \mid Z = C, \mathbf{X} = \mathbf{x}\big) \\
&= \mathrm{E}\big(Y \mid Z = T, \mathbf{X} = \mathbf{x}\big) - \mathrm{E}\big(Y \mid Z = C, \mathbf{X} = \mathbf{x}\big),
\end{aligned}$$

where both term on the right-hand side can be estimated from a random sample of $(\mathbf{X}, Z, Y)$. The average treatment effect can be then estimated using the equality

$$\tau = \mathrm{E}\big(\tau(\mathbf{X})\big).$$

Typically, there are many background characteristics that need to be controlled for estimating the average causal effect $\tau$, and adjusting the estimation for all these covariates can be actually unfeasible. Propensity score technology, introduced by Rosenbaum and Rubin (1983a), addresses this situation by reducing the entire collection of background characteristics to a single "composite" characteristic that appropriately summarizes the collection. Formally, the propensity score is defined as the conditional probability of receiving a treatment given pretreatment characteristics:

$$e(\mathbf{X}) = \Pr\big(Z = T \mid \mathbf{X}\big).$$

The propensity score is a balancing score, that is, treatment assignment and observed covariates are conditionally independent given the propensity score:

$$\mathbf{X} \perp Z \mid e(\mathbf{X}). \tag{2.2}$$

In particular, the propensity score is the coarsest balancing score, i.e., any balancing score $b(\mathbf{X})$ must satisfy the relation $e(\mathbf{X}) = f(b(\mathbf{X}))$, for some function $f$ (Rosenbaum and Rubin, 1983a).

The key feature of propensity score methodology is that, given the strong ignorability assumption, treatment assignment and the potential outcomes are independent:

$$\big(Y(C), Y(T)\big) \perp Z \mid e(\mathbf{X}),$$

and

$$0 < \Pr\big(Z = T \mid e(\mathbf{X})\big) < 1.$$

Thus adjusting for the propensity score removes the bias associated with differences in the observed covariates in the treated and control group. As a result, given the strong ignorability assumption, if the propensity score $e(\mathbf{X})$ is known, it follows that

$$\begin{aligned}
\tau &= \mathrm{E}\big(Y(T) - Y(C)\big) \\
&= \mathrm{E}\Big(\mathrm{E}\big(Y(T) - Y(C) \mid e(\mathbf{X})\big)\Big) \\
&= \mathrm{E}\Big(\mathrm{E}\big(Y(T) \mid Z = T, e(\mathbf{X})\big) - \mathrm{E}\big(Y(C) \mid Z = C, e(\mathbf{X})\big)\Big),
\end{aligned}$$

where the outer expectation is over the distribution of $e(\mathbf{X})$.

The propensity score is a potential matching variable because it does not depend on response information that will be collected after matching. Since exact matching for a known propensity score will on average remove all the bias in $\mathbf{X}$, the propensity score $e(\mathbf{X})$ is in a sense the most important scalar matching variable.

Matching on $e(\mathbf{X})$ balances the observed covariates $\mathbf{X}$; however, unlike randomization, matching on $e(\mathbf{X})$ does not balance unobserved covariates except to the extend that they are correlated with $\mathbf{X}$: we need the strong ignorability assumption. For discussion of methods for addressing the possible effects of unobserved covariates in observational studies, see Rosenbaum and Rubin (1983b) and Rosenbaum (1984a).

In practice, several issues need to be addressed before the propensity score can be used as a matching variable. First, the functional form of $e(\mathbf{X})$ is rarely if ever known, and therefore $e(\mathbf{X})$ must be estimated from the available data. Second, exact matches will rarely be available, and so issues of closeness on $e(\mathbf{X})$ must be addressed. Third, adjustment for $e(\mathbf{X})$ balances $\mathbf{X}$ only in expectation, that is, averaging over repeated studies. In any particular study, further adjustment for $\mathbf{X}$ may be required to control chance imbalances in $\mathbf{X}$. Such adjustments, for example by covariance analysis, are often used in randomized experiment to control chance imbalances in observed covariates.

As noted by Rosenbaum and Rubin (1983a), matching on the propensity score is generalization to arbitrary $\mathbf{X}$ distributions of discriminant matching for multivariate normal $\mathbf{X}$ as proposed by Rubin (1970) and discussed by Cochran and Rubin (1973) and Rubin (1976a, 1976b, 1979, 1980b). Propensity matching is not, however, the same as any of the several procedures proposed by Miettinen (1976): the propensity score is not generally a confounder score (see Rosenbaum and Rubin, 1983a). First, the propensity score depends only on the joint distribution of $\mathbf{X}$ and $Z$, whereas a confounder score depends additionally on the conditional distribution of a discrete outcome variable given $\mathbf{X}$ and $Z$, and is not defined for continuous outcome variables. Second, the propensity score is the coarsest function of $\mathbf{X}$ that has the balancing property (2.2) (Rosenbaum and Rubin, 1983a), so unless a confounder score is finer than the propensity score, it will not have this balancing property.

To conclude we briefly describe how we can estimate and use propensity score methodology in practical applications.

First suppose to know the propensity score. In general, exact matches on propensity score is impossible to obtain, so methods which seek approximate matches must be used. Therefore, it can be useful to study properties of some matching methods based on the propensity score.

The mean bias or expected difference in $\mathbf{X}$ prior to matching is $\mathrm{E}(\mathbf{X} \mid Z = T) - \mathrm{E}(\mathbf{X} \mid Z = C)$, whereas the mean bias in $\mathbf{X}$ after matching is $\mathrm{E}(\mathbf{X} \mid Z = T) - \mathrm{E}_M(\mathbf{X} \mid Z = C)$, where $\mathrm{E}_M(\mathbf{X} \mid Z = C)$ is the expected value of $\mathbf{X}$ in the matched control group. Generally, $\mathrm{E}_M(\mathbf{X} \mid Z = C)$ depends on the matching method used, whereas $\mathrm{E}(\mathbf{X} \mid Z = T)$ and $\mathrm{E}(\mathbf{X} \mid Z = C)$ depend only on population characteristics. A matching method is equal-percent bias reducing (EPBR) if the reduction in bias is the same for each coordinate of $\mathbf{X}$, that is, if

$$\mathrm{E}(\mathbf{X} \mid Z = T) - \mathrm{E}_M(\mathbf{X} \mid Z = C) = \gamma\Big(\mathrm{E}(\mathbf{X} \mid Z = T) - \mathrm{E}(\mathbf{X} \mid Z = C)\Big)$$

for some scalar $0 \leq \gamma \leq 1$ (Rubin, 1976a, b). If a matching methods is not EPBR, then matching actually increases the bias for some linear functions of $\mathbf{X}$. If little is known about the relationship between $\mathbf{X}$ and the response variables that will be collected after matching, then EPBR matching methods are attractive, since they are the only methods that reduce bias in all variables having linear regression on $\mathbf{X}$. Rosenbaum and Rubin (1983a) showed that matching on the population propensity score alone is EPBR whenever $\mathbf{X}$ has a linear regression on some scalar function of $e$; that is, whenever $\mathrm{E}(\mathbf{X} \mid e) = \alpha + \gamma' g(e)$ for some scalar function $g(\cdot)$.

Matched samples can be constructed by using several different methods that matched treated units to control units.[3] Two standard techniques are the "nearest available matching on the propensity score" and "subclassification on the propensity score". Subclassification method consists of dividing experimental and control units on basis of $e(\mathbf{X})$ into subclasses or strata such that within each subclass treated and control units have on average the same propensity score. Then, within each stratum in which both treated and control units are present, the average outcomes of the treated and control units are compared. The average treatment effect of interest is finally obtained as an average of the subclass-specific comparisons. One of the pitfalls of the subclassification method is that it discards observations in strata where either treated or control units are absent. This observation suggests an alternative way to match treated and control units, which consists of taking each treated unit and searching for the control unit with the closest propensity score, i.e.,

[3] In many observational studies, there is a relatively small group of subject exposed to a treatment and a much larger group of control subjects not exposed. When the cost associated with obtaining outcome or response data from subjects are high, some sampling of the control reservoir is often necessary. Matched sampling attempts to choose the control for further study so that they are similar to the treated subjects with respect to background variables measured on all subjects.

the nearest available matching on the propensity score. Although it is not necessary, the method is usually applied with replacement, in the sense that a control unit can be a best match for more than one treated unit. Once each treated unit is matched with a control unit, the difference between the outcome of the treated units and the outcome of the matched units is computed. The average treatment effect of interest is then obtained by averaging these differences.

As noted previously, usually we do not actually know the propensity scores, and so we must estimate them. Propensity scores can be estimated in a number of different ways, including discriminant or CART analysis. In principle, any standard probability model can be used to estimate the propensity score. For instance, $\Pr(Z = T \mid \mathbf{X}) = F(h(\mathbf{X}))$, where $F(\cdot)$ is the normal or the logistic cumulative distribution function and $h(\mathbf{X})$ is a function of covariates with linear or higher order terms. It is critically important to note that the outcome variable plays no role in the estimation of the propensity score; estimating propensity scores only involves the covariates. Consequently, the success of the propensity score estimation must be assessed by the resultant balance of the observed distribution of covariates across the treated and control groups rather than by the fit of the models used to create estimated propensity scores.

Clearly, models for the data, $\Pr(\mathbf{X}, Y(0), Y(1))$, can be an important adjunct to propensity score methods, just as covariance adjustment can be an important adjunct in randomized experiments. There exists a large literature (e.g., Rubin, 1973, 1979; Reinisch et al., 1996; Rubin and Thomas, 2000) indicating the improved estimation that can take place when models are used to refine estimation. But it must be remembered that such modeling is a supplement to modeling the assignment mechanism, and is essentially adding a Bayesian component to the structure as in Rubin (1978). These Bayesian answers create procedures whose frequency performance can be evaluated. Such modeling becomes almost a necessary ingredient in complex situations involving, for example, randomized experiments with noncompliance (see section 5), or when investigating sensitivity to the unconfoundedness assumption.

## 3 Principal Stratification in Causal Inference

Many scientific problems require that treatment comparisons be adjusted for posttreatment variables, but the estimands underlying standard methods are not properly defined causal effects (see Rubin, 2000; and Frangakis and Rubin, 2002 for more discussion on this).

Suppose that after each unit $i$ ($i = 1, \ldots, N$) gets assigned treatment $Z_i$, a posttreatment variable $S_i^{\text{obs}}$ is measured, in addition to measuring the main outcome $Y$. The variable $S^{\text{obs}}$ encodes characteristics of the unit as well as of the treatment. Therefore, an important study goal can be to compare the effects of treatments on $Y$ "after adjusting" for the posttreatment characteristics in a way that the adjusted estimands are causal effects.

A standard method adjusts for the posttreatment variable using a comparison between outcomes under standard versus new treatment for subjects who got a common value $s$ of $S^{\text{obs}}$. The key to understanding such adjustments is to recognize that $S_i^{\text{obs}}$ is $S_i(Z_i)$, i.e., the observed value of one of the two potential values $S_i(C)$, $S_i(T)$, depending on treatment assignment. This comparison is problematic if the treatment has any effect on the posttreatment variable (Rosenbaum, 1984b) because the groups $\{i : S_i(C) = s\}$ (i.e., who get posttreatment value $s$ under standard treatment) and $\{i : S_i(T) = s\}$ (i.e., who get posttreatment value $s$ under active treatment) are not the same groups of subjects. In our view of causality, the causal effect of the assignment on the outcome $Y$ is defined to be a comparison between the ordered sets of potential outcomes on a common set of units. Then, according to this definition, comparing outcomes of the two groups $\{i : S_i(C) = s\}$ and $\{i : S_i(T) = s\}$ does not lead to a properly defined causal estimand.

To address this deficiency, Frangakis and Rubin (2002) proposed a general framework for comparing treatments adjusting for posttreatment variables that yields principal effects based on principal stratification.

**Definition 3.1.** *(a) The basic principal stratification $P_0$ with respect to posttreatment variable $S$ is the partition of units $i = 1, \ldots, N$ such that, within any set of $P_0$, all units have the same vector $(S_i(C), S_i(T))$.*
*(b) A principal stratification $P$ with respect to posttreatment variable $S$ is a partition of the units whose sets are unions of sets in the basic principal stratification $P_0$.*

Generally, a principal stratification generates the following estimands.

**Definition 3.2.** *Let $P$ be a principal stratification with respect to the posttreatment variable $S$ and let $S_i^P$ indicate the stratum of $P$ to which unit $i$ belongs. Then a principal effect with respect to that principal stratification is defined as a comparison of potential outcomes under standard versus new treatment within a principal stratum $\zeta$ in $P$, i.e., a comparison between the ordered sets*

$$\{Y_i(C) : S_i^P = \zeta\} \quad \textit{and} \quad \{Y_i(T) : S_i^P = \zeta\}.$$

The importance of principal effects draws from their conditioning on principal strata. Although the potential variable $S_i(C)$ generally differs from $S_i(T)$, the value of the ordered pair $(S_i(C), S_i(T))$ is, by definition, not affected by treatment. As a result, the central property of principal effects is that they are always causal effects and do not suffer from complications of standard posttreatment-adjusted estimands.

Setting principal causal effects to be the goal helps focus the role of inference. Generally, we cannot directly observe the principal stratum to which a subject belongs because we cannot directly observe both $S_i(C)$ and $S_i(T)$ for any $i$; therefore inference about the principal effects, e.g., in $P_0$, requires prediction of the subjects' missing memberships to the principal strata, as determined by $S^{\text{mis}} = S(Z^c) = (1 - \mathrm{I}\{Z = T\}) \cdot S(T) + \mathrm{I}\{Z = T\} \cdot S(C)$, as well as prediction of the subjects' missing potential outcomes $Y^{\text{mis}}$.

Principal stratification is a powerful framework which allows to address possible complications in a study - such as the presence of missing outcomes - better then other standard methods. In the next sections, we focus on the problem of missing data showing to what extent this complication can affect a study and how it can be addressed. Assignment mechanism and principal stratification are the two key concepts on which our discussion relies.

# 4 Estimating Causal Effects with Missing Background Data

Covariates are variables whose values are not affected by the treatment assignment.

In randomized experiments in which treatment assignment is independent of covariates, these can still be used to improve efficiency of estimation. Specifically, covariates improve prediction of missing potential posttreatment outcome variables, that is, outcome values under treatment for those who are assigned to control, and outcome values under control for those who are assigned to treatment.

Covariate also enhance generalizability of the experimental results. Compared to marginal relationships (such as the mean value of the outcome for the experimental group) the relation between the outcome values and the covariates is more likely to generalize. In other words, we can always recover marginal relationships from conditional relationships, but not vice versa. Therefore, conditional relationships extract more information from the data.

In nonrandomized studies the decision to be assigned to a treatment is not generally independent from the observable characteristics, and so it is likely that units exposed to one treatment differ systematically from units exposed to the other treatment. However, if we assume the strong ignorability assumption of the assignment mechanism, we can attempt to recreate the ideal situation of the randomized experiments where covariates are balanced across treatment group. In fact if the assignment mechanism is strongly ignorable, the observed covariates are sufficient to explain why people chose one treatment or another. As noted previously, the plausibility of this assumption rests on the amount of information contained in the covariates, $\mathbf{X}$, so often the higher the dimension of $\mathbf{X}$, that is, roughly, the greater is the number of covariates in $\mathbf{X}$, the more plausible we might consider the assumption to be. Therefore, covariates play a particularly important role in nonexperimental studies.

Much work has been done in the case where covariate are fully observed (e.g., Hirano et al., 2000; Mealli et al., 2004; Rubin and Thomas, 1992a,b, 1996). In practice, however, some covariate values will be missing.

In randomized studies, missingness of background variables occurs before randomization. In principle, such missingness is also a covariate and so does not directly create imbalance of subjects between randomized arms, although it does create loss in efficiency when background covariates are to be used in the analysis. When there is insight that missingness of pretreatment variables can affect heavily the results, Barnard et al. (2003) suggested to include in the analysis information on the patterns of missing background data. Imbens and Pizer (1999) suggested that with missing covariates, the standard approach (Rubin, 1976c; Little and Rubin, 1987), which assumes that, conditional on treatment and any fully-observed covariates, the data are Missing At Random (MAR) or, alternatively, the missing data process is ignorable, is not necessarily adequate to describe the data. The reason is that the two assumptions (*i*) random assignment of treatment, and (*ii*) missing at random, have implications that can be in conflict. Specifically, Imbens and Pizer (1999) noted that, if we observe that among complete-data observations, those assigned to treatment have different covariate distributions than those assigned to control, we can deduce that the missing data are not missing at random. Motivated by this conflict, Imbens and Pizer (1999) developed alternative models for the analysis from randomized experiments with missing pretreatment variables and outcomes, which are both consistent with the data and preserve the appeal of MAR.

In nonrandomized studies, missingness of background variables is a more serious problem, because the missingness itself may be predictive about which treatment is received. Consider the problem of estimating and using propensity scores with partially missing covariate data.

Estimation of propensity score in the complete data case is generally straightforward since it uses standard methods (e.g., logistic regression or discriminant analysis) and relies on diagnostics that are relatively easy to calculate and interpret (see section 2.3).

When missing covariate data exist, it is no longer obvious how to estimate propensity scores. Any technique will have to either make a stronger assumption regarding ignorability of the assignment mechanism or will have to make an assumption about the missing data mechanism.

As described in section 2.3, propensity score matching relies heavily on the assumption of ignorability of the assignment mechanism, which depends on the relationship between $Z$ and all of the other study variables. In order to maintain this ignorability of the assignment mechanism any existing technique for estimating propensity scores with incomplete data needs to assume at least that

$$\Pr\big(Z \mid \mathbf{X}, R_x, Y(C), Y(T)\big) = \Pr\big(Z \mid \mathbf{X}, R_x\big), \tag{4.1}$$

where $R_x$ is the missing covariate indicator ($R_x = 1$ for observed, and $R_x = 0$ for missing).

Several possible approaches to the missing data problem exist. Complete case and complete variables (and combinations thereof) are extremely common approaches to missing data. Little and Rubin (1987) outline the potential problems with reliance on these simple but unprincipled approaches. Rosenbaum and Rubin (1984) and D'Agostino and Rubin (2000) developed two strategies precisely for the issue of estimating propensity score in presence of covariate missing data. Hill (2004) proposed some of the several possible combination of Multiple Imputation (MI) and propensity score matching. Now we briefly discuss the assumptions implicit in each of these possible methodologies, assuming that the assumption (4.1) holds.

Complete-case analyses use only observations where all variables are observed, and it is based on the Missing Completely At Random assumption (MCAR) (Rubin, 1976c; Little and Rubin, 1987). In our context, given the assumption (4.1), the missing data mechanism is MCAR if

$$\Pr\bigl(R_x \mid \mathbf{X}, Z\bigr) = \Pr\bigl(R_x\bigr).$$

This assumption, fairly strong and in many cases implausible, has several testable implications and is often rejected by the data.

A complete-variables analysis uses only the fully observed variables, denoted by $\mathbf{X}_f$. This type of analysis will fail if any of the covariates excluded are not independent of treatment assignment conditional on $\mathbf{X}_f$ and $R_x$, and are also related to the potential outcomes (again, conditional on $\mathbf{X}_f$ and $R_x$). Formally, we need

$$\begin{array}{rcl} \Pr\bigl(Z \mid \mathbf{X}, R_x\bigr) & = & \Pr\bigl(Z \mid \mathbf{X}_f, R_x\bigr) \\ & \text{or} & \\ \Pr\bigl(Y(C), Y(T) \mid \mathbf{X}, R_x\bigr) & = & \Pr\bigl(Y(C), Y(T) \mid \mathbf{X}_f, R_x\bigr). \end{array}$$

In words, we would have to believe either the variables removed were independent of treatment assignment or that ignorability of the assignment mechanism depends in fact only upon the variables retained.

This approach makes no assumption about the missing data mechanism. However, the omission of any variables with missing data will generally throw away too much information to continue to justify the ignorability of the assignment mechanism.

Rosenbaum and Rubin (1984) considered using a "pattern mixture" model (Little 1993; Rubin 1986) for propensity score estimation with missing covariate data. Appendix B of Rosenbaum and Rubin (1984) defined a "generalized" propensity score as the probability of treatment assignment given $\mathbf{X}^*$, a $K$-coordinate vector, where the $j$th element of $\mathbf{X}^*$ is a covariate value if the $j$th covariate was observed, and is an asterisk if the $j$th covariate is missing (formally, $\mathbf{X}^*$ is an element of $\{\mathbb{R}, *\}^K\}$). This is equivalent to conditioning on the observed values of $\mathbf{X}$, $\mathbf{X}^{\text{obs}}$, and the missing covariate indicator $R_x$; with discrete covariates, this is equivalent to adding an additional missing category to each covariate. Rosenbaum and Rubin (1984) proved that adjustment for the generalized propensity score in expectation balances the observed covariate information and the pattern of missing covariates. In addition, they suggested that in large enough samples, one can estimate this generalized propensity score by estimating a separate logit model using the subset of covariates fully observed for each pattern of missing data. The practical problem is that typically there are many patterns of missing data with only a few individuals from each of the two treatment groups, thereby making the straightforward pattern mixture approach infeasible.

D'Agostino and Rubin (2000) proposed a solution to this problem. Their modeled the joint distribution of $(Z, \mathbf{X}, R_x)$ using a general location model (Olkin and Tate, 1961) accounting for the missing data (Schafer, 1997). This modeling implies a conditional distribution for $Z$ given $(\mathbf{X}^{\text{obs}}, R_x)$; that is, the generalized propensity score: probabilities of $Z = T$ versus $Z = C$ for each unit as a function of its observed covariate values $\mathbf{X}^{\text{obs}}$ and missing pattern $R_x$. Because $\mathbf{X}$ is missing when $R_x = 0$, a saturated model for $(\mathbf{X}, R_x)$ cannot be fit, even with the general location model, so D'Agostino and Rubin (2000) imposed log-linear constraints on the categorical variables which include the missing value indicators for covariates whose missingness is related to treatment assignment, and estimated the propensity scores using the ECM algorithm (Meng and Rubin, 1993). D'Agostino and Rubin (2000) suggested that in the special case of no missing data and only continuous covariates, their approach reduces to estimating propensity score by discriminant analysis.

These two methods developed for estimating the generalized propensity scores by Rosenbaum and Rubin (1984) and D'Agostino and Rubin (2000) respectively, rely on either one of the following assumptions:

$$\Pr\bigl(Z \mid \mathbf{X}, R_x\bigr) \quad = \quad \Pr\bigl(Z \mid \mathbf{X}^{\text{obs}}, R_x\bigr)$$

or

$$\Pr\bigl(Y(C), Y(T) \mid \mathbf{X}, R_x\bigr) \quad = \quad \Pr\bigl(Y(C), Y(T) \mid \mathbf{X}^{\text{obs}}, R_x\bigr).$$

One way of thinking of these assumptions is as follows. Within each missing data pattern defined by $R_x$, we either need assignment to be independent of the covariates unobserved in that pattern, or we need ignorability to be satisfied just on the basis of those covariates observed in that pattern.

The strength of these two methods is that, in principle, they do not make any assumption about the missing data process, yet still makes weaker assumptions on the assignment mechanism and response surface than the complete variables approach. However, they do assume that either all missing covariate values are already balanced across treatment groups or that they are independent of the potential outcomes conditional on the observed covariate values and missing data patterns.

Another potential weakness of these methods is that since they specify one model for both handling missing data and estimating propensity scores they will not always have the possibility to incorporate $Y$ into this model, even though it might provide useful information about missing values.

To overcome this weakness, we can think of handling the incomplete data using Multiple Imputation (MI) techniques (Rubin, 1978). MI is a Monte Carlo technique in which each missing value is replaced by $m > 1$ simulated versions, where $m$ is typically small (e.g., 3-10). Each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Most of the techniques presently available for creating MIs assume that the missing data mechanism is ignorable, but it is important to note that the MI paradigm does not require ignorable nonresponse. Imputation may in principle be created under any kind of model for the missing-data mechanism, and the resulting inferences will be valid under that mechanism (see chapter 6, Rubin, 1987).

Hill (2004) suggested that the combination of MI and propensity score matching implicitly assumes the latent ignorability of the assignment mechanism. Latent ignorability was first introduced by Frangakis and Rubin (1999) as an extension of standard ignorability in the context of missing data mechanism. It describes a situation where the mechanism is ignorable only when conditional on certain latent or missing values, in addition to the observed values. In our case, the assignment mechanism is ignorable only conditional on complete covariate data (which includes, of course, values that in practice are missing). Computationally, this is achieved by filling in the missing covariate values using MI. Hill (2004) illustrated two approaches to combining propensity score matching with multiple imputation and discussed the required structural assumptions. Furthermore, she evaluated the potential relative performance of this methods using simulation models with compatible assumptions. Hill (2004) found that the MI methods seem to outperform not only the complete case and the complete variables analyses, but also the D'Agostino-Rubin method, and suggested that the MI techniques can accomodate a broader range of missing data models, matching methods, and analysis models.

## 5 Extensions to Accomodate Missing Outcome Data

It is unusual to have missing data for baseline characteristics but have fully observed outcomes. When outcome data is incomplete we must also consider the mechanism behind that missingness, that is, we need to consider assumptions about

$$\Pr\big(R_y(C), R_y(T) \mid Z, Y(C), Y(T), \mathbf{X}, R_x\big).$$

Complete case analyses in the presence of missing data require the additional assumption

$$\Pr\big(R_y(C), R_y(T) \mid Z, Y(C), Y(T), \mathbf{X}, R_x\big) = \Pr\big(R_y(C), R_y(T) \mid Z\big).$$

This means that, as long as the outcome missing data mechanism is independent of the covariates and potential outcomes then the observations removed from the sample will be a random sample of the entire dataset with respect to those variables. If these distributions remain the same in the observations remaining then the analyses can proceed appropriately. Similar to the discussion about complete cases with covariate missing data, this MCAR assumption, which has several testable implications, is often rejected by the data

In this scenario, strictly speaking, complete variables analyses cannot be performed because $Y$ has been removed from the dataset. However, we can use a combination of complete covariate variables and cases that have the outcome observed. The additional assumption needed here then is the same as for complete cases.

Another potentially more plausible assumption is the Missing At Random (MAR) model proposed by Rubin (1976c), which essentially allows the probability of nonresponse to depend on observed but not on unobserved variables, that is, it assumes that missing data values carry on information about the probability of missingness. Formally, we require the following two assumptions

$$\begin{aligned} \Pr\bigl(R_x \mid Z, \mathbf{X}\bigr) &= \Pr\bigl(R_x \mid Z, \mathbf{X}^{\text{obs}}\bigr) \\ &\text{and} \\ \Pr\bigl(R_y(C), R_y(T) \mid Z, Y(C), Y(T), \mathbf{X}, R_x\bigr) &= \Pr\bigl(R_y(C), R_y(T) \mid Z, \mathbf{X}^{\text{obs}}, R_x\bigr). \end{aligned}$$

In randomized experiments, the covariates $\mathbf{X}$ and the missing covariate indicator $R_x$ are independent of the treatment indicator $Z$, so the MAR assumption implies that $R_x$ can be ignored in the analysis. Moreover, under MAR, in randomized experiments we have that[4]

$$\Pr\bigl(R_y(C), R_y(T) \mid Z, \mathbf{X}^{\text{obs}}, R_x\bigr) = \Pr\bigl(R_y(C), R_y(T) \mid \mathbf{X}^{\text{obs}}, R_x\bigr).$$

The MAR assumption is convenient because it allows us to avoid an explicit model of nonresponse and is often relatively plausible. In addition, if the parameters of a MAR missing data process are distinct from those of the data distribution, then the missing data process is ignorable. Unfortunately, data can never provide any direct evidence against MAR, so that MAR is not testable without auxiliary information.

[4] See Barnard et al., 2003.

The framework of multiple imputation can be easily extended to handle jointly missing outcomes and covariates. As noted previously, MI can be created under any kind of assumption about the missing data process; the resulting inferences will be valid under that missing data mechanism (Rubin, 1987).

In observational studies, methodologies precisely developed for the issue of estimating propensity scores in the presence of covariates missing data, such as the general location model proposed by D'Agostino and Rubin (2000) and the pattern mixture model developed by Rosenbaum and Rubin (1984), need to be combined with complete cases for outcomes because the missing data model cannot incorporate outcomes in general since it is also used to compute propensity scores. Alternatively, missing outcomes could be imputed through separate process.

Concerning experimental studies, in addition to missing background and outcome data, they often suffer from noncompliance with the randomly assigned treatment.

Noncompliance occurs when the actual treatment that subjects receive differ from their nominal assignment. Here we assume all or none noncompliance: after randomization, some subjects assigned to the new treatment will not take it, but effectively take control, whereas some those assigned control receive the new treatment. In such a case, the compliance behavior is a variable defined by the joint vector of treatment receipt under both treatment assignments, say $(D(C), D(T))$. Specifically, this variable identifies four strata of people: compliers, those who take the treatment if so assigned and take the control if so assigned; never-takers, those who never take the treatment no matter the assignment; always-takers, those who always take the treatment no matter the assignment; and defiers, who would do the opposite of the assignment no matter its value. These strata are not fully observed, however, by randomization, their distribution is the same in both treatment arms. Such stratification, which dates back at least to Imbens and Rubin (1997) for randomized trials with noncompliance, is a direct application of the idea of principal stratification (Frangakis and Rubin, 2002) using the framework of the Rubin Causal Model.

In randomized studies with compliance as the only partially uncontrolled factor, and where there is full outcome data, 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 effect (comparing all units by their assignment rather than by the treatment actually received). More recently researchers have also focused on the intention-to-treat effect for subpopulation of compliers (Bloom, 1984; Sommer and Zenger, 1991; Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Imbens and Rubin, 1997; 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 a randomized experiment suffers from both noncompliance and missing outcome data, then these two complications have to be jointly taken into account and modeled in some principled way.

With respect to the response behavior, an often appealing assumption that has been proposed to link noncompliance with nonresponse is Latent Ignorability (Frangakis and Rubin, 1999). Under Latent Ignorability, if we knew the compliance type for all the subjects, the missing data mechanism would be ignorable, that is, potential outcomes and potential response indicators are assumed to be independent within each level of the compliance variable (with the same value of the fully observed covariates). This assumption alone does not allow full identification of the ITT effect for compliers, the effect of primary interest.

Along with the no-defier assumption, also called monotonicity assumption (Angrist and Imbens, 1994; Angrist et al., 1996), one choice for additional assumptions, which lead to full identification, is the “compound exclusion” restriction for never-takers and always-takers (Frangakis and Rubin, 1999; Barnard et al., 2003): when assignment has no effect on the treatment taken (for never-takers and always-takers), it has no effect on outcomes or response behavior as well. Of course, the plausibility of such an assumption depends on the context of the application.

Alternative assumptions can also be made. For example, one can assume that the exclusion restrictions on the outcome hold for never-takers and always-takers, but the exclusion restrictions on the response behavior hold for always-takers and compliers. The rationale for this is that those who would decline participation (the never-takers) might be induced, by the offer to participate, into not providing information on their outcome, which would have otherwise (Mealli et al., 2004).

This framework can be extended to allow for missing covariate. We can find a challenging example in Barnard et al. (2003), where a randomized study suffering from complications due to missing background and outcome data, and noncompliance with the randomly assigned treatment is described. They address these complications using a Bayesian approach with the framework of principal stratification (Frangakis and Rubin, 2002).

## 6 The Three Major Approaches to Causality

Recent years have seen an increased discussion about causation and a variety of causal models in the fields of economics, statistics, computer science, epidemiology, and sociology. Leaving the century-lasting discourse on accounts of causation in philosophy aside, this increased research on matters of causation in the above-mentioned fields has led to three major approaches to modeling causation currently dominating the debate on causal inference. These are Structural Equation Models (SEMs), Rubin Causal Model (RCM), and Direct Acyclic Graphs (DAGs). We have broadly described the Rubin Causal Model in the previous sections, being the statistical framework that we adopt for causal inference. In this section, we will give brief introductions to the other two approaches, and discuss somewhat further how the three frameworks are perceived in the academic community.

Structural Equation Models (SEMs) as an approach to causation are mainly used in economics and the social sciences. The SEMs have their origin in path analysis developed by geneticists (Wright, 1921, 1934). Founding work in SEMs was done by Haavelmo (1943, 1944) and Koopmans and Hood (1953), work that set the stage for modern econometrics (see Morgan, 1990; and Heckman, 2000 for further discussion). In fact, SEMs have remained the paradigm of causal modeling in contemporary econometrics and the social and behavioral sciences. Structural equation models rely on the specification of systems of equations with parameters and variables that attempt to capture behavioral relationships and specify the causal links between variables. Specifically, Goldberger (1972) defined SEMs as "stochastic models in which each equation represent a causal link".

The use of Direct Acyclic Graphs (DAGs) to assess causal questions is a rather recent phenomenon. Currently, the main proponent of graphical approaches to causation are Spirtes, Glymour, and Scheines (2000, first edition 1993), and Pearl (1995, 1998, 2000). It is beyond the scope of this section to discuss the functioning and mechanisms of DAGs, and difficult to do so in just a few phrases - for an introduction see the mentioned papers and books. Rather, we want to describe what their advocates think DAGs are aimed at: they are aimed at making causal relations and assumptions and implications in causal models more explicit, in particular more explicit than - in the view of their proponents - other approaches. For instance, Pearl (2000) claims that recent advantages in DAGs have transformed causality from "a concept shrouded in mystery" into a mathematical object with well-defined semantics and well-founded logic. This is another aim of the graphical approach, namely to provide causal talk with a common language helping researches communicate (Pearl 1995, 1998), an aim that DAGs do not yet live up to in the view of everybody - see the discussion of Pearl (1995), in particular Imbens and Rubin (1995), Rosenbaum (1995). Pearl (2000) strongly emphasizes the gain in clarity and explicitness gained from causal models based on DAGs in his view. For better or worse, his conclusion is that due to DAGs "causality has been mathematized" (Pearl, 2000).

Naturally, different approaches to questions of causation are viewed differently by proponents of different approaches. For instance, Rubin and Imbens (1995) argue that graphical models in general, and DAGs in particular, with their nodes, directed arrows, undirected arrows, absence of arrows, etc., are extremely seductive, but they provide a framework for causal inference that is inherently less revealing than the potential outcomes framework because it tends to bury essential scientific and design issue. Also perceptions SEMs as an adequate approach to causation diverge strongly. Pearl (1998) unfolds the idea that the original conceptual strength of SEMs along with the clear conception of it among its founding father has been lost since, or at least become "obscured". In his belief, "the causal content of SEMs has gradually escaped the consciousness of SEMs practitioners" (Pearl, 1998) for two reason: (*i*) SEMs practitioners have kept causal assumptions implicit in order to gain respectability for SEMs, because statisticians, "the arbiters of respectability", abhor assumptions that are not directly testable, and (*ii*) SEMs lack the notational facility needed to make causal assumptions, as distinct from statistical assumptions, explicit. The latter point means that the SEMs founding fathers thought of the equality sign as the asymmetrical relation "is determined by" rather than an algebraic equality, but did not invent a distinct sign for this relation. They were aware of the distinction, but now their descendants seem to have lost this clear conception - for more on this issue see Pearl (1998), who evidently develops this idea to contrast it with DAGs as a more coherent tool of causal language.

On the other hand, Heckman (2000) is a clear proponent of SEMs and forcefully stresses the major role that econometric analysis played in the twentieth century analysis of causal parameters:

> A major contribution of twentieth century econometrics was the recognition that causality and causal parameters are most fruitfully defined within formal economic models and that comparative statics variations within these models formalize the intuition in Marshall's (notion of a

ceteris paribus change) and most clearly define causal parameters.

This is how economists define causal effects.[5] Heckman (1996, 2000) argues that the statistical RCM, based on potential outcomes, is simply a version of the econometric causal model. This is in line with his finding that the definition of a causal parameter does not require any statement about what is actually observed or what can be identified from data. A finding that also the SEM founding fathers would have subscribed to, as Pearl (1998) refers to Haavelmo (1943) who explicitly interprets each structural equation as a statement about a hypothetical controlled experiment.

In his comment on Angrist et al.'s paper (1996), Heckman (1996) argues that RCM is a version of the widely used econometric switching regression model. On the contrary, Angrist et al. (1996) view the term Rubin Causal Model (coined by Holland, 1986 for work by Rubin, 1974, 1978) as referring to a model for causal inference where causal effects are defined explicitly by comparing potential outcomes. This comparison can be in the context of a randomized experiment or an observational study. Any element of the set of the potential outcomes could have been observed by manipulation of the treatment of interest, even though ex-post only one of them is actually observed. Moreover, the RCM defines the assignment mechanism, which determines which potential outcome are observed, as the conditional probability of each possible treatment assignment given the potential outcomes and possibly other variables. In contrast, the switching regression model as defined by Quandt (1958, 1972) is a time series model where the first part of the sample comes from one regression model and the second part from a separate regression model with an unknown switching point (Angrist et al., 1996 - rejoinder).

A second example mentioned by Heckman (1996) is Roy (1951) who studied the distribution of observed incomes in a world where individuals always choose the occupation with the highest income. Neither Roy (1951) nor Quandt (1958, 1972) discussed causal effects. As Angrist et al. (1996) argue, what makes the

[5] A different concept of causation in time series econometrics is Granger causation, which we will not discuss (Granger, 1969).

Roy model and the switching regression model technically closer to the RCM than many other models used in econometric evaluation studies is their explicit focus on potential outcomes as distinct from observed outcomes. Only recently has the RCM potential outcome framework been adopted in economic models for causal effects (e.g., Maddala, 1983; Bjorklund and Muffit, 1987; Heckman, 1990; and Manski, 1990).

Recent years have seen substantial convergence of methods from statistics and econometrics. For instance, Angrist et al. (1996) provide a link between the RCM potential outcomes framework and the Instrumental Variables (IV) approach, capitalizing on the strengths of each. Inference in structural equation models often exploits the presence of instrumental variables. These are variables that are explicitly excluded from some equations and included in others, and therefore correlated with some outcomes only through their effect on other variables. Angrist et al. (1996) show how the IV estimand can be given a precise and straightforward causal interpretation in the potential outcomes framework, despite nonignorability of treatment received. This interpretation avoids the drawbacks of the standard structural equation framework, such as constant effects for all units, and delineates critical assumptions needed for a causal interpretation. Specifically, Angrist et al. (1996) show that the IV estimand can be embedded within the RCM and that under some simple and easily interpretable assumptions, the IV estimand is the average causal effect for a subgroup of units, the compliers. Without these assumptions the IV estimand is simply the ratio of the intention-to-treat causal estimands with no interpretation as an average causal effect.

Standard IV procedures rely on judgments regarding the correlation between functional-form-specific disturbances and instruments. Typically the researcher does not have a firm idea what these disturbances really represent, and therefore it is difficult to draw realistic conclusions or communicate results based on their properties. In addition, the SEMs are sensible to critical assumptions (see Little, 1985) and apparently unable to reproduce experimental results (see Lalonde, 1986). The causal interpretation of the IV estimand using the potential outcomes framework, allows for a formulation of the critical assumptions in a more transparent manner and so it makes these models more accessible to statisticians. Moreover, by separating and defining the critical assumptions, the potential outcomes framework allows for a clear assessment of the consequences of violations of these assumptions trough sensitivity analysis under more general models (Angrist et al., 1996).

In summary, one cannot but highly appreciate the vivid debate on causation in various fields, the expanding amount of causal models suggested, and the analogies, connections and distinctions that have been drawn among models from different fields.

# 7 Conclusion

Just as Neyman's notation for randomized experiments was not obvious and Fisher's suggestion to physically randomize units was not obvious, so too the transition to use potential outcomes as the definition of causal effects, whether or not the assignment mechanism was randomized, was not obvious.

The framework that we describe here, using potential outcomes to define causal effects and thereby the assignment mechanism, has been called the Rubin Causal Model (RCM) by Holland (1986) for work initiated in the 1970's (Rubin, 1974, 1977, 1978). This perspective conceives of all problems of statistical inference as missing data problems with a mechanism for creating missing data (Rubin, 1976). The RCM has the following salient features for causal inference: (1) causal effects are defined as comparisons of a priori observable potential outcomes without regard to the choice of assignment mechanism that allows the investigator to observe particular values; (2) interference between units and variability in efficacy of treatments can be incorporated in the notation so that the commonly used "stability" assumption can be formalized as deviations from it; (3) models for the assignment mechanism are viewed as methods for creating missing data, thereby allowing nonrandomized studies to be considered using the same notation as used for randomized experiments; (4) potential outcomes and covariates can be given a joint distribution, thereby allowing both randomization-based methods, traditionally used for randomized experiments, and model-based methods, traditionally used for observational studies, to be applied to both kinds of studies; (5) there are explicit mathematical results showing the role of randomization for both randomization-based and Bayesian inference (Rubin, 1977, 1978).

After a brief description of the RCM potential outcomes framework, this paper focuses on describing and addressing complications due to missing background and outcomes data in randomized and nonrandomized studies using this framework.

Randomized experiments offer many benefits to the researcher. The randomization of treatment assignment ensures that treatment and control groups are comparable, and therefore causal inferences regarding the average causal effects of the treatment of interest can be drawn without additional assumptions. Specifically, randomization avoids the need for modeling the outcome distributions because it ensures that average causal effects can be estimated by the difference between average treatment outcomes and average control outcomes.

These benefits, however, require that we have complete data on treatment and response for all units. In presence of missing data, it is necessary to make assumptions regarding the dependence of the missing data mechanism on both treatment assignment and values of missing variables. In addition, if the study also suffers from noncompliance, compliance behavior and response behavior have to be jointly taken into account and modeled in some principled way.

Observational data with missing data, both for covariates and outcomes variables, are prevalent in the social sciences. Studies using such data to make causal inferences are increasingly making use of propensity score techniques as a means for controlling for observable differences between treatment groups. These methods are complicated, however, by the addition of missing data.

In this paper, we illustrated some of the possible existing approaches to the missing data problem and discussed the underlying assumptions. Specifically, we review and propose different sets of assumptions, and discuss which assumptions seem to be more appropriate for different settings. An important lesson is that there are no universally appropriate assumptions; the most plausible assumptions are specific to each context.

## References

Angrist, J. D. (1990) Lifetime earnings and the Vietnam era draft lottery: evidence form social security administrative records. *American Economic Review*, **80**, 313–335.

Angrist, J. D., and Krueger, A. (1991) Does compulsory school attendance affect schooling and earnings. *Quarterly Journal of Economics*, **106**, 979–1014. item[] 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.

Baker, S. G. and Lindeman, K.S. (1994) The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statistics in Medicine*, **13**, 2269–2278.

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 missing binary outcome, an auxiliary variable, and all-or-none compliance. *Journal of the American Statistical Association*, **95**, 43–50.

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(2)**, 285–317.

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.

Bjorklund, A., and Moffitt, R. (1987) Estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics*, **69**, 42–49.

Bloom H. (1984) Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, **164**, 1837–1846.

Cochran, W. G., and Rubin, D. B. (1973) Controlling bias in observational studies: a review. *Sankhya*, Series A **35**, 417–446.

Connors A. F. et al. (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, **276**, 889–97

Frangakis, C. E. and Rubin, D. B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-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.

Goldberger, A. S. (1972) Structural equation methods in the social sciences. *Econometrika*, **40**, 979–1001.

Granger, C. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Haavelmo, T. (1943) The statistical implication of a system of simultaneous equations. *Econometrica*, **11**, 1–12.

Heckman, J. J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrika*, **46**, 931–959.

Heckman, J. J. (1990) Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **58**, 1121–1150.

Heckman, J. J. (1996) Comment on "Identification of causal effects using instrumental variables" by Angrist J. D., Imbens, G. W., and Rubin D. B. *Journal of the American Statistical Association*, **91**, 459–462.

Heckman, J. J. (2000) Causal parameters and policy analysis in economics: a twentieth century retrospective. *Quarterly Journal of Economics*, **115**, 45–97.

Hill J. (2004) Reducing bias in treatment effect estimation in observational studies suffering from missing data. Working paper series, *School of International and Public Affairs*, Columbia University, NY.

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 Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–476.

Imbens, G. W., and Pizer, W. A. (1999). The analysis of randomized experiments with missing data. Working paper.

Imbens, G. W. and Rubin, D. B. (1994) *Causal inference with instrumental variables* Discussion paper 1976. Cambridge, Massachusetts, Harvard Institute of Economic Research.

Imbens, G. W. and Rubin, D. B. (1995) Comment on "Causal diagrams for empirical research" by J., Pearl, J. *Biometrika*, **82**, 694–695.

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

Koopmans, T., Hood, W. (1953) *The estimation of simultaneous linear economic relationships* in W. Hood and T. Koopmans (eds), Studies in Econometric Method, New York, Chapman & Hall.

Lalonde, R. (1986) Evaluating the econometric evaluations of training programs. *American Economic Review*, **76**, 604–620.

Little, R. J. A. (1985) A note about models for selectivity bias. *Econometrica*, **53**, 1569–1474.

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

Little, R., and Yau, L. (1998) Statistical techniques for analyzing Data from prevention trials: treatment of no-shows using Rubin's Causal Model. noncompliance. *Psychological Methods*, **3**, 47–159.

Maddala, G. S. (1983) *Limited-dependent and qualitative variables in econometrics.* Cambridge, U.K., Cambridge University Press.

Manski, C. F. (1990) Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, **80**, 319–323.

Manski, C. F., Sandefur, G. D., McLanahan, S., and Powers, D. (1992) Alternative estimates of the effects of family structure during adolescence on high school graduation. *Journal of the American Statistical Association*, **87**, 25–37.

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.

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.

Meng, X. L., and Rubin, D. B. (1993) Maximum likelihood estimation via ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Miettinen, O. (1976) Stratification by multivariate confounder score. *American Journal of Epidemiology*, **104**, 609–620.

Morgan, M. (1990) *The history of econometric ideas*, Cambridge, Cambridge University Press.

Neyman, J. (1923) On the application of probability theory to agricultural experiments: essay on principles, section 9. *Translated in Statistical Science*, **5**, 465–480, 1990.

Olkin, I., and Tate R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–465.

Pearl, J. (1995) Causal diagrams for empirical research (with discussion). *Biometrika*, **82** 669–710.

Pearl, J. (1998) Graphs, causality, and structural equation models. *Sociological Methods and Research*, **27** 226–284.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*, Cambridge, Cambridge University Press.

Quandt, R. E. (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of American Statistical Association*, **53**, 873–880.

Quandt, R. E. (1972) A new approach to estimating switching regressions. *Journal of American Statistical Association*, **67**, 306–310.

Reinisch, J., Sanders, S., Mortensen, E., and Rubin, D. B. (1996) In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association*, **274**, 1518–1525.

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 American Statistical Association*, **89**, 737–749.

Rosenbaum, P. R. (1984a) From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, **79**, 41–48.

Rosenbaum, P. R. (1984b) The consequences of adjustment for a concomitant variable that has been affected by the treatment *Journal of the Royal Statistical Society*, Series A **147**, 656–666.

Rosenbaum, P. R. (1985) Comment on "Causal diagrams for empirical research" by J., Pearl, J. *Biometrika*, **82**, 656–666.

Rosenbaum, P. R., and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biomatrika*, **70**, 41–55.

Rosenbaum, P. R., and Rubin, D. B. (1983b) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. studies for causal effects. *Journal of the Royal Statistical Society*, Series B **45**, 212–218.

Rosenbaum, P. R., and Rubin, D. B. (1984) Reducing bias in observational studies using subclassication on the propensity score. *Journal of the American Statistical Association*, **79**, 516–524.

Rosenbaum, P. R., and Rubin, D. B. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. the propensity score. *The American Statistician*, **39**, 33–38.

Roy, A. D. (1951) Some thoughts on the distribution of earnings. *Oxford Economic Papers*, **3**, 135–146.

Rubin, D. B. (1970) The use of matched sampling and regression adjustment in observational studies. Unpublished PH.D. dissertation, Harvard University, Dept. of Statistics.

Rubin, D. B. (1973) The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29(1)**, 185–203.

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

Rubin, D. B. (1976a) Matching to remove bias in observational studies. *Biometrics*, **29**, 185–203; Prints correction (1974), **30**, 728.

Rubin, D. B. (1976b) The use of matching and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

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

Rubin, D. B. (1977) Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1–26.

Rubin, D. B. (1978) Bayesian inference for causal effects. *Annals of Statistics*, **6**, 34–59.

Rubin, D. B. (1979) Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*, **74**, 318–328.

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

Rubin, D. B. (1980b) Bias reduction using Mahalanobis metric matching. *Biomatrics*, **36**, 293–298.

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

Rubin, D. B. (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**, 472–480.

Rubin, D. B. (1998) More powerful randomization-based $p$-values in double-blind trials with noncompliance. *Statistics in Medicine*, **17**, 371–387.

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

Rubin, D. B. and Thomas N. (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, **95**, 573–585.

Schafer, J. L. (1997) *Analysis of incomplete data*, Chapman and Hall.

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.

Spirtes P., Glymour C., and Scheines R. (2000) *Causation, prediction, and search.* New York, Springer, 2nd edition.

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

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

Tinbergen, J. (1930) Determination and Interpretation of Supply Curves: An Example. In *The Foundations of Econometric Analysis*, eds. D. Hendry and M. Morgan, Cambridge: Cambridge University Press, reprinted from Zeitschrift fur Nationalokonomie.

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

Wright, S. (1934) The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161–215.

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

## 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". We present an extended framework for the analysis of data from randomized experiments which suffer from these complications. 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 (BSE). In the study two methods of teaching BSE, consisting of either mailed information about BSE (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 suffers from the complication 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 posttest questionnaire. In addition, quality of self-exam execution is "truncated by death", in the sense that there is no hidden value of the quality outcome for women who do not practice BSE, the truncating event. Our analysis reveals a positive, even if not highly significant, effect on quality of self exams for women who always comply with their assignment and would practice BSE under both treatment arms.

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

## 1 Introduction

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. Although BSE is simple, non invasive and inexpensive, its effectiveness is heavily debated in spite of more then 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 teaching methods, 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 el., 2000; Strickland et al., 1997; Mishra et al., 1998; Giles et al., 2001).

In this paper, we will consider one of such studies, a randomized experiment conducted between January 1988 and December 1990 in Faenza (Italy). In this study, 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 with the aim of assessing whether teaching programs could increase BSE practice and improve examination skills.

As in most research involving human subjects, our study also suffered from complications due to missing data and noncompliance with the randomly assigned treatment. In general, noncompliance is selective in the sense that noncompliers and compliers generally differ in background characteristics. Moreover, missing outcomes, caused by loss to follow-up in our study, may also be selective in the analogous sense (Farwell et al., 1990). These complications are rarely fully under the experimenter's control, and there is currently substantial awareness among researchers that such complications in a study compromise the ability to draw clear conclusions. Therefore, 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 our application, we have to face on another complication which is linked to the topic of "truncation by death". Quality of self exams 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; therefore the estimation of the causal effects of the enhanced BSE training course on quality of self exams requires that treatment comparisons are adjusted for BSE practice status. The solution 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, because they ignore the fact that the quality outcome for women who do not practice BSE is neither "censored" nor "missing"; it should be treated as being defined on an extended sample space (Rubin, 2000; Frangakis and Rubin, 2002; Zhang and Rubin, 2003). We call the quality outcome "truncated by death" because there is no hidden value of the outcome variable masked by the truncating event.

Here we focus on describing and addressing these complications in our study using a Bayesian approach with the framework of principal stratification (Frangakis and Rubin, 2002). Principal stratification is a general framework for comparing treatments where the estimands are adjusted for posttreatment variables and yet are always causal effects.

The Faenza BSE study offers us a good opportunity to develop an extended framework for the analysis of randomized experiments which require that treatment comparisons be adjusted for (*i*) noncompliance with the randomly assigned treatment, (*ii*) missing outcomes (dropout) following treatment noncompliance, and (*iii*) "truncation by death".

We describe the study in section 2, and summarize its data complications in section 3. Section 4 places the study in the context of broken randomized experiments, a phrase apparently first coined by Barnard, Du, Hill, and Rubin (1998). We present the framework we use in section 5 and section 6, and discuss our model's structural assumptions in section 7. Section 8 describes our parametric model specification, and main results of the analysis are presented in section 9. We discuss model building and checking in section 10 and conclude in section 11.

## 2 The Faenza Randomized Experiment on Breast Self-Examination

In this paper we reanalyze data of a randomized trial on Breast Self-Examination (BSE) conducted between January 1988 and December 1990 at the Oncologic Center of the Faenza Health District in Italy (see previous analyses by Ferro et al., 1996; and Mealli et al., 2004). In the 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. Both treatment levels were selected on the basis of their practical feasibility and their acceptability according to the cultural profile of the area.

Participants in this experiment were a random sample of 825 women, with ages ranging from 20 to 64, 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, presence of known risk factors for breast cancer, preventive beliefs, level of knowledge, practice and examination skills of BSE, and other individual characteristics. To evaluate whether BSE was correctly performed, according to the criteria of the Canadian National Breast Screening Study, the questionnaire included the following items (yes/no): preliminary visual examination; lying in a prone position; fingers used flatly; circular motion of the fingers; circular palpation; most of breast examined; axillae examined; check on nipple discharge; frequently (monthly or not); and BSE practice following menstruation. Each response was assigned an a-priori score; an overall index computed as the total score for each subject was used as indicator of quality of BSE practice.

Respondents to the pretest 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 an explicative 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 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 session, a group discussion and a fifteen-minute individual practice session. Women were invited to the course in small groups according to their education level by a letter and a telephone call in order to motivate them, often reluctant for work engagement, or little interest, or lack of time. Actually, 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 received 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, 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.

In the Faenza study, the question of interest was the effect of an enhanced training class on BSE practices and quality of self-exam execution. This quality outcome was assessed using the difference between the overall score obtained at the pretest and posttest, 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 an 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.

# 3 Data Complications

The study presented above is a two-arm randomized experiment that compares a new, enhanced teaching program to a standard treatment with access to the new training course only available to those in the enhanced treatment group. This study suffers from a number of complications that may compromise the analysis and require additional assumptions. In this section we describe these data complications and how they can bias a standard causal analysis. First consider the complications due to noncompliance with the randomly assigned treatment and the presence of missing outcome data.

The data we use include the background covariates, pre-test data and post-test data. For each individual $i$ who participates in the study we observe: a binary variable $Z_i^{\text{obs}}$, the treatment assignment, equal to $T$ if woman $i$ is assigned to the enhanced treatment, and $C$ otherwise; a binary variable $D_i^{\text{obs}}$, the actual treatment received, equal to $P$ if person $i$ participates in the BSE training program and $p$ otherwise ($D_i^{\text{obs}} = p$ if $Z_i^{\text{obs}} = C$, by definition); and two binary outcome variables $S_i^{\text{obs}}$, equal to $B$ if woman $i$ practices BSE and $b$ otherwise, and $Y_i^{\text{obs}}$ equal to $H$ if the quality of individual $i$'s posttreatment BSE practice exceeds the designed threshold (the overall study sample median of the quality indicator), and $L$ otherwise. Lastly, we observe $R_i^{\text{obs}}$, the response indicator (1 if a subject responds to the posttest questionnaire, 0 for non-responders). We consider only one indicator $R_i^{\text{obs}}$ for missingness of outcomes, because the outcomes on BSE practice and quality of self exams were either jointly observed or jointly missing. In addition, three covariates are observed: $X_{i1}^{\text{obs}}$, a binary indicator of previous BSE practice, $X_{i2}^{\text{obs}}$ a binary indicator of good knowledge of breast pathophysiology, and $X_{i3}^{\text{obs}}$ age in years. Table 1 presents some summary statistics for the sample, classified by assignment, $Z_i^{\text{obs}}$, and treatment status, $D_i^{\text{obs}}$.

As we can see in Table 1, the randomization of the assignment leads to the pretreatment variables being closely balanced in the two subsamples defined by assignment. The randomization does not, however, imply that the pretreatment variables are balanced in the subsamples defined by the actual treatment status. Knowledge of breast pathophysiology, $X_{i2}^{\text{obs}}$, prior to the program, for example, is significantly higher for those women who attended the course than those who did not. This imbalance indicates that attendance of the self-exam course was not perfectly correlated with assignment, so that treatment comparisons have to be adjusted for treatment status to obtain credible estimates of the effect of BSE training course. Specifically, if there was perfect compliance - and the outcomes were observed for each subject - the effect of teaching program on BSE practice could be simply estimated comparing BSE practice outcomes by treatment status; but since compliance is imperfect, this estimator can be coarse and even misleading if taken as summarizing the evidence in the data for the effects of treatment. As we will see, the estimation of causal effects of the enhanced BSE teaching program on quality of

Table 1: Faenza BSE study - Summary statistics.

| | Means | | | | | |
|---|---|---|---|---|---|---|
| | Grand mean | $Z_i^{\text{obs}} = C$ | $Z_i^{\text{obs}} = T$ | $Z_i^{\text{obs}} = T$ $D_i^{\text{obs}} = p$ | $Z_i^{\text{obs}} = T$ $D_i^{\text{obs}} = P$ | $D_i^{\text{obs}} = p$ |
| $N$ | 657 | 327 | 330 | 148 | 182 | 475 |
| Assignment ($Z_i^{\text{obs}}$) | 0.502 | 0 | 0 | 1 | 1 | 0.312 |
| Course attendance ($D_i^{\text{obs}}$) | 0.277 | 0 | 0.551 | 0 | 1 | 0 |
| Response ($R_i^{\text{obs}}$) | 0.653 | 0.688 | 0.618 | 0.399 | 0.797 | 0.598 |
| BSE practice ($S_i^{\text{obs}}$)$^*$ | 0.785 | 0.796 | 0.774 | 0.475 | 0.897 | 0.729 |
| BSE quality ($Y_i^{\text{obs}}$)$^*$ | 0.492 | 0.402 | 0.594 | 0.250 | 0.669 | 0.381 |
| Prior BSE practice ($X_{i1}^{\text{obs}}$)$^{**}$ | 0.585 | 0.591 | 0.579 | 0.551 | 0.601 | 0.579 |
| Knowledge of breast pathophysiology ($X_{i2}^{\text{obs}}$) | 0.554 | 0.560 | 0.548 | 0.439 | 0.637 | 0.522 |
| Age ($X_{i3}^{\text{obs}}$) | 41.4 | 41.5 | 41.3 | 41.7 | 41.0 | 41.6 |

($^*$) Computed on respondents only. ($^{**}$) Available for 615 women.

self exams is more problematic, because treatment comparisons should be adjusted not only for compliance status, but also for posttreatment BSE practice.

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%). Analyses that would be limited to complete cases would discard more that half of the units. Moreover, standard adjustments for outcome missingness ignore its potential interaction with the other complications and generally make implicit and unrealistic assumptions. For instance, we observe that within the group assigned to receive the active treatment, response rates significantly differ between those who complied with their assignment ($D_i^{\text{obs}} = P$) and those who did not ($D_i^{\text{obs}} = p$). This suggests that the compliance behavior may be related to the willingness to respond of the subjects. In our study, it seems reasonable that missingness could be also related to BSE behavior, that is, the potential BSE practice indicators and the associated potential non-response indicators could be dependent within each level of the latent compliance covariate regardless of the assigned treatment. This non ignorability of the missing data is crucial in our analysis, above all when we focus on evaluating the causal effects of the enhanced BSE training program on the quality outcome.

Recall that quality of BSE practice can only be observed for women who practice BSE ($S_i^{\text{obs}} = B$), and it is not only unobserved but also undefined on the usual sample space for women who do not practice the self exams. The quality for those who do not practice BSE can be defined as $*$ on the extended space $\{L, H, *\}$, although sometimes it is treated as "missing" or "censored" which would imply a hidden value on the sample space. Therefore, we need to "account for" BSE practice status (i.e., the possible occurrence of an observed $*$) when addressing the causal effects on quality. A common and seemingly obvious approach to adjust for BSE practice status is to compare the BSE practice groups under the two treatment arms, either through direct mean comparison or through regression adjusted comparison. This common approach can not extract the right information from observed data. In principle, a causal estimand of interest would be the effect of the treatment on the quality of self exam for those women who practice BSE under both assignments. In a randomized experiment with noncompliance, such a causal estimand would be the effect of the treatment for women who would comply with their treatment assignment no matter which assignment they would be given and would practice BSE under both treatment arms.

An additional complication limits our analysis sample. The indicator of previous BSE practice, $X_{i1}^{\text{obs}}$, is not available for all the women. In principle, missingness of background variables is also a covariate because it occurs before randomization. So, such missingness does not directly create unbalance of subjects between randomized arms, although it can create loss in efficiency when background covariates are to be used in the analysis. However, in the Faenza experiment, it seems reasonable to ignore the missingness of pretreatment variables; so all analyses in this paper are limited to results for the 615 women who gave complete information on all the covariates.

# 4 The Faenza Study as a Broken Randomized Experiment

The above deviations from the study's protocol clarify that our experiment does not really randomized attendance in the self-exam course, but that it randomizes the encouragement, using a letter and a telephone call, to attend the BSE teaching program. As in most encouragement studies, interest here focuses not only on the effect of encouragement itself, which will depend on what percentage of people encouraged would actually partecipate if experiment were to be implemented more broadly, but also on the effect of the treatment that is being encouraged, here, attending a "hands-on" training course on BSE techniques. If there were perfect compliance, so that all those encouraged to attend the BSE training program actually did so, then the standard intention-to-treat effect on BSE practice, being estimated typically, and the intention-to-treat effect on quality of self exams properly adjusted for BSE practice would be attributed to attendance of the course, rather than simply to the encouragement.

We focus on defining and estimating the causal effects on BSE practice rate and quality of self-exam execution using a framework for comparing treatments adjusting for the data complications in our study that yields properly defined causal estimands.

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 (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 of subjects belonging to specific latent groups defined by the compliance behavior and by the value of the posttreatment variable "BSE practice" under the two treatment arms (BSE practice behavior).

The causal estimands on BSE quality that we are interested to evaluate are: (1) the ITT effect for all women who would practice BSE under both assignments; and (2) the average causal effect for compliers who would practice BSE under both treatments (CACE on BSE quality). These quantities will be defined more formally in the next two Sections.

In recent years, there has been substantial progress in the analysis of encouragement designs, based on building bridges between statistical and econometric approaches to causal inference. In particular, the widely accepted approach in statistics to formulating causal questions is in terms of "potential outcomes". Although this approach has roots dating back to Neyman and Fisher in the context of perfect randomized experiment (Neyman 1923; Rubin 1990), it is generally referred as Rubin's Causal Model (Holland, 1986) for work extending the framework to observational studies (Rubin, 1974, 1977) and including modes of inference other than randomization-based, in particular, Bayesian (Rubin 1978a, 1990). In economics, the technique of "Instrumental Variables" (IV), due to Tinbergen (1930) and Haavelmo (1943), has been a main tool of causal inference in the type of non-randomized studies prevalent in that field. Angrist, Imbens and Rubin (1996) showed how the approaches can be view as completely compatible, that is, how the econometric IV methods can be interpreted as estimating a well-defined causal effect under the potential outcome approach. In particular, their main result was the interpretation of the IV technology as a way to approach a randomized experiment that suffers from noncompliance, such as a randomized encouragement design.

In encouragement designs with compliance as the only uncontrolled factor, and where there are full outcome data, Imbens and Rubin (1997) extended the Bayesian approach to causal inference in Rubin (1978a) to handle simple randomized experiments with noncompliance, and Hirano, Imbens, Rubin, and Zhou (2000) extended further the approach to handle fully observed covariates.

In encouragement designs with more than one partially uncontrolled factor, as with noncompliance and missing outcomes, defining and estimating treatment effects of interest is more challenging. Barnard, Frangakis, Hill, and Rubin (2003) developed a fully Bayesian analysis with the framework of principal stratification (Frangakis and Rubin 2002) to address complications due to missing background and outcome data, and noncompliance with the randomly assigned treatment in an encouragement design.

Principal stratification is a powerful framework for comparing treatments adjusting for posttreatment variables that yields properly defined causal effects. Principal stratification with respect to a posttreatment variable is a cross-classification of subjects defined by the joint potential values of that posttreatment variable under each of the treatments being compared. It generates principal effects, which are causal effects within a principal stratum. The key property of principal strata is that they are not affected by treatment assignment and therefore can be used just as any pretreatment variables. As a result, the central property of principal effects is that they are always causal effects and do not suffer from complications of standard posttreatment-adjusted estimands.

As in Barnard, Frangakis, Hill, and Rubin (2003), we use a Bayesian approach with the framework of principal stratification to address the complications in our study. Our principal strata are defined by the potential compliance status and by the joint potential values of BSE practice under each of the treatment conditions. In each principal stratum there are respondents and non-respondents. As stated previously, in our study, willingness to respond of the subjects can be related to both the compliance behavior and the BSE practice behavior. In principle, reasons for missing outcomes can be different for subjects who belong to different principal strata, and also, can be affected by treatment assignment, creating even more disparity between the type of people being compared. As the results shown by Frangakis and Rubin (1999) suggest, in such cases the respondent-based estimators are generally biased for the causal estimands of interest. Using the framework of principal stratification, we construct a new missing data model that explicitly allows both nonignorable distribution of units among principal strata and nonignorable nonresponse.

We fully develop a Bayesian framework that yields valid estimates of quantities of interest and also properly account for our uncertainty about these quantities.

## 5 Principal Stratification and Role for Causal Inference

In order to address better the complications discuss above, 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 practices BSE and $b$ otherwise; and $Y_i(z)$ for the potential quality outcome, where $Y_i(z) = H$ if the woman practice 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 practice BSE with "low" quality, that is, whether her BSE quality is lower than the fixed threshold. Lastly, we denote with $R_i(z)$ the indicator equal to 1 if the woman $i$ responds to the posttest questionnaire, and 0 otherwise. As noted previously, $Y_i(z)$ could take value in the extended set $\{L, H\} \cup \{*\}$, where $Y_i(z) \in \{L, H\}$ if $S_i(z) = B$, and $Y_i(z) = *$ if $S_i(z) = b$. 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 $Z_i^{\text{obs}}$ the observed treatment assignment, the observed data are

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

which we will denote by $(Z_i^{\text{obs}}, D_i^{\text{obs}}, R_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}})$, $i = 1, \ldots, N$, as suggested in section 3. In addition, 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 $i$th row equal to $\mathbf{X}_i^{\text{obs}} = (X_{i1}^{\text{obs}}, X_{i2}^{\text{obs}}, X_{i3}^{\text{obs}})$, the three observed background variables previously defined.

The potential outcome $D_i(T)$, that is, the treatment woman $i$ would received if assigned to the active treatment, is particularly important because it 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 $\mathrm{I}\{D_i^{\text{obs}} = P\} = \mathrm{I}\{Z_i^{\text{obs}} = Z\}$, 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 $T$. For our experimental setting, this compliance status $D_i(T)$ can be viewed as a covariate which is observed only for women with $Z^{\text{obs}} = T$ (Angrist et al., 1996); by randomization, however, it is guaranteed to have the same distribution in both treatment arms. Each of 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 $\pi$(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 $\pi$(PbB) of all women;

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 $\pi$(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 $\pi$(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 $\pi$(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 $\pi$(pbB) of all women;

Table 2: 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 $= \{i : D_i(T) = p, S_i(C) = B, S_i(T) = b\}$ : never-takers who would practice BSE under control but would not practice BSE under treatment, which comprise a proportion $\pi$(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 $\pi$(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 2.

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. Formally, we could regard the potential response indicator as another posttreatment variable respect to classify the subjects in the trial. In such case, we would have 32 principal strata. Thus, in our framework each stratum is actually union of four strata. For example,

$$\begin{aligned}\text{PBB} \;=\; & \bigcup_{r=0,1} \Big( \big\{ i : D_i(T) = P, R_i(C) = r, R_i(T) = r, S_i(C) = B, S_i(T) = B \big\} \cup \\ & \qquad \big\{ i : D_i(T) = P, R_i(C) = r, R_i(T) = 1 - r, S_i(C) = B, S_i(T) = B \big\} \Big).\end{aligned}$$

This means that the presence of non response implies that our strata comprise different types of people, therefore ignoring the missingness of outcomes would not lead to properly defined causal estimands. In addition, we know that when compliance is imperfect and outcomes are not observed for all units, an analysis based only on complete observations can lead to biased causal estimands (Frangakis and Rubin, 1999). To address these complications, we propose a new missing data model, which allows us to properly estimate the causal effects of interest using the classification of participants given above. 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 $(D_i(T), S_i(C), S_i(T))$.

As stated above, by definition, the quantity $D_i(T)$ is fixed for individual $i$, and therefore it is a covariate, the true compliance status covariate (Angrist, Imbens, and Rubin, 1996; Rubin, 1998), though it is only partially observed in the sample. The observed posttreatment compliance behavior, $D_i^{\text{obs}}$, is completely determined by the values of the covariate $D_i(T)$ and the assignment $Z_i^{\text{obs}}$. Naive attempts to condition on $D_i^{\text{obs}}$, the observed treatment received, generally, lead to biased conclusions because $D_i^{\text{obs}}$ is not a true covariate. Likewise, the observed BSE practice indicator $S_i^{\text{obs}}$ encodes characteristics of participants $i$ as well as the treatment assignment $Z_i^{\text{obs}}$, so it is not a true covariate. The pair of potential BSE practice indicator $(S_i(C), S_i(T))$, however, is not affected by treatment assignment $Z_i^{\text{obs}}$, so it only reflects characteristics of subject $i$, and can be regarded as a covariate.

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. If such adjustment 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. Thus there is exchangeability within each treatment arm for the group who attends the self-exam course and practices BSE, the group who attends the self-exams course but does not practice BSE, the group who does not attend the self-exam course and practices BSE, and the group who neither attends the self-exam course nor practices BSE. In other words, it is assumed that conditional on pretreatment variables, in each treatment arm, which principal stratum a person belongs to is equivalent to the outcome from choosing at random a ball in a large urn filled with eight identical balls numbered from 1 to 8. 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.

Principal stratification gives us a formal perspective on why a standard direct comparison between the distributions

$$\Pr\big(Y_i^{\text{obs}} \mid Z_i^{\text{obs}} = C, S_i^{\text{obs}} = B, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i\big)$$

and

$$\Pr\big(Y_i^{\text{obs}} \mid Z_i^{\text{obs}} = T, S_i^{\text{obs}} = B, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i\big)$$

which compares quality outcomes under standard versus new treatment for subjects who practices BSE, given the pretreatment variables is misleading for inferences about the causal effects on quality of BSE practice. The reason is that the groups $\{i : Z_i^{\text{obs}} = C, S_i^{\text{obs}} = B, \mathbf{X}_i^{\text{obs}} = \mathbf{x}\}$ and $\{i : Z_i^{\text{obs}} = T, S_i^{\text{obs}} = B, \mathbf{X}_i^{\text{obs}} = \mathbf{x}\}$ could involve different combination of principal strata, and thus might be different groups of people. For example, if either PbB group or pbB group exists, within cells defined by pretreatment covariates, the group of subjects who get posttreatment value $S_i^{\text{obs}} = B$ under control is a combination of the principal strata PBB, PBb pBB, and pBb; in contrast the group of subjects who get posttreatment value $S_i^{\text{obs}} = B$ under treatment is a mixture of the principal strata PBB, PbB pBB, and pbB.

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. Recall that principal strata 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 than standard conditions.

In our study, the most meaningful inferences about the effect of assignment on $Y$ can be drawn for the PBB group and the pBB group, since $Y_i(C)$ and $Y_i(T)$ are both clearly defined only for these groups. Thus, the primary causal effects on quality of BSE practice are formally defined by

$$\text{ITT(PBB)} = \text{E}\big(Y_i(T) - Y_i(T) \mid i \in \text{PBB}\big)$$

and

$$\text{ITT(pBB)} = \text{E}\big(Y_i(T) - Y_i(T) \mid i \in \text{pBB}\big),$$

that is, the ITT effects of $Z$ on $Y$ for compliers and never-takers, respectively, who would practice BSE under both treatment arms. These estimands focus on the causal effect of assignment of treatment rather than the causal effect of receipt of treatment. Here, we are also interesting in the causal effect of treatment. 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, so we focus on the ITT effect for the PBB group. Even for this subpopulation, however, inferring causal effects is controversial. We will discuss this further in section 7.

# 6 Observed Gruops

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 behavior, though we cannot observe the value of the BSE practice indicator under the unassigned standard treatment. Therefore, in the treatment arm, we can observe the following groups:

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

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

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

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

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

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

As the compliance status is only observed for women who are assigned to the new treatment, the observed group of women assigned to control in general comprises a mixture of compliers ($D_i(T) = P$) and never-takers ($D_i(T) = p$). Therefore, what we can observe in the control arm are the following three groups:

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 (?=data missing).

| Observed Group | | | | | Latent Group | | | |
|---|---|---|---|---|---|---|---|---|
| OBS($Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}}$) | $Z_i^{\text{obs}}$ | $D_i^{\text{obs}}$ | $R_i^{\text{obs}}$ | $S_i^{\text{obs}}$ | $G_i$ | | | |
| 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<br>pBB | PbB<br>pbB | PBb<br>pBb | Pbb<br>pbb |

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

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

OBS($C, p, 0, ?$) $= \{i : Z_i^{\text{obs}} = C, D_i^{\text{obs}} = p, R_i^{\text{obs}} = 0, S_i^{\text{obs}} =?\}$ : 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. Let $G_i$ represent the latent principal stratum indicator for subject $i$. The $N$-dimensional vector of $G_i$'s will be denoted by $\mathbf{G}$. If all eight principal strata exist, that is, if $\pi(g) > 0$, for each $g \in \{\mathrm{PBB}, \mathrm{PbB}, \mathrm{PBb}, \mathrm{Pbb}, \mathrm{pBB}, \mathrm{pbB}, \mathrm{pBb}, \mathrm{pbb}\}$, each observed group $\mathrm{OBS}(Z^{\mathrm{obs}}, D^{\mathrm{obs}}, R^{\mathrm{obs}}, S^{\mathrm{obs}})$ would be a mixture of two or more principal strata. For example, $\mathrm{OBS}(T, P, 1, B)$ would be a mixture of the PBB group and the PbB group, and $\mathrm{OBS}(C, p, 1, B)$ 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 3.

# 7 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.

## 7.1 SUTVA

A standard assumption made in causal analysis is the Stable Unit Treatment Value Assumption (SUTVA), formalized with potential outcomes by Rubin (1978a, 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 program to those performed in our experiment. Generalizability of results to other methods for teaching breast self-exam techniques would have to be judged separately.

## 7.2 Ignorability of Treatment Assignment

The study design of the Faenza randomized trial implies

Assumption 1. (Ignorability of Treatment Assignment)

$$\begin{aligned}\Pr\bigl(Z_i \mid D_i(T), S_i(C), S_i(T), R_i(C), R_i(T), Y_i(C), Y_i(T), \mathbf{X}_i^{\text{obs}}, \theta\bigr)\\ = \Pr\bigl(Z_i \mid \mathbf{X}_i^{\text{obs}}, \theta\bigr) = \Pr\bigl(Z_i \mid \mathbf{X}_i^{\text{obs}}\bigr),\end{aligned}$$

where $\theta$ is generic notation for the parameters governing the distribution of all the variables. There is no dependence on $\theta$ 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.

## 7.3 Monotonicity Assumptions

We impose two monotonicity assumptions which rule out the existence of two principal strata. These assumptions are based on the idea that the treatment or the control could be at least neutral on BSE practice rate for specific subpopulations.

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 two assumptions imply that the BSE practice indicator as function of of the assignment indicator, $\mathrm{I}\{Z_i = T\}$, is monotonous nondecreasing within the subpopulations of compliers, and monotonous noncreasing within the subpopulations of never-takers.

The monotonicity assumptions, 2 and 3, 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 different sign. In such analyses, we focused on the ITT effects on BSE practice without considering the quality of self-exam execution, and we investigated the consequences of an econometric exclusion restriction, that rules out for the subpopulation of never-takers, for whom there is no effect of assignment on receipt of treatment, any systematic effect of assignment on BSE practice. As we can see in Table 4, our analyses suggested that this econometric exclusion restriction could be questionable for the Faenza experiment. We did not find a strong evidence against this assumption; in any case the course did not seem to have a significant effect on BSE practice for compliers. However, relaxing the exclusion restriction led to more plausible ITT effects: we found a small and not much significant positive for compliers, and a quite strong, though insignificant, negative effect for never-takers. In addition, examining the joint distribution of the two ITT effects, they appeared somewhat negatively correlated (see Figure 1).

Table 4: Intention-To-Treat analysis on BSE practice using weakly identified models$^{(*)}$ - Summary statistics of posterior distribution.

| | BSE practice | | | |
|---|---|---|---|---|
| | Excl. Rest. Never-Takers | | | |
| Estimand | Yes | | No | |
| | Mean | s.d. | Mean | s.d. |
| ITT for Compliers | -0.040 | (0.050) | 0.058 | (0.117) |
| ITT for Never-Takers | 0 | 0 | -0.179 | (0.228) |
| Global ITT | -0.022 | (0.028) | -0.047 | (0.048) |

$(^{*})$ There are two key feature of these models: we imposed the missing data model developed by Mealli et al. (2004), which bases on two assumptions, namely, “latent ignorability”, and “response exclusion restriction for compliers”, and we used a proper prior distribution with a simple coniugate form.

We could also formulate the two monotonicity assumptions the other way around, that is, we could assume that there are nor PbB group or pBb group, but these assumptions seem be very unlikely in our experimental setting.

In our study, the monotonicity assumption seems more plausible for compliers than for never-takers. Compliers are women who follow the protocol in their assignment. If the PBb group exists, by definion, the ITT effect on BSE practice for this subpopulation of compliers is negative, but it is difficult to understand why a program designed to encourage BSE would have a negative effect among compliers. Our monotonicity assumption for compliers implies that there is a non negative ITT effect on BSE practice for all the compliers.

Never-takers, on the other hand, would not attend the BSE teaching course when assigned to the new treatment. Our preliminary analyses suggested that probably there is a some effect of the assignment on BSE practice for this type of units; however, it is hard to make reasonable assumptions on such ITT effect. In other words, it is not implausible that there are pbB groups, that is, never-takers who

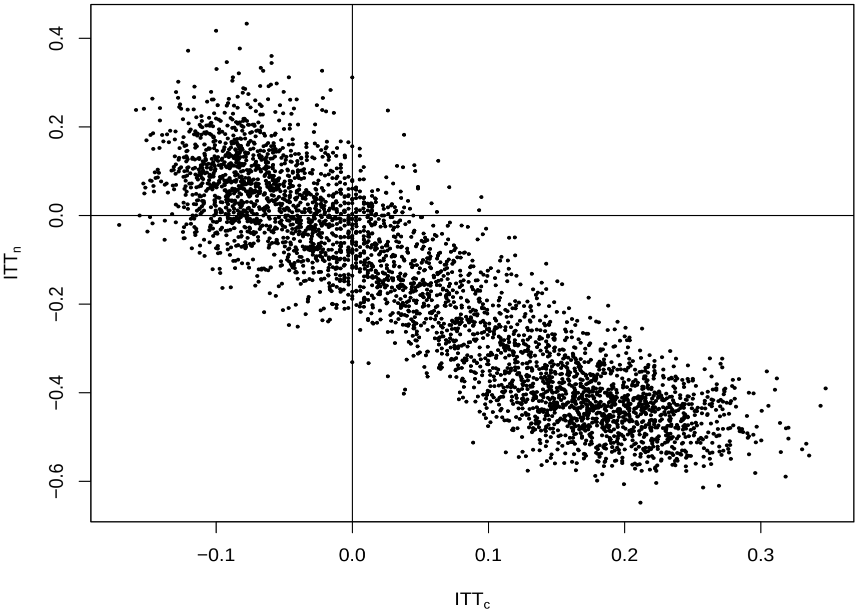

Figure 1: Simulation scatterplot of the joint posterior distribution of ITT effects on BSE practice for compliers (ITT$_c$) and never-takers (ITT$_n$).

would practice BSE under treatment but would not practice it under control.

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 10.

As we can see from Table 5, under the two monotonicity assumptions, 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. Despite the two monotonicity assumptions, the most of observed groups are still a mixture of two or more principal strata, and so their observed

Table 5: 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 for compliers (assumption 2) and never-takers (assumption 3).(?=data missing).

| Observed Group | | | | | | Latent Group | |
|---|---|---|---|---|---|---|---|
| OBS($Z^{\text{obs}}, D^{\text{obs}}, R^{\text{obs}}, S^{\text{obs}}$) | $Z_i^{\text{obs}}$ | $D_i^{\text{obs}}$ | $R_i^{\text{obs}}$ | $S_i^{\text{obs}}$ | | $G_i$ | |
| 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<br>pBB | PbB<br>pBb | Pbb<br>pbb |

distributions are a mixture of two or more distributions.

Our monotonicity assumptions are analogous to that typically exploited in Instrumental Variables (IV) analysis. Similar to monotonicity assumption in Angrist et al. (1996), the two monotonicity assumptions here rule out specific “defier” groups. However, in Angrist et al. (1996), the monotonicity assumption was imposed on the compliance behavior in a experimental setting where all the units had access to the active treatment. Since it is difficult to think of someone who would always go against treatment assignment, that is, taking the control when assigned to the treatment, while taking the treatment when assigned to control, the monotonicity of compliance appears very plausible in many applications of encouragement designs. In our application, defiers with respect to assignment do not exist by definition, as women assigned to the standard treatment had not access to the BSE training course. In our study, the monotonicity assumptions concern the BSE practice behavior and appear less based because it is plausible that all the principal strata exist. We will focus on assessing the sensitivity on inference to these assumptions and evaluating the influence of the model using Bayesian posterior predictive checks (section 10).

## 7.4 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 distributions 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)

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

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 $Y_i(z)$.

Assumption 4 is closely related to "exclusion restriction" assumptions in the traditional instrumental variables approach (Durbin, 1954; Goldberger, 1972; Angrist et al., 1996), which concern noncompliers. In our study, the exclusion restriction applies to a specific subpopulations of noncompliers, never-takers who would practice BSE under both treatment and control.

As stated previously, the intention-to-treat comparison on BSE quality makes sense only for women who practice the self exams under both assignments. In our study, this group of women is a mixture of PBB and pBB principal strata; therefore, the ITT effect on $Y$ for women practicing BSE under both assignments, say ITT(BB), can be defined as the weighted average of ITT effects across these two groups:

$$\text{ITT(BB)} = \pi(\text{PBB}) \times \text{ITT(PBB)} + \pi(\text{pBB}) \times \text{ITT(pBB)}.$$

The quality outcome exclusion restriction for the pBB group implies that ITT(pBB) = 0, and thus it formalizes the notion that any ITT effect of assignment on quality of BSE practice should be mediated by an effect of assignment on the treatment received. Moreover, the PBB principal stratum is the only group of women who would attend the enhanced BSE teaching course if and only if encouraged. This implies that, in our experimental setting, the ITT effect for compliers who would practice BSE under both treatment arms, ITT(PBB), is the only intention-to-treat 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 the pBB groups there is no direct effect of the assignment on quality outcome, it seems plausible to attribute the effect of assignment for the PBB group to the effect of the receipt the treatment.

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 another intermediate outcome. 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)

$$\begin{aligned}&\Pr\big(R_i(T) = 1 \mid D_i(T) = P, S_i(C), S_i(T), \mathbf{X}_i\big)\\ &\quad = \Pr\big(R_i(C) = 1 \mid D_i(T) = P, S_i(C), S_i(T), \mathbf{X}_i\big).\end{aligned}$$

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.

### 7.5 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*

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

*(b) when assigned new treatment*

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

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. Recall that $Y_i(z) = *$ almost sure for women who would not practice BSE when assigned to treatment $z$, so we do not consider LI when BSE quality is not defined on the usual sample space.

Since our study is a randomized experiment, LI implies that

$$\begin{aligned}&\Pr\big(R_i, Y_i \mid D_i(T), S_i(Z_i) = B, S_i(Z_i^c), Z_i, \mathbf{X}_i\big)\\&= \quad \Pr\big(R_i \mid D_i(T), S_i(Z_i) = B, S_i(Z_i^c), Z_i, \mathbf{X}_i\big) \Pr\big(Y_i \mid D_i(T), S_i(Z_i) = B, S_i(Z_i^c), Z_i, \mathbf{X}_i\big),\end{aligned}$$

where $Z_i^c = C$ if $Z_i = T$ and $Z_i^c = T$ if $Z_i = C$. This means that, if principal strata were observed for all subjects and the parameters of missing data process are distinct from those of the outcome distribution, the missing data process would be ignorable. But because principal strata are only partially observed, the missing data mechanism is in fact nonignorable, also under the response exclusion restriction for compliers defined in section 7.4.

We make the latent ignorability assumption here, first because it is more plausible than the assumption of standard ignorability (SI) (Rubin, 1978a; Little and Rubin, 1987) and second, because making it leads to different likelihood inference.

LI is more plausible than SI to the extend that it provides a closer approximation to the missing data mechanism. The intuition behind this 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 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. Our LI assumption arises from the combination of two ideas. First, it is widely supported by the literature for noncompliance (see, for example, The Coronary Drug Project Research Group, 1980) that the compliance status $D(T)$ is a strong factor in outcome missingness, even when other covariates are included in the model, therefore assuming that compliers and never-takers have a different response behavior appears quite plausible. Explorations of raw data from our study across individuals with known compliance status provide empirical support that $D(T)$ may be related to the willingness to respond of the subjects (see Table 1). Second, it seems likely that within subpopulations defined by the compliance covariate $D(T)$, response rates significantly differ among women who would practice BSE under either assignment, who would practice BSE only if assigned to control but not if assigned to treatment, who would practice BSE if assigned to treatment but not if assigned to control, and who not practice BSE regardless of assignment.

Regarding improved estimation, when LI and the preceding structural assumptions hold but the likelihood is constructed assuming SI, the posterior distribution of the underlying causal effects converges to the truth with increasing sample size if the additional assumption is made that, within subclasses of subjects with similar observed variables, the partially missing principal stratum $G$ is not associated with potential outcomes. However, as noted previously, this assumption is not plausible.

## 7.6 Additional Assumptions

For never-takers, the treatment actually received would be the same no matter what their treatment assignment, namely, these women would not attend the BSE teaching course in any case. Since our study lacks a comparable set-up in the standard treatment, 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 easily lower their subsequent response probability. In comparison to those never-takers receiving the standard treatment, their explicit refusal to comply with the assigned (active) treatment may plausible induce them to refuse to respond in the posttest questionnaire as well. This suggests that never-takers could be affected in their response behavior by the treatment assignment. On another hand, 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, so it might be reasonable to assume that the willingness to respond of these women is not related to latent covariate defined by the vector $(S(C), S(T))$. Thus assumption 7 comes into play

Assumption 7.

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

This assumption implies that never-takers have the same response behavior irrespective of their BSE practice behavior, given pretreatment variables conditional on the treatment assignment, thus that

$$\begin{aligned}\Pr\big(S_i(C), S_i(T), \mid D_i(T) = p, \mathbf{X}_i\big) \Pr\big(R_i \mid Z_i, D_i(T) = p, S_i(C), S_i(T), \mathbf{X}_i\big)\\ = \Pr\big(S_i(C), S_i(T) \mid D_i(T) = p, \mathbf{X}_i\big) \Pr\big(R_i \mid Z_i, D_i(T) = p, \mathbf{X}_i\big).\end{aligned}$$

Assumption 7 is not directly testable and could be questionable. In our experimental context, it is plausible that the response rates differ among the subpopulations of never-takers defined by the partially unobserved covariate $(S(C), S(T))$. Assessment of assumption 7 is thus crucial for any sensible inference based on it. In particular, we should assess if this assumption is only a way to simplify the inferential task, given the relatively small sample size, or it could be actually plausible.

We will discuss this further in section 10.

Finally, we consider two stochastic dominance assumptions which rank the self-examination skills among compliers. These assumptions are an extension 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 treatment comparisons should be only adjusted for one posttreatment variable.

We distinguish two components of our stochastic dominance assumption. In the first component we assume that when assignment 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)

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

In the second component of the stochastic dominance assumption we assume that when assigned to standard 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 9. (Stochastic Dominance of the PBB Group over the PBb Group)

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

These two assumptions formalize the notion that the PBB group is more capable or has higher motivation than either the PbB group or the PBb group; the PBB group would practice BSE under either treatment arm, whereas the PbB group or the PBb group would practice the self exam under only one treatment arm. Since we impose the monotonicity assumption for compliers, the second component of the stochastic dominance assumption is obviously satisfied, being $\pi(\text{PBb}) = 0$; so we actually consider only assumption 8. 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]

In the next section, we present the model we used to obtain our inferential results. Recall that a model has the role of assisting, not creating, inference in the sense that results should be robust to different parametric specifications (Frangakis and Rubin, 2001 - rejoinder).

## 8 Parametric Pattern Mixture Model

Generally speaking, constrained estimation of separate analyses within missing data patterns is the motivation behind pattern mixture modeling. A variety of authors use pattern mixture model approaches to missing data including Little (1993, 1996), Rubin (1978b), Glynn, Laird, and Rubin (1993). Typically pattern mixture models partition the data with respect to the missingness of the variables. Here, 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 $X$. This represents a partial pattern mixture model approach. One argument for this approach is that it focuses the model on the quantities of interest, in such a way that parametric specifications for the marginal distributions of $X$ can be ignored.

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

[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).

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

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

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

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 such a way, we have to model four things: the conditional distribution of the compliance variable $D(T)$ given the pretreatment variables $\mathbf{X}^{\text{obs}}$; the joint conditional distribution of the intermediate potential outcomes $S(C)$ and $S(T)$ given $\mathbf{X}^{\text{obs}}$ and $D(T)$; the conditional distribution of the potential response indicator $R$ given the covariates $\mathbf{X}^{\text{obs}}$ and the principal stratum variable $G$ defined by the vector $(D(T), S(C), S(T))$, and the conditional distribution of the quality potential outcome $Y$, also given $\mathbf{X}^{\text{obs}}$, and $G$.

In our experimental setting the compliance covariate is dichotomous, therefore we assume that its distribution has a logistic regression form:

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

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 arms. 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:

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

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 $\gamma_P^{\text{BB}}$ 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:

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

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:

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

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

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

and

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

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, 1978a).

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

$$\begin{aligned}\pi_{iz}^{R}(g) &= \Pr\bigl(R_i = 1 \mid Z_i = z_i, G_i = g, \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \beta_z(g)\bigr) \\ &= \frac{\exp(\beta_{0z}(g) + \beta_{1z}(g)'\mathbf{x}_i)}{1+\exp(\beta_{0z}(g) + \beta_{1z}(g)'\mathbf{x}_i)},\end{aligned}$$

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.

Consider 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 $\mathbf{G}$. Under the monotonicity assumptions 2 and 3, and the latent ignorability assumption 6 the complete-data likelihood function is

$$
\begin{aligned}
&\mathcal{L}_{\text{COMP}}\big(\theta \mid \mathbf{Z}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{R}^{\text{obs}}, \mathbf{S}^{\text{obs}}, \mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}}, \mathbf{G}\big) = \\
&\quad \prod_{i \in \text{PBB}} \pi_i^P \pi_i^{\text{BB}}(P) \\
&\qquad \Bigg( \Big( \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\}} \\
&\qquad\quad \Big(1 - \pi_{iC}^R(\text{PBB})\Big)^{\text{I}\{R_i^{\text{obs}}=0\}} \Bigg)^{\text{I}\{Z_i^{\text{obs}}=C\}} \\
&\qquad \Bigg( \Big( \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\}} \Big)^{\text{I}\{R_i^{\text{obs}}=1\}} \\
&\qquad\quad \Big(1 - \pi_{iT}^R(\text{PBB})\Big)^{\text{I}\{R_i^{\text{obs}}=0\}} \Bigg)^{\text{I}\{Z_i^{\text{obs}}=T\}} \\
&\times \quad \prod_{i \in \text{PbB}} \pi_i^P \pi_i^{\text{bB}}(P) \Bigg( \Big(\pi_{iC}^R(\text{PbB})\Big)^{\text{I}\{R_i^{\text{obs}}=1\}} \Big(1 - \pi_{iC}^R(\text{pbB})\Big)^{\text{I}\{R_i^{\text{obs}}=0\}} \Bigg)^{\text{I}\{Z_i^{\text{obs}}=C\}} \\
&\qquad \Bigg( \Big( \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\}} \Big)^{\text{I}\{R_i^{\text{obs}}=1\}} \\
&\qquad\quad \Big(1 - \pi_{iT}^R(\text{PbB})\Big)^{\text{I}\{R_i^{\text{obs}}=0\}} \Bigg)^{\text{I}\{Z_i^{\text{obs}}=T\}} \\
&\times \quad \prod_{i \in \text{Pbb}} \pi_i^P \pi_i^{\text{bb}}(P) \Bigg( \Big(\pi_{iC}^R(\text{Pbb})\Big)^{\text{I}\{R_i^{\text{obs}}=1\}} \Big(1 - \pi_{iC}^R(\text{Pbb})\Big)^{\text{I}\{R_i^{\text{obs}}=0\}} \Bigg)^{\text{I}\{Z_i^{\text{obs}}=C\}} \\
&\qquad \Bigg( \Big(\pi_{iT}^R(\text{Pbb})\Big)^{\text{I}\{R_i^{\text{obs}}=1\}} \Big(1 - \pi_{iT}^R(\text{Pbb})\Big)^{\text{I}\{R_i^{\text{obs}}=0\}} \Bigg)^{\text{I}\{Z_i^{\text{obs}}=T\}}
\end{aligned}
$$

$$
\begin{aligned}
\times & \prod_{i \in \mathrm{pBB}} (1-\pi_i^P)\pi_i^{\mathrm{BB}}(p) \\
& \quad \left( \left( \pi_{iC}^R(\mathrm{pBB}) \big(\pi_{iC}^Y(\mathrm{pBB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=H\}} \big(1-\pi_{iC}^Y(\mathrm{pBB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=L\}} \right)^{\mathrm{I}\{R_i^{\mathrm{obs}}=1\}} \right. \\
& \quad \quad \left. \Big(1-\pi_{iC}^R(\mathrm{pBB})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=0\}} \right)^{\mathrm{I}\{Z_i^{\mathrm{obs}}=C\}} \\
& \quad \left( \left( \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\}} \right)^{\mathrm{I}\{R_i^{\mathrm{obs}}=1\}} \right. \\
& \quad \quad \left. \Big(1-\pi_{iT}^R(\mathrm{pBB})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=0\}} \right)^{\mathrm{I}\{Z_i^{\mathrm{obs}}=T\}} \\
\times & \prod_{i \in \mathrm{pBb}} (1-\pi_i^P)\pi_i^{\mathrm{Bb}}(p) \\
& \quad \left( \left( \pi_{iC}^R(\mathrm{pBb}) \big(\pi_{iC}^Y(\mathrm{pBb})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=H\}} \big(1-\pi_{iC}^Y(\mathrm{pBb})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=L\}} \right)^{\mathrm{I}\{R_i^{\mathrm{obs}}=1\}} \right. \\
& \quad \quad \left. \Big(1-\pi_{iC}^R(\mathrm{pBb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=0\}} \right)^{\mathrm{I}\{Z_i^{\mathrm{obs}}=C\}} \\
& \quad \left( \Big(\pi_{iT}^R(\mathrm{pBb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=1\}} \Big(1-\pi_{iT}^R(\mathrm{pBb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=0\}} \right)^{\mathrm{I}\{Z_i^{\mathrm{obs}}=T\}} \\
\times & \prod_{i \in \mathrm{pbb}} (1-\pi_i^P)\pi_i^{\mathrm{bb}}(p) \left( \Big(\pi_{iC}^R(\mathrm{pbb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=1\}} \Big(1-\pi_{iC}^R(\mathrm{pbb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=0\}} \right)^{\mathrm{I}\{Z_i^{\mathrm{obs}}=C\}} \\
& \quad \left( \Big(\pi_{iT}^R(\mathrm{pbb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=1\}} \Big(1-\pi_{iT}^R(\mathrm{pbb})\Big)^{\mathrm{I}\{R_i^{\mathrm{obs}}=0\}} \right)^{\mathrm{I}\{Z_i^{\mathrm{obs}}=T\}} .
\end{aligned}
$$

The complete-case likelihood function has a simple form with twenty-five factors: one for each of the six quality outcome distributions, one for each of twelve response outcome submodels, and six plus one involving the parameters of the principal stratum distribution. Under the response exclusion restriction for compliers, $\pi_{iC}^R(\mathrm{PBB}) = \pi_{iT}^R(\mathrm{PBB})$, $\pi_{iC}^R(\mathrm{PbB}) = \pi_{iT}^R(\mathrm{PbB})$, and $\pi_{iC}^R(\mathrm{Pbb}) = \pi_{iT}^R(\mathrm{Pbb})$; and under the quality outcome exclusion restriction for the pBB group, $\pi_{iC}^Y(\mathrm{pBB}) = \pi_{iT}^Y(\mathrm{pBB})$. In addition, assumption 7 implies that $\pi_{iz}^R(\mathrm{pBB}) = \pi_{iz}^R(\mathrm{pBb}) = \pi_{iz}^R(\mathrm{pbb})$, for $z = C, T$. Therefore, under our structural assumptions described in section 7, the complete-data likelihood function consists of only seventeen distinct factors. If we also impose the stochastic dominance assumption of the PBB group over the PbB group (assumption 8), the complete-data likelihood function is the same as above if $\pi_{iT}^{Y}(\text{PBB}) \geq \pi_{iT}^{Y}(\text{PbB})$ and equal to 0 otherwise.

For inference based on the observed data, we cannot work directly with this complete-data likelihood function, because we do not observe the principal stratum $G_i$ of each unit. However, we can exploit the complete-data likelihood function by using missing data methods such as the EM algorithm (Dempster et al., 1977), and the Data Augmentation (DA) algorithm (Tanner an Wong, 1987). In Appendix A, we describe the numerical methods used to generate the inference reported in the next section.

As shown in Table 5, under the monotonicity assumptions 2 and 3, and the latent ignorability assumption 6, there are nine possible patterns of missing data and observed data in $(D_i(T), S_i(C), S_i(T), R_i, Y_i)$, corresponding to the nine possible values for $(Z_i^{\text{obs}}, D_i^{\text{obs}}, R_i^{\text{obs}}, S_i^{\text{obs}})$. We can then write the actual (observed) likelihood function in terms of the observed data as

$$
\begin{aligned}
&\mathcal{L}_{\text{OBS}}\big(\theta \mid \mathbf{Z}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{R}^{\text{obs}}, \mathbf{S}^{\text{obs}}, \mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}}\big) = \\
&\quad \prod_{i \in \text{OBS}(T,P,1,B)} \pi_i^P \Bigg( \pi_i^{\text{BB}}(P)\pi_{iT}^{R}(\text{PBB})\big(\pi_{iT}^{Y}(\text{PBB})\big)^{\text{I}\{Y_i^{\text{obs}}=H\}}\big(1-\pi_{iT}^{Y}(\text{PBB})\big)^{\text{I}\{Y_i^{\text{obs}}=L\}} + \\
&\qquad\qquad \pi_i^{\text{bB}}(P)\pi_{iT}^{R}(\text{PbB})\big(\pi_{iT}^{Y}(\text{PbB})\big)^{\text{I}\{Y_i^{\text{obs}}=H\}}\big(1-\pi_{iT}^{Y}(\text{PbB})\big)^{\text{I}\{Y_i^{\text{obs}}=L\}} \Bigg) \\
&\times \prod_{i \in \text{OBS}(T,P,1,b)} \pi_i^P \pi_i^{\text{bb}}(P)\pi_{iT}^{R}(\text{Pbb}) \\
&\times \prod_{i \in \text{OBS}(T,P,0,?)} \pi_i^P \Bigg( \pi_i^{\text{BB}}(P)\Big(1-\pi_{iT}^{R}(\text{PBB})\Big) + \\
&\qquad\qquad \pi_i^{\text{bB}}(P)\Big(1-\pi_{iT}^{R}(\text{PbB})\Big) + \pi_i^{\text{bb}}(P)\Big(1-\pi_{iT}^{R}(\text{Pbb})\Big) \Bigg) \\
&\times \prod_{i \in \text{OBS}(T,p,1,B)} (1-\pi_i^P)\pi_i^{\text{BB}}(p)\pi_{iT}^{R}(\text{pBB})\big(\pi_{iT}^{Y}(\text{pBB})\big)^{\text{I}\{Y_i^{\text{obs}}=H\}}\big(1-\pi_{iT}^{Y}(\text{pBB})\big)^{\text{I}\{Y_i^{\text{obs}}=L\}} \\
&\times \prod_{i \in \text{OBS}(T,p,1,b)} (1-\pi_i^P)\Bigg( \pi_i^{\text{Bb}}(p)\pi_{iT}^{R}(\text{pBb}) + \pi_i^{\text{bb}}(p)\pi_{iT}^{R}(\text{pbb}) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
\times \prod_{i \in \mathrm{OBS}(T,p,0,?)} & (1-\pi_i^P)\Bigg(\pi_i^{\mathrm{BB}}(p)\Big(1-\pi_{iT}^R(\mathrm{pBB})\Big) + \\
& \qquad \pi_i^{\mathrm{Bb}}(p)\Big(1-\pi_{iT}^R(\mathrm{pBb})\Big) + \pi_i^{\mathrm{bb}}(p)\Big(1-\pi_{iT}^R(\mathrm{pbb})\Big)\Bigg) \\
\times \prod_{i \in \mathrm{OBS}(C,p,1,B)} & \Bigg(\pi_i^P \pi_i^{\mathrm{BB}}(P)\pi_{iC}^R(\mathrm{PBB})\big(\pi_{iC}^Y(\mathrm{PBB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=H\}}\big(1-\pi_{iC}^Y(\mathrm{PBB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=L\}} + \\
& (1-\pi_i^P)\pi_i^{\mathrm{BB}}(p)\pi_{iC}^R(\mathrm{pBB})\big(\pi_{iC}^Y(\mathrm{pBB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=H\}}\big(1-\pi_{iC}^Y(\mathrm{pBB})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=L\}} + \\
& (1-\pi_i^P)\pi_i^{\mathrm{Bb}}(p)\pi_{iC}^R(\mathrm{pBb})\big(\pi_{iC}^Y(\mathrm{pBb})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=H\}}\big(1-\pi_{iC}^Y(\mathrm{pBb})\big)^{\mathrm{I}\{Y_i^{\mathrm{obs}}=L\}}\Bigg) \\
\times \prod_{i \in \mathrm{OBS}(C,p,1,b)} & \Bigg(\pi_i^P \pi_i^{\mathrm{bB}}(P)\pi_{iC}^R(\mathrm{PbB}) + \pi_i^P \pi_i^{\mathrm{bb}}(P)\pi_{iC}^R(\mathrm{Pbb}) + (1-\pi_i^P)\pi_i^{\mathrm{bb}}(p)\pi_{iC}^R(\mathrm{pbb})\Bigg) \\
\times \prod_{i \in \mathrm{OBS}(C,p,0,?)} & \Bigg(\pi_i^P \pi_i^{\mathrm{BB}}(P)\big(1-\pi_{iC}^R(\mathrm{PBB})\big) + \pi_i^P \pi_i^{\mathrm{bB}}(P)\big(1-\pi_{iC}^R(\mathrm{PbB})\big) + \\
& \pi_i^P \pi_i^{\mathrm{bb}}(P)\big(1-\pi_{iC}^R(\mathrm{Pbb})\big) + (1-\pi_i^P)\pi_i^{\mathrm{BB}}(p)\big(1-\pi_{iC}^R(\mathrm{pBB})\big) + \\
& (1-\pi_i^P)\pi_i^{\mathrm{Bb}}(p)\big(1-\pi_{iC}^R(\mathrm{pBb})\big) + (1-\pi_i^P)\pi_i^{\mathrm{bb}}(p)\big(1-\pi_{iC}^R(\mathrm{pbb})\big)\Bigg).
\end{aligned}
$$

The first three factors in the likelihood represent the contribution of compliers assigned to the treatment, including both respondents and nonrespondents. The second three factors represent the contribution for never-takers assigned to the treatment, including respondents and nonrespondents. The last three factors represent the contribution to the likelihood function for those assigned to the standard treatment, and they include both compliers and never-takers. Each of these factors, except the second and the fourth ones, includes women with possible different attitudes towards BSE practice; therefore each likelihood contribution consists of averages over the distribution of principal strata. Because the observed-data likelihood function has this mixture structure over a large amount of missing data, 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 $g \in (\mathrm{PBB}, \mathrm{PbB}, \mathrm{Pbb}, \mathrm{pBB}, \mathrm{pBb},\ \mathrm{pbb})$; 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 value of the pretreatment variables, $\mathbf{X}_i^{\text{obs}}$. More formally, the prior distribution is proportional to

$$p(\theta) \propto \prod_{i=1}^{N} \times \prod_{g \in \{\text{PBB},\text{PbB},\text{Pbb},\text{pBB},\text{pBb},\text{pbb}\}} \times \prod_{z=C,T} \prod_{r=0,1} \prod_{y=L,H}$$

$$\left[ \pi_i(g) \left( \pi_{iz}^{R}(g) \left( \left(\pi_{iz}^{Y}(g)\right)^{\mathrm{I}\{y=H\}} \left(1 - \pi_{iz}^{Y}(g)\right)^{\mathrm{I}\{y=L\}} \right)^{\mathrm{I}\{S(z)=B\}} \right)^{r} \left(1 - \pi_{iz}^{R}(g)\right)^{(1-r)} \right]^{\frac{3}{k(g)N}} .$$

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.

Under assumption 5 and 4 (the response exclusion restriction for compliers and the quality outcome exclusion restriction for the pBB group, respectively), assumption 7, and the prior equalities given above, the number of parameters reduces to 29. The simulated posterior distributions of these parameters are summarized in Table 7.

To demonstrate that our proper prior distribution does not lead to highly informative prior distribution for the estimands of interest, Table 6 presents summary statistics, obtained by the methods described in Appendix A, of the marginal prior distributions of the estimands of primary interest: the ITT effect on BSE practice

Table 6: Summary statistics: prior distribution.

| Estimand | Mean | s.d. | Percentiles | | | | |
|---|---|---|---|---|---|---|---|
| | | | 2.5% | 25% | 50% | 75% | 97.5% |
| CACE on BSE practice | 0.31 | 0.18 | 0.03 | 0.18 | 0.29 | 0.43 | 0.70 |
| NACE on BSE practice | -0.34 | 0.19 | -0.73 | -0.46 | -0.33 | -0.20 | -0.03 |
| ITT on BSE practice | 0.00 | 0.17 | -0.33 | -0.11 | -0.01 | 0.11 | 0.33 |
| CACE on BSE quality adjusted for BSE practice | -0.15 | 0.54 | -0.94 | -0.59 | -0.24 | 0.29 | 0.90 |
| ITT on BSE quality adjusted for BSE practice | -0.09 | 0.33 | -0.74 | -0.30 | -0.09 | 0.10 | 0.64 |

for compliers (CACE on BSE practice), the ITT effect on BSE practice for never-takers (NACE on BSE practice), and the overall ITT effect on BSE practice (ITT on BSE practice); the ITT effect on BSE quality for the PBB groups (CACE on BSE quality) and the overall ITT effect for women who would practice BSE under both assignments (ITT on BSE quality). The joint distributions of these ITT effects were obtained using the same computational techniques used to obtain the actual posterior distribution with the data. The comparison of the standard deviations in Table 6 for the ITT effects with the corresponding values in Tables 8 and 10, reported in the next sections, indicates that the prior distribution is relatively uninformative about quantities of interest.

Table 7: Posterior distribution for parameters.

| Estimand | Mean | s.d. | Percentiles | | | | |
|---|---|---|---|---|---|---|---|
| | | | 2.5% | 25% | 50% | 75% | 97.5% |
| $\alpha_0$ | -0.33 | 0.22 | -0.77 | -0.48 | -0.33 | -0.19 | 0.10 |
| $\alpha_{X_1}$ | 0.33 | 0.24 | -0.14 | 0.16 | 0.33 | 0.49 | 0.79 |
| $\alpha_{X_2}$ | 0.64 | 0.23 | 0.19 | 0.48 | 0.64 | 0.80 | 1.09 |
| $\gamma_0^{\text{bB}}(P)$ | -0.46 | 1.48 | -4.65 | -1.05 | -0.19 | 0.46 | 1.53 |
| $\gamma_{X_1}^{\text{bB}}(P)$ | -0.39 | 1.48 | -2.80 | -1.15 | -0.59 | 0.08 | 3.93 |
| $\gamma_{X_2}^{\text{bB}}(P)$ | -2.11 | 1.57 | -6.83 | -2.54 | -1.80 | -1.29 | 0.29 |
| $\gamma_0^{\text{bb}}(P)$ | -0.10 | 0.67 | -1.42 | -0.53 | -0.10 | 0.33 | 1.21 |
| $\gamma_{X_1}^{\text{bb}}(P)$ | -1.32 | 0.51 | -2.38 | -1.65 | -1.31 | -0.98 | -0.35 |
| $\gamma_{X_2}^{\text{bb}}(P)$ | -1.01 | 0.60 | -2.16 | -1.38 | -1.02 | -0.63 | 0.27 |
| $\gamma_0^{\text{Bb}}(p)$ | 1.89 | 1.21 | -0.90 | 1.25 | 1.97 | 2.65 | 4.12 |
| $\gamma_{X_1}^{\text{Bb}}(p)$ | -2.26 | 1.29 | -4.76 | -3.05 | -2.29 | -1.56 | 0.49 |
| $\gamma_{X_2}^{\text{Bb}}(p)$ | -3.27 | 1.43 | -6.73 | -4.01 | -3.09 | -2.31 | -1.07 |
| $\gamma_0^{\text{bb}}(p)$ | 1.33 | 1.29 | -1.61 | 0.62 | 1.46 | 2.15 | 3.75 |
| $\gamma_{X_1}^{\text{bb}}(p)$ | -4.00 | 1.29 | -6.98 | -4.70 | -3.87 | -3.12 | -1.82 |
| $\gamma_{X_2}^{\text{bb}}(p)$ | -0.81 | 1.24 | -3.30 | -1.59 | -0.89 | -0.05 | 1.68 |
| $\beta_{0\cdot}(\text{PBB})$ | 3.76 | 2.43 | 0.19 | 1.97 | 3.42 | 5.19 | 9.51 |
| $\beta_{0\cdot}(\text{PbB})$ | 0.66 | 2.17 | -3.55 | -0.44 | 0.53 | 1.65 | 5.55 |
| $\beta_{0\cdot}(\text{Pbb})$ | 0.53 | 1.26 | -1.14 | -0.30 | 0.25 | 1.04 | 3.74 |
| $\beta_{X_1\cdot}(P)$ | -0.26 | 0.81 | -1.68 | -0.78 | -0.28 | 0.22 | 1.24 |
| $\beta_{X_2\cdot}(P)$ | -0.02 | 0.87 | -2.01 | -0.47 | 0.02 | 0.54 | 1.51 |
| $\beta_{0C}(p)$ | 0.35 | 0.36 | -0.34 | 0.11 | 0.35 | 0.58 | 1.07 |
| $\beta_{0T}(p)$ | -0.30 | 0.28 | -0.86 | -0.48 | -0.30 | -0.11 | 0.23 |
| $\beta_{X_1\cdot}(p)$ | 0.18 | 0.30 | -0.43 | -0.03 | 0.18 | 0.38 | 0.77 |
| $\beta_{X_2\cdot}(p)$ | -0.27 | 0.29 | -0.85 | -0.46 | -0.27 | -0.07 | 0.31 |
| $\delta_{0C}(\text{PBB})$ | 0.88 | 0.38 | 0.24 | 0.63 | 0.85 | 1.09 | 1.67 |
| $\delta_{0T}(\text{PBB})$ | 4.52 | 7.06 | 1.06 | 1.48 | 1.83 | 2.66 | 32.37 |
| $\delta_{0T}(\text{PbB})$ | -1.02 | 1.67 | -4.44 | -1.97 | -1.02 | -0.02 | 2.26 |
| $\delta_{0\cdot}(\text{pBB})$ | -0.22 | 0.38 | -0.98 | -0.47 | -0.21 | 0.04 | 0.52 |
| $\delta_{0C}(\text{pBb})$ | -0.48 | 1.01 | -2.81 | -0.97 | -0.36 | 0.14 | 1.21 |

# 9 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 reports results for the ITT and CACE estimands on quality of self-exam execution and for outcome response rates. The reported estimands are not parameters of the models but functions of parameters and data.

## 9.1 Proportions in Principal Strata and Causal Effects on BSE Practice

Table 8 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 $\theta$, $D_i(T)$, and $(S_i(C), S_i(T))$ given $D_i(T)$, $i = 1, \ldots, N$, from the posterior distribution (see Appendix A); (2) calculate $\Pr(D_i(T) = P \mid \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \theta)$, and $\Pr(S_i(C), S_i(T) \mid D_i(T), \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \theta)$, for each subject based on the models in section 8; 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 $\Pr(S_i(C), S_i(T) \mid D_i(T); \theta)$.

Combining the marginal posterior distribution of the compliance status with the conditional posterior distributions of $(S(C), S(T))$ given the compliance status $D(T)$, we can easily obtain the posterior distributions of the estimands of the probability of being in a stratum defined by the vector $(D(T), S(C), S(T))$, $\Pr(D(T), S(C), S(T) \mid \theta)$. In particular, to draw from this distribution, we use the same steps (1) and (2) described above, and then calculate $\Pr(D_i(T), S_i(C), S_i(T) \mid \mathbf{X}_i^{\text{obs}} = \mathbf{x}_i; \theta)$ as product of the distributions in (2), and average the resulting values over all the subjects. The simulated posterior distributions of these estimands are

Table 8: 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% |
| $\pi^P$ | 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 |

summarized in Table 9.

The clearest pattern revealed by Tables 8 and 9 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. This pattern is reasonable. Compliers are women who attend the enhanced BSE course if so assigned. Such subjects appear to be very aware of the risk of breast cancer; in our analysis this is revealed by the positive logistic coefficients of the compliance submodel on knowledge of breast pathophysiology and prior BSE practice (see Table 7), which imply that compliers tend to have a good knowledge of breast pathophysiology and to be used to practice BSE. In contrast, never-takers

Table 9: Summary statistics of the posterior distributions of principal strata.

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

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. Given this attitude towards BSE practice, never-takers assigned to the new treatment arm might feel frustrated with the extra burden of attending the course and hence, other to not comply with their assignment, they were more likely to choose to not practice BSE anymore. Consequently, it could be reasonable to expect that the invitation to partecipate to the BSE teaching course had a negative effect on BSE practice for these women.

Under our model, the intention-to-treat effect on BSE practice for never-takers (NACE on $S$) is negative by construction. As we can easily show, the NACE on BSE practice estimand for individual $i$ is defined as

$$\mathrm{E}\big(S_i(T) - S_i(C) \mid D_i(T) = p; \theta\big) = \pi_i^{\mathrm{bB}}(p) - \pi_i^{\mathrm{Bb}}(p),$$

where $\pi_i^{\mathrm{bB}}(p) = \Pr(S_i(C) = b, S_i(T) = B \mid D_i(T) = p; \theta)$ is the probability of being a never-takers who would not practice BSE if assigned to control, but would practice

BSE if assigned to treatment. Under assumption 3 (monotonicity for never-takers), there is no PbB group, and so $\pi_i^{\mathrm{bB}}(p) \equiv 0$, $i = 1, \ldots, N$. This implies that NACE on the intermediate outcome $S$ is equal to minus $\pi_i^{\mathrm{Bb}}(p)$, and so it is non positive. As shown in Table 8, our data seem to support this assumption; our analysis suggests that there is a non-negligible negative ITT effect on BSE practice for never-takers. The presence of this 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 8 - has mass primarily ($> 75\%$) to the left of zero. Note that, in our application, this overall ITT effect can be defined as the weighted average of the ITT effects across never-takers and compliers:

$$\begin{aligned}
&\mathrm{E}\big(S_i(T) - S_i(C) \mid \theta\big) \\
&\quad = \pi_i^P \mathrm{E}\big(S_i(T) - S_i(C) \mid D_i(T) = P; \theta\big) + (1 - \pi_i^P)\mathrm{E}\big(S_i(T) - S_i(C) \mid D_i(T) = p; \theta\big) \\
&\quad = \pi_i(\mathrm{PbB}) - \pi_i(\mathrm{pBb}).
\end{aligned}$$

As it is well known, the standard intention-to-treat analysis focuses on the causal effect of assignment of treatment rather than the causal effect of receipt of treatment. Therefore, in our application, the global ITT effect on BSE practice represents the impact of being invited to attend the enhanced teaching course on posttreatment BSE practice and, due to the presence of imperfect compliance, it cannot be taken as summarizing the evidence in the data for the effects of treatments. If we want to learn something about the effect of attending the course on BSE practice, we should focus only on the compliers, the CACE on $S$. The corresponding estimand for individual $i$ is defined as

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

which under the monotonicity assumption for compliers (assumption 2), is equal to $\pi_i^{\mathrm{bB}}(P)$, the probability of being a complier who would practice BSE under treatment, but would not practice BSE under control. Our model, therefore, implies that the complier ITT effect on BSE practice is not negative by construction. As noted previously, this assumption seems plausible, because 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 to cause significant decreases in posttreatment BSE. Our analysis suggests 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. In addition, examining their joint distribution in Figure 2, we see that these effects are negatively correlated. Although this result necessarily relies more heavily on the specific form of the likelihood function and prior distribution, it casts considerable doubt on the scientific validity of the practical inference that would be drawn from a standard ITT analysis which drops subjects with missing data outcomes and ignores compliance information.

## 9.2 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, $Y$?

As noted previously, the ITT estimand on BSE quality is the effect of the assignment on the quality of self exams for those women who would practice BSE under both assignments (women belonging to the PBB group or to the pBB group), and the

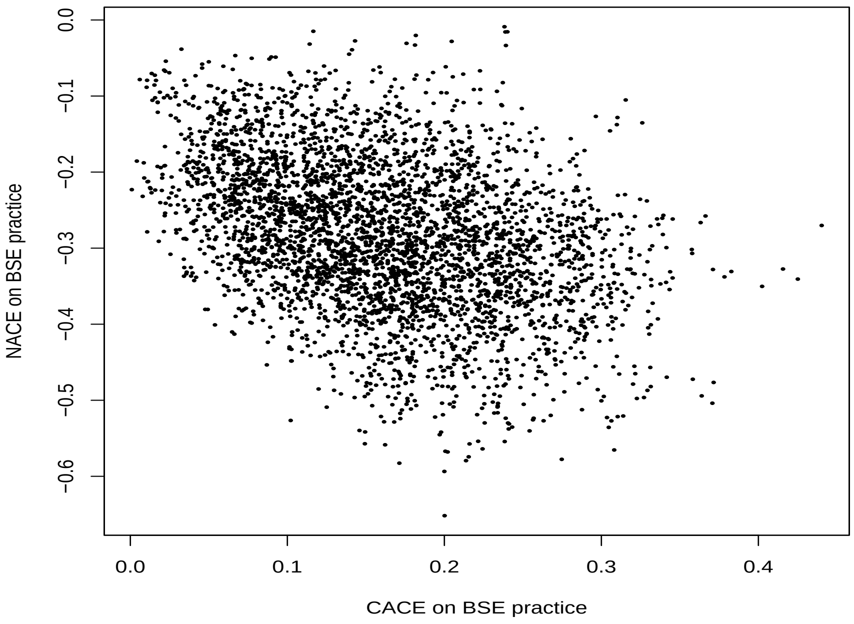

Figure 2: Simulation scatterplot of the joint posterior distribution of CACE and NACE on BSE practice.

CACE estimand on $Y$ is the effect of the treatment for compliers who would practice BSE under both treatments, namely, for women belonging to the PBB group.

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. However, we can infer about the proportion of women who would practice BSE with high quality in the PbB group when assigned to treatment, and about the proportion of women who would practice BSE with high quality in the pBb group when assigned to control. As concern the Pbb group and the pbb group, any inference can be drawn about quality of self-exam execution, being both $Y_i(C) = *$ and $Y_i(T) = *$ for these two

Table 10: Summary statistics of the posterior distributions of the quality 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 |

types of women. Table 10 shows summary statistics of the posterior distributions of the estimands of interest about 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 post-treatment quality of self exams adjusted for the intermediate outcome "BSE practice", $S$. The corresponding estimand for individual $i$ is defined as

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

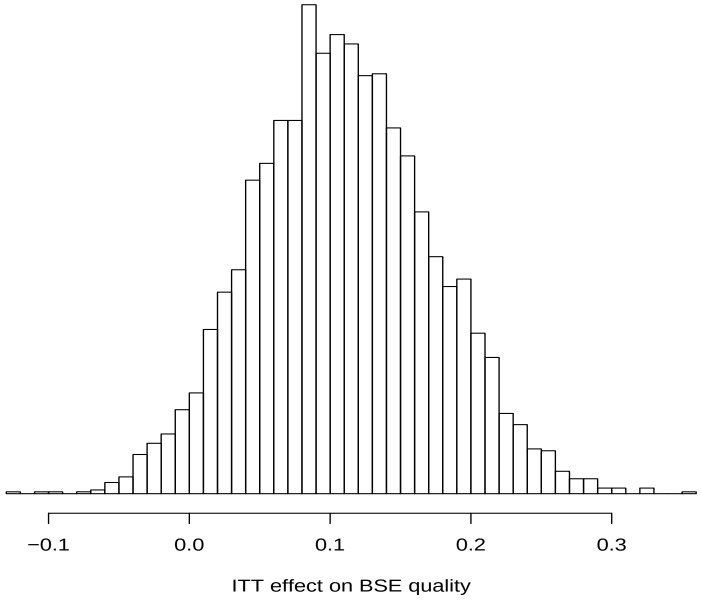

Figure 3: Simulation histogram of the posterior distribution of the intention-to-treat effect on BSE quality.

The simulated posterior distribution of this ITT effect is summarized in the second row in Table 10. Figure 3 shows its histogram. To draw form this distribution, we use the same step (1) described in the previous section and then calculate the expected causal effect $\mathrm{E}\big(Y_i(T) - Y_i(C) \mid S_i(C) = B, S_i(T) = B, \mathbf{X}_i^{\mathrm{obs}} = \mathbf{x}_i; \theta\big)$ for each subject based on the BSE quality submodel (see section 8), and average these values over the subjects whose current draw of $(S_i(C), S_i(T))$ is $(\mathrm{B}, \mathrm{B})$.

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

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

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 10, and Figure 4 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.

It should be noted that, in general, posterior distributions of the causal estimands on BSE quality are associated with a quite large uncertainty, because they are only defined for subgroups of all women, those who would practice BSE under both assignments.

As can be seen in Table 10, 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

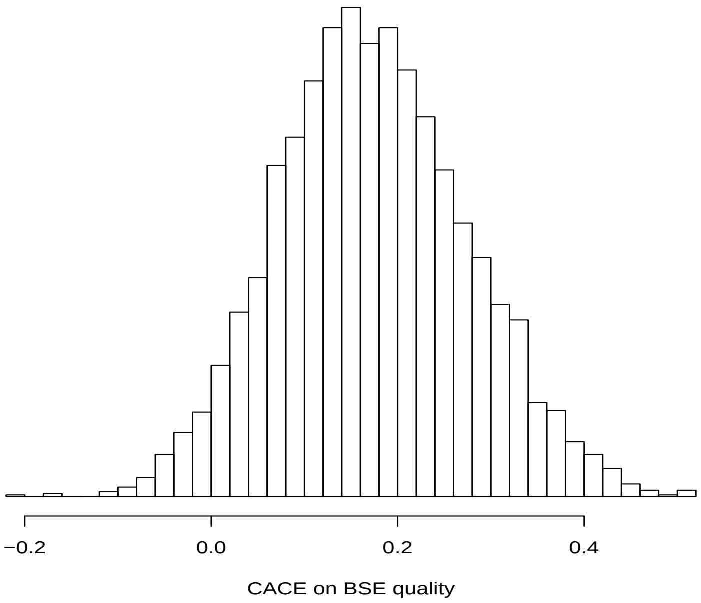

Figure 4: Simulation histogram of the posterior distribution of the complier average causal effect on BSE quality.

belonging to the PBB group are very sensitive to the risk of breast cancer, and so they tend to practice BSE correctly.

### 9.3 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

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

Table 11: 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 |

To draw from the distributions of these estimands, we use step (1) described in section 9.1 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 be the principal stratum indicator $G_i$. Table 11 presents some summary statistics of these posterior distributions. Recall that $\Pr\big(R_i(z) \mid G_i = g;\theta\big) = \Pr\big(R_i(z) \mid D_i(T) = p;\theta\big)$ for each $g \in \{\text{pBB}, \text{pBb}, \text{pbb}\}$ under assumption 7, and under the response exclusion restriction for compliers (assumption 5), given the BSE practice behavior, compliers have the same response behavior irrespective of the treatment arm they are assigned to.

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.

# 10 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. The possibility that the model is only weakly identified implies that its specification is more important than usual, as we have discussed in section 8.

## 10.1 Convergence Checks

Because posterior distributions were simulated from an MCMC algorithm (Appendix A), it is important to assess its convergence. To do this, we ran three 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 the 46 estimands (parameters and functions of parameters) that serve as building block for all other estimands. 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 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.

## 10.2 Model Checks

We evaluate the influence of the model presented in section 8 using posterior predictive checks. A posterior predictive check generally involves: ($a$) choosing a discrepancy measure, that is, a function of observed data and possibly of missing data and the parameter vector $\theta$, and ($b$) computing a posterior predictive p-value (PPPV), which is the probability, over the posterior predictive distribution of the missing data and $\theta$, that the discrepancy measure in a new study drawn with the same $\theta$ as in our study would be as or more extreme than in our study (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996). More formally, 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$^{\text{rep}}$ 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). As a result, properties of PPPVs are not exactly the same as properties of classical p-values under frequency evaluations over both levels of uncertainty, namely, the drawing of $\theta$ from the prior distribution and the drawing of data from the likelihood given $\theta$. For example, over frequency evaluations of the latter type, a PPPV is stochastically less variable than, but with the same mean as, the uniform distribution and so tends to be more conservative than the classical p-value, although the reverse can be true over frequency evaluations of the first type (Meng, 1994). For more details on the interpretation and properties of PPPVs, see also Rubin (1984) and Gelman et al. (1996).

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 $S^{\text{rep}}_i$, $R^{\text{rep}}_i$, 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. In each replicated study, $S^{\text{rep}}_i$ and $Y^{\text{rep}}_i$ are either jointly observed or jointly missing as well as in the Faenza study; so $R^{\text{rep}}_i$ is equal to 1 if and only if $S^{\text{rep}}_i$ and $Y^{\text{rep}}_i$, the BSE practice outcome and the quality outcome, are jointly observed (with $Y^{\text{rep}}_i = *$ if $S^{\text{rep}}_i = b$), and 0 otherwise. 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

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

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

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

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 9. Concerning the quality outcome, the causal effect of primary interest is the ITT effect for women belonging to the PBB group, so that our checks focus on the quality outcome distribution within this subpopulation. More preferable measures might have been the posterior mean and standard deviation for the actual estimands in section 9 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 7. 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. A good posterior predictive check for assessing our monotonicity assumptions might be built by fitting other three alternative models: a model in which only the monotonicity assumption for never-takers (assumption 3) is made; a model in which only the monotonicity assumption for compliers (assumption 2) is made; and a model in which neither of the two monotonicity assumptions is made. In such a case, we could draw replicated data from the posterior distribution under each model and compare realizations. This approach would require high time-consuming; in addition, relaxing one or both the monotonicity assumptions implies that the number of parameters increases. This would not complicate

Table 12: 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 |

the computational methodology greatly but, given the relatively small sample size, would lead to imprecise estimates and make difficult the convergence of the MCMC algorithm used to simulate the posterior distributions.

Although the posterior distributions for parameters in alternative models are not readily available, we can use a posterior predictive check distribution for monitoring our model. To determine if the monotonicity assumptions 2 and 3 are supported by our data, we could consider, for example, the log-likelihood ratio discrepancy:

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

where $g \in \{\mathrm{PBB}, \mathrm{PbB}, \mathrm{Pbb}, \mathrm{pBB}, \mathrm{pBb}, \mathrm{pbb}\}$, 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. Figure 5 shows a scatterplot of the realized discrepancy $\mathrm{L}^2(\mathrm{data}, \mathbf{G}, \theta)$ versus the predictive discrepancy $\mathrm{L}^2(\mathrm{data}^{\mathrm{rep}}, \mathbf{G}^{\mathrm{rep}}, \theta)$, in which each point represents a different value of $(\mathrm{data}^{\mathrm{rep}}, \mathbf{G}^{\mathrm{rep}}, \theta)$ drawn from the posterior distribution. In this case, the PPPV is equal to 0.783, the proportion of points above the 45° line in the figure. 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

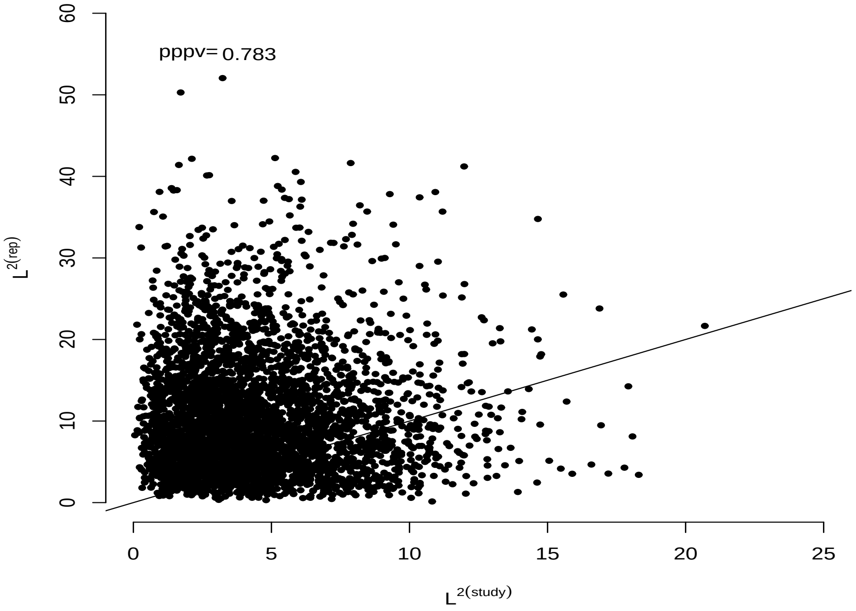

Figure 5: Scatterplot of predictive vs realized loglikelihood ratio discrepancies under the joint posterior distribution.

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 12 and in Figure 5, provide no special evidence for such influence of the model.

## 11 Concluding Remarks

In this paper, we defined the framework for principal stratification in broken 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 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: probably, they consider the course only as a waste of time. The presence of this strong negative ITT effect for never-takers leads to a negative (even if small and non-significant) global ITT effect on BSE practice. These results suggest that, in applications such as ours, the interpretation of the global intention-to-treat estimates as summarizing the evidence in the data for the effects of treatments on BSE practice can be too coarse and even misleading.

Concerning BSE quality, our results do not indicate strong treatment effects. The posterior distribution for the CACE estimand on BSE quality, which measures the effect of the attendance in the training program versus only mailed informational leaflet, is larger than the corresponding ITT effect but is also associated with greater uncertainty.

The results from this randomized study are not subject to the selection bias in the way that nonrandomized study are. Nevertheless, although we use the CACE on BSE practice and the CACE on BSE quality, well defined causal effects, to represent the effect of attendance in the enhanced training program versus only a mailed informational leaflet, it is important to remember that they are defined on specific subsets of the study women: CACE on BSE practice is defined on women who would have complied with either assignment, and CACE on BSE quality is defined on on a subgroup of these women: those who belong to the PBB group. For the other women there is no information on such effects of attendance in this study;

therefore, as with any randomized trial that is based on subpopulations, external information, such as background variables also needs to be used when generalizing these causal effects to other target groups of women.

Our results also reveal 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, response is increasing in the following order: the PBB group, the PbB group, and the Pbb group. These suggest that women who are probably more sensible to the risk of breast cancer and have likely higher motivation are more willing to respond to the survey.

# Appendix A

## 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 $\theta$ 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 $j+1$ following 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 $\theta$ and W, the $D_i(T)$ are independent of $D_j(T)$, $Z_j^{\text{obs}}$, $D_j^{\text{obs}}$, $R_j^{\text{obs}}$, $S_j^{\text{obs}}$, $Y_j^{\text{obs}}$ for all $j \neq i$. Then,

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

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

$$\begin{aligned}
&\Pr\bigl(D_i(T) = P \mid Z_i^{\text{obs}} = C, D_i^{\text{obs}} = p, R_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) \propto \\
&\pi_i^P \Biggl[ \Bigl( \pi_i^{\text{BB}}(P) \pi_{iC}^R(\text{PBB}) \bigl(\pi_{iC}^Y(\text{PBB})\bigr)^{\mathrm{I}\{Y_i^{\text{obs}}=H\}} \bigl(1 - \pi_{iC}^Y(\text{PBB})\bigr)^{\mathrm{I}\{Y_i^{\text{obs}}=L\}} \Bigr)^{\mathrm{I}\{R_i^{\text{obs}}=1, S_i^{\text{obs}}=B\}} \\
&\qquad \Bigl( \pi_i^{\text{bB}}(P) \pi_{iC}^R(\text{PbB}) + \pi_i^{\text{bb}}(P) \pi_{iC}^R(\text{Pbb}) \Bigr)^{\mathrm{I}\{R_i^{\text{obs}}=1, S_i^{\text{obs}}=b\}} \Bigl( \pi_i^{\text{BB}}(P) \bigl(1 - \pi_{iC}^R(\text{PBB})\bigr) + \\
&\qquad\qquad \pi_i^{\text{bB}}(P) \bigl(1 - \pi_{iC}^R(\text{PbB})\bigr) + \pi_i^{\text{bb}}(P) \bigl(1 - \pi_{iC}^R(\text{Pbb})\bigr) \Bigr)^{\mathrm{I}\{R_i^{\text{obs}}=0\}} \Biggr].
\end{aligned}$$

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 $\Pr(S(C), S(T) \mid D(T)^{(j+1)}, \theta^{(j)}, \mathrm{W})$. By the monotonicity assumption for compliers (assumption 2),

$$\begin{aligned}
\Pr\bigl(S_i(C) = b, S_i(T) = b \mid D_i(T)^{(j+1)} = P, \text{OBS}(T, P, 1, b), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) &= 1; \\
\Pr\bigl(S_i(C) = B, S_i(T) = B \mid D_i(T)^{(j+1)} = P, \text{OBS}(C, p, 1, B), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) &= 1;
\end{aligned}$$

and by the monotonicity assumption for never-takers (assumption 3),

$$\begin{aligned}
\Pr\bigl(S_i(C) = B, S_i(T) = B \mid D_i(T)^{(j+1)} = p, \text{OBS}(T, p, 1, B), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) &= 1; \\
\Pr\bigl(S_i(C) = b, S_i(T) = b \mid D_i(T)^{(j+1)} = p, \text{OBS}(C, p, 1, b), Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) &= 1.
\end{aligned}$$

In addition, because $\{i : D_i(T)^{(j+1)} = P, Z_i^{\text{obs}} = T, D_i^{\text{obs}} = p\} = \emptyset$,

$$\Pr\bigl(S_i(C), S_i(T) \mid D_i(T)^{(j+1)} = P, Z_i^{\text{obs}} = T, D_i^{\text{obs}} = p, R_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) = 0.$$

Similarly, being $\{i : D_i(T)^{(j+1)} = p, Z_i^{\text{obs}} = T, D_i^{\text{obs}} = P\} = \emptyset$,

$$\Pr\bigl(S_i(C), S_i(T) \mid D_i(T)^{(j+1)} = p, Z_i^{\text{obs}} = T, D_i^{\text{obs}} = P, R_i^{\text{obs}}, S_i^{\text{obs}}, Y_i^{\text{obs}}, \mathbf{X}_i^{\text{obs}}\bigr) = 0.$$

It remains to consider the observed groups OBS$(T, P, 1, B)$, OBS$(T, P, 0, ?)$, OBS$(C, p, 1, b)$, and OBS$(C, p, 1, ?)$ for observations with $D_i(T)^{(j+1)} = P$, and the observed groups OBS$(T, p,\ 1, b)$, OBS$(T, p, 0, ?)$, OBS$(C, p, 1, B)$, and OBS$(C, p, 1, ?)$ for observations with $D_i(T)^{(j+1)} = p$. A subject with $Z_i^{\text{obs}} = T$, $D_i^{\text{obs}} = P$, $R_i^{\text{obs}} = 1$, $S_i^{\text{obs}} = B$ is a complier who would practice BSE under both assignments or who would practice BSE if assigned to treatment but not if assigned to control. Therefore,

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

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

and

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

for each $(s_C, s_T) \in \{(B, b), (b, b)\}$. The drawing of $(S_i(C), S_i(T))$ for the other subjects is done in a similar way.

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

For the parameters $\delta_{0C}(\text{PBB})$, $\delta_{0\cdot}(\text{pBB})$, and $\delta_{0C}(\text{pBb})$ we know the full conditional distributions: they are Beta distributions; so we can directly draw from them.[2] The parameter $\delta_{0T}(\text{PbB})$ is drawn from a truncated Beta distributions. For the other subvectors of $\theta$, in our specification, it is rather difficult to draw directly

[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 $\delta_{0T}(\text{PBB})$ and $\delta_{0T}(\text{PbB})$ would be Beta distributions.

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 $\alpha$, we draw *candidate* values $\alpha^*$ from a density $\mathrm{g}(\alpha \mid \theta^{(j)})$. The candidate draw is accepted with probability

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

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

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

simplifying the expression for $\tau$ 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., Du, J., and Rubin, D. B. (1998) A broader template for analyzing broken randomized experiments. *Sociological Methods and Research*, **27**, 285–317.

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.

Dempster A. P., Laird, N., and Rubin, D. B. (1977) Maximum 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, 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 treatment-non-compliance and subsequent missing outcomes. *Biometrika*, **86**, 365–379.

Frangakis, C. E., and Rubin, D. B. (2001) Addressing the idiosyncrasy in estimating survival curves using double-sampling in the presence of self-selected right censoring, (with discussion). *Biometrics*, **57**, 333–353.

Frangakis, C. E., and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

Gelfand, A. E, and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

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.

Geman, S., and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993) Multiple imputation in mixture models for nonignorable nonresponse with follow-up. *Journal of the American Statistical Association*, **88**, 984–993.

Haavelmo, T. (1943) The statistical implication of a system of simultaneous equations. *Econometrica*, **11**, 1–12.

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.

Manski, C. F. (1990) Non-parametric bounds on treatment effects. *American Economic Review, Papera and Proceedings*, **80**, 319–323.

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.

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.

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. (1977) Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1–26.

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

Rubin, D. B. (1978b) Multiple imputation in sample survey: a phenomenological Bayesian approach to nonresponse (C/R: P29-34). *ASA Proceedings of Survey Research Methods Section*, 20–28.

Rubin, D. B. (1979) Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*, **74**, 318–328.

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. (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. (1998) More powerful randomization-based $p$-values in double-blind trials with noncompliance. *Statistics in Medicine*, **17**, 371–387.

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.

The Coronary Drug Project Research Group. (1980) Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine*, **303**, 1038–1041.

Tinbergen, J. (1930) Determination and Interpretation of Supply Curves: An Example. In *The Foundations of Econometric Analysis*, eds. D. Hendry and M. Morgan, Cambridge: Cambridge University Press, reprinted from Zeitschrift fur Nationalokonomie.

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*.

# Assessing the Impact of Demographic Events on Poverty using a Quasi-Experimental Approach

## Abstract

In this paper we analyze the impact of childbearing events on individuals'wellbeing in Indonesia, using a sample of women from the Indonesia Family Life Survey. We consider the impact of having children on wellbeing as a quasi-experimental problem. In this approach the endogenous variable of interest - having children - is considered as the treatment variable, and wellbeing level is the outcome variable. The main issue in this approach is that subjects who experience childbearing events might somewhat be self-selected, and so large differences may exist between the treatment and control groups on observable as well as unobservable covariates, which can lead to badly biased estimates of treatment effects. Therefore, to be able to estimate the causal effects of interest additional assumptions have to be made. An assumption often made in such a study is the strong ignorability or unconfoundedness of the assignment mechanism given the observed covariates, which requires that all variables that affect both outcome and the probability of receiving the treatment are observed. Although this assumption is not testable, it clearly is a very strong assumption, and one that need not generally be applicable. In our application, we impose the strong ignorability assumption; we consider the assumption that all relevant variables are observed as a reasonable approximation. In addition, any

alternative assumptions that not rely on unconfoundedness, while allowing for consistent estimation of the causal effects of interest, must make alternative untestable assumptions, which can be even more difficult to justify. Unconfoundedness validates comparisons between different units with identical values of the observed background characteristics. When there are many background covariates, as in our study, balancing the distribution of all the covariates between treated and control groups can be hard. To address this problem Rosenbaum and Rubin (1983) developed the propensity score methodology. The key insight was that given the strong ignorability assumption, treatment assignment and the potential outcomes are independent given the propensity score only. Thus, adjusting for the propensity score removes the bias associated with differences in the observed covariates in the treated and control groups. To estimate propensity scores, which are the conditional probabilities of being treated given a vector of observed covariates, we must model the distribution of the treatment indicator given these observed covariates. Much work has been done in the case where the covariates are fully observed. We address the problem of calculating propensity scores when covariates can have missing values, which is the case in our application. We consider three different approaches of handling missing background data in the estimation and use of propensity scores: a complete-case analysis, a multiple imputation approach, and a pattern-mixture model based approach. For each approach, we use the resulting propensity scores to construct comparison groups to the group of treated subjects, and then we estimate the causal effect of interest, and compare the results.

Keywords: IFLS, Longitudinal data, Childbearing, Wellbeing, Propensity Score Matching, Missing data, Multiple imputation, Pattern-mixture model.

# 1 Introduction

The relationship between poverty dynamics and fertility in developing countries is one of the most disputed topic among economists and demographers. The general empirical observation that poorer countries tend to have higher population growth rates and that larger households tend to be poorer, underlies the common presumption of a negative causal relation between poverty and fertility at the national and household levels respectively.

Existing studies on this long-standing issue have relied on either cross sectional micro data or aggregate level data, which, no matter what techniques applied, are unlikely to provide robust causal information about the relationship between the occurrence of life events and poverty, as well as the impact of intermediate variables. Only longitudinal surveys can provide information on the timing and duration of poverty spells. Existing research on poverty dynamics based on panel data for developing countries is not yet common, although some exception exist. However, none of these studies analyzes explicitly the relationship between poverty and demographic events such as fertility.

Using a sample of women drawn from the Indonesia Family Life Survey (IFLS), a longitudinal survey conducted by RAND in collaboration with other socio-economic and demographic institutes, this paper makes an attempt to identify causal relationship between poverty and fertility in Indonesia, by focussing on the extent to which childbearing events lead to changes in wellbeing. Indonesia is the fourth most populous country in the world, with an estimated population level at 207 million in 2000. Although Indonesia has a positive net population growth, the total fertility rate declined consistently during the last decades. Interestingly, the nineties represent also a period in which Indonesia benefited from sustained high economic growth, reducing the overall poverty rates. At the same time Indonesia has faced increasing participation in education. Another interesting feature of the Indonesia setting is the financial crisis that occurred in 1997, producing a dramatic reversal in educational enrolment and poverty. The IFLS, with its four waves ranged between 1993 and 2000, therefore covers an eventful period in the history of Indonesia - providing detailed information about individuals' behavior prosperity as well as sudden hardship. Consequently the IFSL provides an extremely useful data source given our research objective.

Longitudinal information requires use of appropriate methods that take into account units being measured at different points in time. Standard methods of panel data analysis, such as hazard regression, can be quite deceptive for the aim of establishing causal relationships, because they provide no warnings about their property. An alternative approach is to consider the impact of demographic events on poverty as quasi experiments, that is, one considers the endogenous variable of interest (e.g., change in fertility) as treatment variable and divides individuals into two groups: those who experienced a childbirth - the treatment group, and those who did not - the control group.

The main issue in this approach is that those women experiencing a childbearing might be somewhat self-selected, because they are not necessarily selected in a random fashion. Therefore there is the need to control for naturally occurring systematic differences in background characteristics $\mathbf{X}$ between the treatment group and the control group (e.g., in age or sex distributions). To do this, throughout our analysis, we make the strong ignorability or unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983), which asserts that conditional on the pretreatment variables, the treatment indicator is independent of the potential outcomes. In other words, within subpopulation defined by values of the covariates, we have random assignment of treatment. This assumption may be controversial. In fact, it requires that all variables that affect both outcome and the likelihood of receiving the treatment are observed or that all the others are perfectly collinear with the observed ones. Although this assumption is not testable, it is a very strong assumption, and one that need not generally be applicable. Clearly selection may also take place on the basis of unobservable characteristics. However, we view it as a useful starting point for two reasons. One is that in our study, we have carefully investigated which variables are most likely to confound any comparison between treated and control units, and so we believe that the assumption that all relevant variables are observed may be a reasonable approximation. Second, any alternative assumptions that not rely on unconfoundedness, while allowing for consistent estimation of the causal effects of interest, must make alternative untestable assumptions. Whereas the unconfoundedness assumption implies that the best matches are units that differ only in their treatment status, but otherwise are identical, alternative assumptions implicitly match units that differ in the pretreatment characteristics. Often such assumptions are even more difficult to justify. For instance, the technique of instrumental variables is sometimes considered as an alternative to assuming unconfoundedness, but in our setting the use of this approach is not particular useful since finding valid instruments (variables that influences the demographic event, but do not influence poverty) is difficult. The strong ignorability assumption therefore may be a natural starting point after comparing average outcomes for treated and control units to adjust for observable pretreatment differences.

The unconfoundedness assumption validates the comparison of treated and control units with the same value of the covariates. Now the problem is that in our study there are many background characteristics that need to be controlled. To address this situation, we use propensity score matching, introduced by Rosenbaum and Rubin (1983), to construct comparison groups to the group of treated individuals - in our case those who experience a change in fertility status. The key insight in this methodology is that given the strong ignorability assumption, treatment assignment and the potential outcomes are independent on propensity score. Thus, adjusting for the propensity score removes the bias associated with differences in the observed covariates in the treated and control groups. In other words, propensity scores are a one-dimensional summary of multidimensional covariates, $\mathbf{X}$, such that when the propensity scores are balanced across the treatment and comparison groups, the distribution of all the pretreatment variables, $\mathbf{X}$, is balanced in expectation across the two groups. This reduction from many characteristics to one composite characteristic allows the straightforward assessment of whether the treated and control groups overlap enough on background characteristics to lead to sensible estimation of treatment versus control effects from the available data.

To estimate propensity scores, which are the conditional probabilities of being treated given a vector of observed covariates, e must model the distribution of the treatment indicator given these observed covariates. Much work has been done in the case where covariates are fully observed (e.g., Gu and Rosenbaum, 1993; Rubin and Thomas, 1992a, 1992b, 1996). Unfortunately, in our study some covariates have missing values. In such a case, which commonly arises in practice, it is not clear how the propensity score should be estimated when some covariate values are missing. In addition, the missingness itself may be predictive about which treatment is received in the sense that the treatment assignment mechanism is ignorable (Rubin, 1978) given the value of $\mathbf{X}$ and the pattern of missing covariates but not ignorable given only the former.

In this paper we compare a complete-data analysis which drops subjects with missing background data, with an analysis based on the "generalized" propensity score as defined in Appendix B of Rosenbaum and Rubin (1984). Rosenbaum and Rubin (1984) proved that adjustment for the "generalized" propensity score in expectation balances the observed covariate information and the pattern of missing covariates.

We also consider multiple imputation (MI) to address the problem of missing covariate values. MI is a Monte Carlo technique in which each missing value is replaced by a set of plausible values that represent the uncertainty about the right value to impute. Then any complete-data analysis can be performed on each of the imputed datasets and the results can be combined in a straightforward manner to yield the final estimands. No matter which complete-data analysis is used, the process of combining results from different datasets is essentially the same. The results is valid statistical inferences that properly reflect the uncertainty due to the missing values. Therefore, we create $m$ multiply imputed datasets by filling in $m$ times the missing covariate data, and then we apply propensity score matching in each simulated complete dataset. Finally, we combine the results from the $m$ complete datasets for inference.

Having estimates of the propensity scores, if the treated and control groups overlap enough on background characteristics, we may proceed to the matching stage, when treated and comparison units are paired according to their scores. Here, we use what is called “Stratification Matching” to perform the matching and to estimate the causal effect of interest; this is a particularly straightforward technique for estimating treatment versus control effects that reflects adjustment for differences in all observed background characteristics.

The paper is outlined as follows. Section 2 summarizes the existing literature on the analysis of the relationship between poverty dynamics and fertility - with a particular emphasis on the motivation of this study, and section 3 provides notation and describes the methodological approach here applied. Section 4 gives a brief description of the IFLS data. Section 5 explains how we define wellbeing and puts particular emphasis on the choice of consumption rather than income as an indicator of living standard in Indonesia. Using this wellbeing definition, the section provides some descriptive patterns of poverty for different family types. Section 6 applies the methodological strategies described in section 3 to the IFLS data and section 7 presents the results. Section 8 concludes with a summary and suggestions for direction of future research.

## 2 Background on Poverty Dynamics and Fertility

The link between population growth and economic wellbeing is one of the most disputed research areas among economist and demographers (Birdsall et al., 2001). The general empirical observation that poorer countries tend to have higher population growth rates and that larger households tend to be poorer, underlies the common presumption of a negative causal relation between poverty and fertility at the national and household levels respectively. The macro level argument on this issue relies on the neoclassic paradigm that a higher population growth rate depresses capital accumulation and wages. Poverty in turn is consider as a key factor in driving high fertility and therefore high rates of population growth. Consequently, it is seen as a crucial element in delaying the demographic transition. However, these theoretical assertion are not sufficient; the interpretation of the link between poverty and fertility cannot neglects the institutional settings as well as other important factors such as time-lags, feedback mechanisms, nonlinearities and reverse causation. A similar argument applies at the micro-level. Individual level fertility behavior adjusts to changes in perceived and actual cost and benefits of children. Economic forces, social organizations and cultural patterns in turn influence prices that determine costs and benefits of children.

While existing studies have relied on either cross sectional micro data or aggregate level data, our study revisit this long-standing issue by exploiting a recent longitudinal dataset - the Indonesia Family Life Survey data. Cross sectional data, no matter what techniques applied, is unlikely to provide robust causal information about the relationship between the occurrence of life events (here births) and poverty, as well as the impact of intermediate variables. Only longitudinal surveys can provide information on the timing and duration of poverty spells, implying that the panels will provide much richer information on issue such as the permanent nature of the poverty, changes in poverty status of individuals over time and about the events related to entry into and escape from income poverty (Muffels, 2000). Recent research on poverty dynamics in industrialized countries has provided striking advantages in our understanding of poverty and policy making in general (e.g., Rose, 2000; Bane and Ellwood, 1986; Huff-Stevens, 1999). Inspired by this progress, here, we analyze the dynamic nature of poverty in the context of a developing country - Indonesia. An important contribution of our study is the recognition that life events such as childbearing, education, health, and employment are not exogenous with respect to poverty transitions. The aim of this study is therefore to implement a treatment effect model, which can identify causal relationships between poverty and fertility, and consequently inform policy makers about what policies may - or may not - work in reducing poverty. Though the relationship between poverty and fertility at the micro level becomes more complex, it certainly provides a more correct framework if the aim is to provide useful policy recommendations. As noted in Jalan and Ravallion (2000) and as investigated in Aassve et al. (2003a) for industrialized countries, there is substantial movement in and out of poverty over one's life cycle - with transitions very much related to other life cycle events - including fertility events.

Existing research on poverty dynamics based on panel data for developing countries is not yet common, although some exception exist. Jalan and Ravallion (2000) using a panel from rural China indicate that there are important differences between transient and chronic poverty. They suggest that policies aimed at reducing transient poverty may not necessarily imply a reduction in chronic poverty. Dercon and Krishman (2000) using the three waves of a panel of rural Ethiopia shows that individual consumption level varies widely by year and season, and indicate that a much larger proportion of households are vulnerable to poverty than what cross sectional poverty statistics may suggest. Mcculloch and Baulch (2000) use a five-year panel of households from rural Pakistan and show that large reductions in poverty can be achieved through policies aiming at smoothing household incomes - simply because large part of poverty is indeed transitory. At the same time, they also show that transitory shocks may have a long-term consequences on household poverty status. Baulch and Hoddinott (2000) summarizes these findings by suggesting that the pool of poor households consist of both chronic poor and transitory poor, where the latter is surprisingly large, and this is the case independent of whether poverty is measured in relative or absolute terms. However, none of these studies analyzes explicitly the relationship between poverty, transitory or not, on other crucial processes and in particular on demographic events such as fertility. For example, they do not answer the question of whether the number of children has any impact on households' experiences of poverty. Existing micro-level studies concerning the relationship between poverty and fertility rely exclusively on cross sectional data. They show very mixed results, indicating that the relationship does not appear to be unidirectional. Some studies suggest a positive relationship between fertility and poverty level, others find it to be negative, and yet others find it to have an inverse J-shaped relationship. Yet other studies find very little evidence of any relationship at all (Schoumaker and Tabutin, 1999).

In general, past studies seem to have been using deficient data sources, and often too simple econometric techniques. Although hazard regression seems to be the most natural choice in modeling poverty dynamics, more sophisticated methods are needed if the aim is to establish causal relationships between poverty and other life course events. For instance a simple binary regression of poverty on demographic variables does not enable the researches to say much about causality, unless specific assumptions hold. If fertility events are indeed endogenous with respect to poverty status, then any regression will yield biased results, and consequently will not produce very useful results for policy analysis. The use of the instrumental variable approach is not particular helpful in this setting since finding valid instruments, that is, variables which influences the demographic event, but does not influence poverty given the demographic event, is difficult. In general, parameter estimate tend to be sensitive if instruments are not particulary strong (e.g., Klepinger et al., 1995). A different approach is to consider the impact of demographic events on poverty (or vice versa) as quasi experiments. This refers to situations in which sample individuals are divided into a treatment group and a comparison group, but where the former is not necessarily selected in a random fashion. By doing this one considers the endogenous variable of interest (e.g., change in fertility) as treatment variable. This is similar to the way policy analysts study the impact of poverty-reducing policies (e.g., the impact of improved school facilities on poverty). The main issue in these studies is that those individuals benefiting from the poverty-reducing policy might somewhat be self-selected. To overcome this problem propensity score matching is used to construct comparison groups to the group of treated individuals - in this case those who experienced a change in the demographic status (Rosenbaum and Rubin, 1983).

# 3 The Propensity Score Methodology

## 3.1 Estimation of Propensity Scores

Since they were introduced by Rosenbaum and Rubin (1983), propensity scores have been used in observational studies in many fields to adjust for imbalances on pretreatment variables, $\mathbf{X}$, between a treated group, indicated by $Z = T$, and a control group, indicated by $Z = C$ (e.g., D'Agostino, 1998; Dehejia and Wahba, 1999; Rubin, 1997). Propensity scores are one-dimensional summary of multidimensional covariates, $\mathbf{X}$, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all the covariates, $\mathbf{X}$, are balanced in expectation across the two groups.

The propensity score for an individual $i$ $(i = 1, \dots, N)$ is the conditional probability of receiving a particular treatment $(Z_i = T)$ versus control $(Z_i = C)$ given a vector of observed covariates, $\mathbf{x}_i$,

$$e_i = e_i(\mathbf{x}_i) = \Pr\big(Z_i = T \mid \mathbf{X}_i = \mathbf{x}_i\big), \tag{3.1}$$

where it assumed that, given the $\mathbf{X}$'s, the $Z_i$ are independent,

$$\begin{aligned}\Pr\big(Z_1 = z_1, \dots, Z_N = z_N \mid \mathbf{X}_1 = \mathbf{x}_1 \dots \mathbf{X}_N = \mathbf{x}_N\big)\\ = \prod_{i=1}^{N} e(\mathbf{x}_i)^{\mathrm{I}\{z_i=T\}}\big(1 - e(\mathbf{x}_i)\big)^{(1-\mathrm{I}\{z_i=T\})}.\end{aligned}$$

Rosenbaum and Rubin (1983) showed that for a specific value of the propensity score, the difference between the treatment and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score, if the treatment assignment is strongly ignorable given the covariates. Thus matching, subclassification, or regression (covariance) adjustment on propensity score tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable, which occurs when the treatment assignment, $Z$, and the potential outcomes, say $(Y(C), Y(T))$, are conditionally independent given the covariates $\mathbf{X}$: $\Pr(Z \mid \mathbf{X}, Y(C), Y(T)) = \Pr(Z \mid \mathbf{X})$.

## 3.2 Estimating Propensity Scores with Incomplete Data

Usually we do not actually know the propensity scores, and so we must estimate them. To estimate propensity scores for all individuals, one must model the distribution of $Z$ given the observed covariates, $\mathbf{X}$. There is a recent and large literature on propensity score methods with complete data (e.g., Gu and Rosenbaum, 1993; Rubin and Thomas, 1992a, 1992b, 1996). In practice, typically some covariate values will be missing, and it is not clear how the propensity score should be estimated when some covariate values are missing. In addition, the missingness itself may be predictive about which treatment is received in the sense that the treatment assignment mechanism might be ignorable (Rubin, 1978) given the value of $\mathbf{X}$ and the observed pattern of missing covariates but not ignorable given only the observed value of $\mathbf{X}$.

Rosenbaum and Rubin (1984) considered using a "pattern mixture" model (Little 1993; Rubin 1986) for propensity score estimation with missing covariate data. Appendix B of Rosenbaum and Rubin (1984) defined a "generalized" propensity score as the probability of treatment assignment given $\mathbf{X}^*$, a $K$-coordinate vector, where the $j$th covariate of $\mathbf{X}^*$ is a covariate value if the $j$th covariate was observed, and is an asterisk if the $j$th covariate is missing. This is equivalent to conditioning on the observed values of $\mathbf{X}$, $\mathbf{X}^{\text{obs}}$, and a missing covariate indicator $R$.

More formally, let the response indicator be $R_{ij}$, $(j = 1, \ldots, K)$, which is 1 when the value of the $j$th covariate for the $i$th subject is observed and 0 when it is missing; $R_{ij}$ is fully observed by definition. Also, let $\mathbf{X} = (\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{mis}})$, where

$\mathbf{X}^{\text{obs}} = \{X_{ij} : R_{ij} = 1\}$ denotes the observed parts and $\mathbf{X}^{\text{mis}} = \{X_{ij} : R_{ij} = 0\}$ denotes the missing component of $\mathbf{X}$. Then, the generalized propensity score for the subject $i$, which conditions on all of the observed covariate information, is

$$e_i^* = e_i^*(\mathbf{X}_i^{\text{obs}}, R_i) = \Pr\big(Z_i = T \mid \mathbf{X}_i^{\text{obs}}, R_i\big). \tag{3.2}$$

Rosenbaum and Rubin (1984) showed that with missing covariate data and strongly ignorable treatment assignment given $\mathbf{X}^{\text{obs}}$ and $R$, the generalized propensity score $e_i^*$ in (3.2) plays the same role as the usual propensity score $e_i$ in (3.1) with no missing covariate data. Treatment assignment is strongly ignorable given $(\mathbf{X}^{\text{obs}}, R)$ if $\Pr(Z \mid \mathbf{X}, R, Y(C), Y(T)) = \Pr(Z \mid \mathbf{X}^{\text{obs}}, R)$. It is important to emphasize that, just as with propensity score matching with no missing data, the success of a propensity score estimation method is to be assessed by the quality of the balance in the $(\mathbf{X}^{\text{obs}}, R)$ distribution between the treatment group and the control group that has been achieved by matching on it.

Rosenbaum and Rubin (1984) suggested that in large enough samples, one can estimate the generalized propensity score by estimating a separate logit model using the subset of covariates fully observed for each pattern of missing data. In addition, they noted that with discrete covariates, their pattern mixture approach is equivalent to adding an additional "missing" category to each covariate.

An alternative approach to handling incomplete data is imputation, i.e., filling in missing data with plausible values, as it addresses the missing-data problem before beginning the analysis with the hope of working well for essentially all subsequent analyses. In order to incorporate missing-data uncertainty, Multiple Imputation (MI) was proposed by Rubin (1978; also see Rubin, 1987, 1996). MI is a Monte Carlo technique in which each missing values is replaced by $m > 1$ simulated versions, where $m$ is typically small (e.g. 3-10). Each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Most of the techniques presently available for creating MIs assume that the missing data mechanism is ignorable, but it is important to note that the MI paradigm does not require ignorable nonresponse. Imputations may in principle be created under any kind of model for the missing data mechanism, and the resulting inferences will be valid under that mechanism (see chapter 6, Rubin, 1987).

In this paper, we consider both the pattern mixture approach introduced by Rosenbaum and Rubin (1984) and the MI approach to address the problem of estimating and using propensity score with partially missing data, and we compare the results from these methods with results from a complete-data analysis.

As noted previously, propensity score matching relies heavily on the assumption of ignorability of the assignment mechanism. Since this ignorability depends on the relationship between $Z$ and all of the other study variables, it is crucial to examine the assumptions about all of the data generating mechanisms. For all of the techniques here described we will maintain the following assumption:

$$\Pr(Z \mid \mathbf{X}, R, Y(C), Y(T)) = \Pr(Z \mid \mathbf{X}, R),$$

in order to have ignorability of the assignment mechanism.

We now discuss the assumptions implicit in each of the competing methodologies in term of the generating mechanisms of the study variables.

A complete-data analysis uses only observations where all variables are observed. This means that any unit that has any missing data is removed from the study. In the best of circumstances this will be inefficient. In general, this best case scenario assumes that the units removed, those with missing data, are just a simple random sample of the other. This is a strong assumption, formally referred to as data Missing Completely At Random (MACR, Little and Rubin, 1987). It does not allow the missing data mechanism to depend on any other variable. In our specific context, to make valid causal inferences with this approach we require the following assumption regarding the missing data mechanism:

$$\Pr(R \mid \mathbf{X}, Z) = \Pr(R).$$

This a very strong assumption, particulary as the number of covariates, needed for ignorability of the assignment mechanism to hold, grows.

The pattern-mixture approach proposed by Rosenbaum and Rubin (1984) provides a possible solution to this problem. This method relies on either one of the following assumptions:

$$\begin{array}{rcl}\Pr(Z \mid \mathbf{X}, R) & = & \Pr(Z \mid \mathbf{X}^{\text{obs}}, R), \\ & \text{or} & \\ \Pr(Y(C), Y(T) \mid \mathbf{X}, R) & = & \Pr(Y(C), Y(C) \mid \mathbf{X}^{\text{obs}}, R).\end{array}$$

These assumptions imply that within each missing data pattern (defined by $R$), we either need assignment to be independent of the covariates unobserved for that pattern, or we need ignorability to be satisfied just on the basis of those covariates observed in that pattern.

The strength of this method is that, in principle, it does not make any assumption about the missing data mechanism; however it does assume that either all missing covariate values are already balanced across treatment groups or that they are independent of the potential outcomes conditional on observed covariate values and the missing data patterns.

Another potential weakness of this method is that since it specifies one model for both handling missing data and estimating propensity scores it will not always possible to incorporate the outcome variable $Y$ into this model, even though it might provide useful information about missing values.

In addition, if there are many patterns of missing data with only few individuals for each of the two treatment groups, the pattern-mixture approach becomes infeasible. To overcome this possible complication, in our application we adopt a "trick". Because in our dataset most of the covariates are categorical, and only few continuous background variables have missing values, we turn into categorical the continuous variables with some missing data, and then apply the pattern-mixture approach simply by adding an additional "missing" category to each missing covariate.

Finally, we consider the MI approach. To create MIs we assume that the missing observations are missing at random (MAR), that is,

$$\Pr(R \mid \mathbf{X}, Z, Y(C), Y(T)) = \Pr(R \mid \mathbf{X}^{\text{obs}}, Z, Y^{\text{obs}}),$$

where $Y^{\text{obs}}$ is the vector of the observed values of $Y$, with $i$th element equal to $Y_i^{\text{obs}} = \text{I}\{Z_i = T\}Y_i(T) + \text{I}\{Z_i = C\}Y_i(C)$. Note that this MAR assumption involves all the observed variables, not only the pretreatment covariate $\mathbf{X}$, so that we can incorporate in the imputation model also the treatment indicator and the potential outcomes, which may provide useful information about the missing values. This is an important advantage of the MI techniques with respect to the other approaches. Here, we perform MI in two way: including $Y$ in the model and not including $Y$ in the imputation model. In this second case we implicitly assume that $\Pr(R \mid \mathbf{X}^{\text{obs}}, Z, Y^{\text{obs}}) = \Pr(R \mid \mathbf{X}^{\text{obs}}, Z)$.

Using MI to handling incomplete data covariates, we essentially assume the latent ignorability of the assignment mechanism. Latent ignorability was first introduced by Frangakis and Rubin (1999) as an extension of standard ignorability in the context of missing data mechanism. It describes a situation where the mechanism is ignorable only when conditional on certain latent or missing values, in addition to the observed values. In our case, the assignment mechanism is ignorable only conditional on complete covariate data (which includes, of course, values that in practice are missing). Computationally, this is achieved by filling in the missing covariate values using MI. Formally, we require that

$$\Pr(Z \mid \mathbf{X}, R, Y(C), Y(T)) = \Pr(Z \mid \mathbf{X}).$$

We conclude this section with some remarks about the computational techniques used to generate the results reported in section 7.

The propensity score analysis is implemented by the use of the pscore module in STATA written by Becker and Ichino (2000), which estimates the propensity scores and tests the balancing property. This property is needed to estimate the causal effect of interest, here, the Average effect of Treatment on the Treated (ATT). In fact, we know that under the strong ignorability assumption, if the propensity score $e(\mathbf{X})$ is known, the ATT can be estimated as follows:

$$\begin{aligned} \mathrm{ATT} &\equiv \mathrm{E}\big(Y_i(T) - Y_i(C) \mid Z_i = T\big) \\ &= \mathrm{E}\Big(\mathrm{E}\big(Y_i(T) - Y_i(C) \mid Z_i = T, e(\mathbf{X}_i)\big)\Big) \\ &= \mathrm{E}\Big(\mathrm{E}\big(Y_i(T) \mid Z_i = T, e(\mathbf{X}_i)\big)\mathrm{E}\big(Y_i(C) \mid Z_i = C, e(\mathbf{X}_i)\big) \mid Z_i = T\Big), \end{aligned} \tag{3.3}$$

where the outer expectation is over the conditional distribution of $e(\mathbf{X}_i)$ given $Z_i = T$.

Once we have estimated the propensity score for each individual, and tested that the balancing property holds, we may proceed to the matching stage, when treated and comparison units are paired according their scores. Here we use what is called “Stratification Matching”. This method consists of dividing the range of variation of the propensity score in intervals such that within each interval treated and control units have on average the same propensity score. Then, within each interval in which both treated and control units are present, the difference between the average outcomes of the treated and the controls is computed. The ATT of interest is finally obtained as an average of the ATT of each block with weights given by the distribution of treated units across blocks. Formally,

$$\mathrm{ATT} = \sum_{g=1}^{G} \mathrm{ATT}_g \frac{\sum_{i \in G(g)} Z_i}{\sum_i Z_i},$$

where $G(g)$ is the set of units in block g, and

$$\mathrm{ATT}_g = \frac{\sum_{i \in G(g)} Y_i(T)}{N_g^T} - \frac{\sum_{j \in G(g)} Y_j(C)}{N_g^C},$$

with $N_g^T$ and $N_g^C$ being the numbers of treated and control units in block $g$.

Recall that we have to face on the problem of partially missing covariates. Therefore, the estimation of this ATT effect will be based on the estimated generalized propensity score in the pattern-mixture approach. When we consider MIs to handling incomplete data, we estimate a standard propensity score and the corresponding ATT effect for each simulated complete datasets, and then we combine the results to yield the final estimate.

To create MIs we use the mvis module in STATA (Patrick Royston, 2004), which is based on MICE method of multiple multivariate imputation described by van Buuren et al. (1999) and assumes that the missing observations are missing at random (MAR) or missing completely at random (MCAR). MICE stands for Multivariate Imputation by Chained Equations. This method assumes that, for each missing variable, the user can specify a conditional distribution for the missing data given the other data. Under the assumption that a multivariate distribution exists from which these conditional distributions can be derived, MICE constructs a Gibbs sampler from the specified conditionals. The sampler is then used to generate MIs.[1] For each imputed dataset, we calculate propensity scores, pick a matched control group and calculate the causal estimand of interest. Finally, we combine causal estimates across imputed datasets using the method derived by Rubin (1987). Specifically, suppose that we have imputed $m$ complete datasets using an appropriate model. Let $\mathrm{ATT}_l$ and $V_l$ denote the point estimate and variance respectively from the $l$th $(l = 1, \ldots, m)$ dataset. The point estimate $\widehat{\mathrm{ATT}}$ of ATT from multiple imputation is simply the arithmetic mean of $\mathrm{ATT}_1, \ldots, \mathrm{ATT}_m$. Obtaining a valid standard error for this estimate of $\widehat{\mathrm{ATT}}$ requires combining information on within-imputation and between-imputation variation. The latter is important in reflecting uncertainty due to variability between imputation samples. First, a within-imputation variance

[1] MICE is a flexible and general methodology for generating multiple MIs in multivariate data. However, this approach has the theoretical limitation that it is possible to generate incompatible distributions via implicit contradictions in their conditional specifications. In other words, it is possible that there is no joint distribution for the variables which must be imputed, however the MCMC can be implemented, and each conditional specification may be a good empirical fit to the data. In addition, as most of the existing procedures, the mvis module restricts to particular and convenient multivariate distributions, which may result in inferior imputation, especially as the number of variables in the data increases. In our application, we expect the results to be robust to these types of limitations, even if we recognize that our MIs could be improved (e.g., Shen, 2000; Rubin, 2004).

component, $W$, is obtained as the mean of the complete-data variance estimates, $\mathrm{ATT}_1, \ldots, \mathrm{ATT}_m$. Second, a between-imputation variance component, $B$, is calculated as the sum of squares of $\mathrm{ATT}_1, \ldots, \mathrm{ATT}_m$ about $\widehat{\mathrm{ATT}}$, divided by m-1. The (total) variance of $\widehat{\mathrm{ATT}}$ is given by $V = W + B(1 + 1/m)$. Rubin (1987) showed that $(\mathrm{ATT} - \widehat{\mathrm{ATT}})/\sqrt{V}$ is distributed approximately as Student's $t$ with $\nu$ degrees of freedom, where $\nu = (m-1)\{1 + W/[B(1 + 1/m)]\}^2$.

In this paper, we implement the propensity score matching procedures described above in the attempt to assess the impact of childbearing on individuals'wellbeing in Indonesia. We do this using a sample of women drawn from the Indonesia Family Life Survey (IFLS).

## 4 Indonesia Family Life Survey

By the middle of the 1990s, Indonesia had enjoyed over three decades of remarkable social, economic, and demographic change and was on the cusp of joining the middle-income countries. Per capita income had risen more than fifteenfold since the early 1960s, from around US\$50 to more than US\$800. Increases in educational attainment and decreases in fertility and infant mortality over the same period reflected impressive investments in infrastructure.

In the late 1990s the economic outlook began to change as Indonesia was gripped by the economic crisis that affected much of Asia. In 1998 the rupiah collapsed, the economy went into a tailspin, and gross domestic product contracted by an estimated $12 - 15\%$ - a decline rivaling the magnitude of the Great Depression.

The general trend of several decades of economic progress followed by a few years of economic downturn masks considerable variation across the archipelago in the degree both of economic development and of economic setbacks related to the crisis. In part this heterogeneity reflects the great cultural and ethnic diversity of Indonesia, which in turn makes it a rich laboratory for research on a number of individual- and household-level behaviors and outcomes that interest social scientists.

The Indonesia Family Life Survey (IFLS) is designed to provide data for studying these behaviors and outcomes. It is an on-going multi-level longitudinal survey conducted by RAND Corporation in collaboration with UCLA, Lembaga Demografi (University of Indonesia), and the Center for Population and Policy Studies (CPPS) of the University of Gadjan Mada. The survey sample is representative of about 83% of the Indonesian population and contains over 30,000 individuals living in 13 of the 27 provinces in the country. The IFLS consists of three waves plus a special wave. The first wave (IFLS1) was administered in 1993 to individuals living in 7224 household. The second wave (IFLS2) sought to reinterview the same respondents four years later (1997). A follow-up (IFLS2+) was conducted in 1998 with 25% of the sample to measure the immediate impact of the economic and political crisis in Indonesia. The next wave, IFLS3, was fielded on the full sample in 2000. The response rate for the IFLS is impressive: 94% of the original sample from 1993 has been reinterviewed in the third wave of IFLS. In our application, we do not use data from IFLS2+.

The IFLS is a comprehensive multipurpose survey that collects data at both the household and individual levels, as well as at community level. In each wave, the questionnaire was organized in the same way: it was divided into *books*, usually addressed to different respondents, and subdivised into topical *modules*.

All the analyses in this paper are based on the subsample of panel women who responded to book IV. This book was administered to all ever-married women age 15-49 and those women who completed it in previous waves irrespective of age. Book IV collects retrospective life history on marriage, children ever born, pregnancy outcomes and health-related behavior during pregnancy and childbirth, infant feeding practice, and contraceptive use. In our sample, information from book IV is integrated with information on basic socio-demographic characteristic. In addition, we consider some information on key characteristics of household structure, and data concerning household consumption expenditures, that we use to define the individual wellbeing.

# 5 A Measure of Wellbeing

In our study we use a measure of monetary wellbeing, given by the real annual value of the total household consumption expenditures per equivalent basis. As the literature on developing countries suggests, we argue that consumption is better suited than income as an indicator of living standard in Indonesia. The main reason is that consumption is believed to vary more smoothly than income, both within any given year and across the life cycle. Income is notoriously subject to seasonal variability, particulary in developing countries, such as Indonesia, whereas consumption tends to be less variable. Life-cycle theories also predict that individuals will try to smooth their consumption across their low- and high-income years (in order to equalize their marginal utility of consumption across time), through appropriate borrowing and saving. In practice, however, consumption smoothing is far from perfect, in part due to imperfect access to commodity in credit markets and to difficulties in estimating precisely one's "permanent" or life-cycle income.

Sometimes, consumption is also preferred over income because it is deemed to be a more "direct" indicator of achievements and fulfilment of basic needs. A caveat is, however, that consumption is indeed an outcome of individual free choice, an outcome which may differ across individuals of the same income and ability to consume, just like actual functioning vary across people of the same capability sets. At a given capability to spend, some individuals may choose to consume less (or little), preferring instead to give to charity, to vow poverty, or to save in order to give important bequests to their children.

Consumption is also held to be more readily observed, recalled and measured than income (at least in developing countries), to suffer less from underreporting problem. Clearly, this is not to say that consumption is easy to measure accurately.

It should be noted that consumption does not equal expenditures. The value of consumption equals the sum of the expenditure on the goods and services purchased and consumed, plus the value of goods and services consumed but not purchased (such as those received as gifts and produced by the household itself), plus the consumption or services value of assets and durable goods owned. Unlikely expenditures, therefore, consumption includes the value of own-produced goods.

For Indonesia we do not have access to any pre-made consumption variables. There exists, however, SAS code to produce consumption expenditure for the first wave. This code, written by Nga Vuong, is available on the IFLS web site. Following Nga Vuong's SAS code, we wrote a STATA program to produce consumption expenditure for each wave. Some variables we used to generate consumption expenditure had missing values. As in Nga Vuong's SAS code, we replaced these missing values with the Enumeration Area medians based on household size, and if the values were still missing, we replaced them with the Kabupaten medians based on household size. Finally, in order to account for price variability, we considered real consumption expenditures, created by dividing the nominal consumption expenditures by the national consumption price index (International Financial Statistics, 2002).

In developing countries a strong positive correlation between household size and poverty is often reported. This implies that wellbeing of individuals who live in households of different sizes and composition are not directly comparable. In other words, differences in household size and composition can be expected to create differences in household "needs". It is essential to take these needs into account when comparing the wellbeing of individuals living in differing households. We can do this using equivalence scales. In our application, we consider a standard equivalence scale (at least for developing countries): the square root of the number of persons in the household.

Table 1 presents some descriptive statistics of the distribution of the (real) total net equivalised household consumption expenditures at the time of third wave (2000), classified by number of live births born between 1993 and 1997. We find that consumption for households experiencing one or two childbearing events tends to be lower than households who do not experience no childbearing event. The birth of the third child seems to increase the level of consumption; this figure is probably

Table 1: Descriptive statistics of total net equivalised household consumption by number of live births.

| | | Consumption expenditures (Rp in thousands) | | |
|---|---|---|---|---|
| Live births | Obs | mean | sd | median |
| 0 | 3024 | 194.084 | 211.816 | 136.539 |
| 1 | 948 | 163.026 | 168.507 | 119.842 |
| 2 | 128 | 151.812 | 195.366 | 118.244 |
| 3 | 7 | 199.538 | 129.990 | 127.870 |
| At least a live birth | 1083 | 161.936 | 171.604 | 119.827 |

due to the small number of women having three live births in four years. The same trend is evident in Figure 6, which shows the histograms of the distribution of the net equivalised household consumption expenditures for women who have no live birth and women who experience at least a live birth (excluded values greater than 99th percentile). The vertical line shows the mean of the consumption distribution within each subsample.

## 6 Causal Effects of Childbearing on Wellbeing

The descriptive statistics show interesting patterns of poverty for different groups in society. However, the reported statistics do not say much about whether - or to what extent - childbearing events may lead to changes in the consumption levels. For instance, in Figure 6 we notice that households with no live births between 1993 and 1997 have on average a higher consumption level respect to households

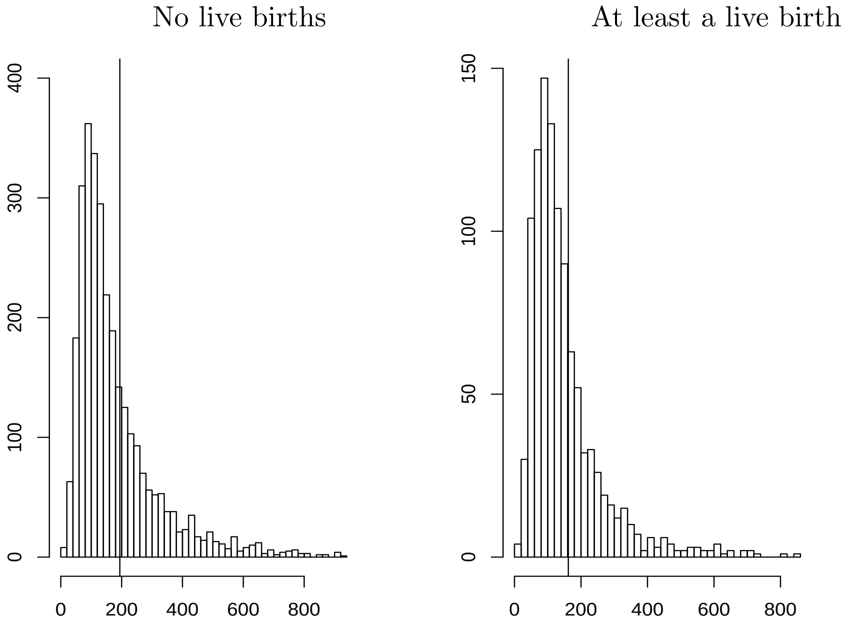

Figure 6: Histogram of the Net equivalised household consumption-Rupiah in thousands (excluding values above the 99th percentile).

who experienced at least a live births in that period, but it is unclear whether the lower consumption level of the latter household is a cause or a consequence of the childbearing. Nevertheless, from a social policy point of view this is an important issue: sensible policies aimed at reducing poverty and improving wellbeing, can only be successfully implemented as long as one knows the causal direction of the effects.

In this section we implement the methods described in section 3 with the aim of establishing whether childbearing events do have a causal impact on poverty.

To assess the impact of childbearing on wellbeing, we compare the wellbeing of women who experience at least a childbearing event, indicated by $Z = T$, to those women who do not experience such an event, indicated by $Z = C$. To a large extend this is what we have done in the presentation of the descriptive statistics in the previous section. However, a quick glance at Tables 2 and 3-4, which present descriptive statistics for the background variables by women who experience a childbearing and women who do not, demonstrates quite clearly that these two group of women are very different in almost all their characteristics. It is important to emphasize that these statistics are descriptive and not inferential, in the sense that they are not aimed at estimating relevant population parameters, but rather simply describe the two samples and their differences. Note that, our dataset suffers from the complication that for some women, some covariates have missing values.

Table 2 presents, for each continuous covariate, the mean and standard deviation using available cases; also presented are standardized percentage differences, defined as the mean difference between women who experience a childbearing event and women who do not, as a percentage of the standard deviation $100(\overline{x}_T - \overline{x}_C)/\sqrt{(s_T^2 + s_C^2)/2}$, where $\overline{x}_T$ and $\overline{x}_C$ are the samples means in the treated and control groups and $s_T^2$ and $s_C^2$ are the corresponding sample variances, again based on available cases. Also presented is the variance ratio, $s_T^2/s_C^2$.

Tables 3 and 4 present, using available case data for the categorical covariates the proportion of women in each category in the two groups defined by $Z$. The third column displays the absolute differences in percent between women who experience a childbearing event and women who do not for each of the categorical covariates. Three covariates, deprivation index, age at first marriage, and years since last live birth, were considered to be either continuous or categorical, depending on the specific approach to the missing data problem and thus appear in both Table 2 and Table 3 (Table 4).

Finally, Table 5 presents the proportion of observed values for the missing data indicators for the 13 covariates with any missing values (either continuous or categorical) by treatment group and the absolute differences in percent between women who experience a childbearing event and women who do not.

The differences between treated and control women summarized in Tables 2-5

Table 2: Means (standard deviations), standardized differences (based on available cases) in percent, and variance ratio for continuous covariate in both treatment groups before matching.

| | $Z = C$ | | $Z = T$ | | Standardized | Variance |
|---|---|---|---|---|---|---|
| Covariate | mean | (s.d.) | mean | (s.d.) | difference (%)[b] | ratio[c] |
| Number of adults | 3.12 | (1.43) | 2.76 | (1.27) | -27 | 0.78 |
| Number of children under 2 years | 0.34 | (0.53) | 0.49 | (0.57) | 28 | 1.16 |
| Number of children 2 to 6 years old | 0.35 | (0.55) | 0.56 | (0.62) | 35 | 1.25 |
| Number of children 6 to 14 years old | 1.33 | (1.11) | 1.06 | (1.13) | -24 | 1.03 |
| Number of women above 14 years | 1.62 | (0.88) | 1.43 | (0.74) | -23 | 0.71 |
| Deprivation index[a] | 0.30 | (0.15) | 0.34 | (0.15) | 20 | 1.03 |
| Consumption (Rp in thousands) | 151.67 | (171.13) | 129.78 | (125.40) | -15 | 0.54 |
| Age of HH head | 41.49 | (9.96) | 35.25 | (9.45) | -64 | 0.90 |
| Yrs of schooling of the HH head[a] | 5.88 | (4.24) | 6.11 | (4.32) | 5 | 1.04 |
| Age | 35.16 | (7.40) | 27.92 | (5.74) | -109 | 0.60 |
| Yrs of schooling[a] | 4.98 | (3.95) | 5.75 | (3.94) | 20 | 1.00 |
| Age at first marriage[a] | 17.89 | (4.34) | 18.49 | (4.15) | 14 | 0.91 |
| Years since last live birth[a] | 6.56 | (5.74) | 2.88 | (2.96) | -81 | 0.27 |
| Number of pregnancies (live births) | 3.60 | (2.28) | 2.55 | (1.97) | -49 | 0.74 |
| Number of miscarriage/still births | 0.32 | (0.69) | 0.23 | (0.53) | -14 | 0.59 |

[a] Covariate suffers from some missing data.

[b] The standardized difference is the mean difference as a percentage of the average standard deviation: $100(\overline{x}_T - \overline{x}_C)/\sqrt{(s_T^2 + s_C^2)/2}$, where for each covariate $\overline{x}_T$ and $\overline{x}_C$ are the samples means in the treated and control groups and $s_T^2$ and $s_C^2$ are the corresponding sample variance.

[c] The variance ratio is $s_T^2/s_C^2$.

indicate the possible extent of biased comparisons of outcomes due to different distributions of observed covariates and patterns of missing data in the two groups of women. That is, ideally all such descriptive statistics should suggest the same distribution in the group of women who have at least a live birth and in the group of women who do not, as they would be in expectation if the treatment indicator (childbearing vs no childbearing) had been randomly assigned. As can be seen from these tables, there exists considerable initial bias between women who experience a childbearing and women who do not. For instance, seven of the continuous covariates have initial standardized differences 25% larger, with age with an initial

Table 3: Table of observed proportions and percent differences for categorical covariate (household characteristics).

| Covariate | | $Z = C$ | $Z = T$ | Difference in % |
|---|---|---|---|---|
| Provincia | Nord Sumatra | 0.05 | 0.10 | 5 |
| | West Sumatra | 0.05 | 0.05 | 0 |
| | South Sumatra | 0.05 | 0.04 | 1 |
| | Lampung | 0.04 | 0.06 | 2 |
| | Jakarta | 0.09 | 0.08 | 1 |
| | West Java | 0.16 | 0.21 | 5 |
| | Central Java | 0.14 | 0.10 | 4 |
| | Yogjakarta | 0.05 | 0.03 | 2 |
| | East Java | 0.16 | 0.12 | 4 |
| | Bali | 0.06 | 0.04 | 2 |
| | West Nusa Tenggara | 0.05 | 0.08 | 3 |
| | South Kalimatan | 0.05 | 0.05 | 0 |
| | South Sulawesi | 0.05 | 0.04 | 1 |
| Area | Urnan | 0.47 | 0.41 | 6 |
| | Rural | 0.53 | 0.59 | 6 |
| Deprivation Index | $\leq 0.18$ | 0.21 | 0.14 | 7 |
| | $0.18 - 0.25$ | 0.2 | 0.21 | 1 |
| | $0.25 - 0.33$ | 0.21 | 0.18 | 3 |
| | $0.33 - 0.44$ | 0.19 | 0.23 | 4 |
| | $> 0.44$ | 0.19 | 0.24 | 5 |
| Sex of HH head | Female | 0.09 | 0.05 | 4 |
| | Male | 0.91 | 0.95 | 4 |
| Education level of HH head | None | 0.12 | 0.10 | 2 |
| | Elementary | 0.55 | 0.55 | 0 |
| | Jr High education | 0.13 | 0.13 | 0 |
| | Sr High education | 0.15 | 0.17 | 2 |
| | College or higher | 0.05 | 0.05 | 0 |
| Marital status of HH head | Married | 0.94 | 0.97 | 3 |
| | Unmarried | 0.06 | 0.03 | 3 |
| Activity last week HH head | Working | 0.93 | 0.94 | 1 |
| | Housekeeping | 0.03 | 0.02 | 1 |
| | Other | 0.04 | 0.04 | 0 |

Table 4: Table of observed proportions and percent differences for categorical covariate (woman's characteristics)

| Covariate | | $Z = C$ | $Z = T$ | Difference (%) |
|---|---|---|---|---|
| Is she HH head? | Yes | 0.08 | 0.03 | 5 |
| | No | 0.92 | 0.97 | 5 |
| Education level | None | 0.17 | 0.10 | 7 |
| | Elementary | 0.58 | 0.58 | 0 |
| | Jr High education | 0.12 | 0.15 | 3 |
| | Sr High education | 0.11 | 0.14 | 3 |
| | College or higher | 0.02 | 0.03 | 1 |
| Marital status | Married | 0.94 | 0.99 | 5 |
| | Unmarried | 0.06 | 0.01 | 5 |
| Activity last week | Working | 0.46 | 0.38 | 8 |
| | Housekeeping | 0.53 | 0.61 | 8 |
| | Other | 0.01 | 0.01 | 0 |
| Age at first marriage | $\leq 15$ | 0.29 | 0.22 | 7 |
| | 15 - 17 | 0.20 | 0.20 | 0 |
| | 17 - 19 | 0.19 | 0.21 | 2 |
| | 19 - 21 | 0.14 | 0.17 | 3 |
| | $> 21$ | 0.18 | 0.20 | 2 |
| Spouse in HH | Yes | 0.91 | 0.95 | 4 |
| | No | 0.09 | 0.05 | 4 |
| Islam | Yes | 0.87 | 0.90 | 3 |
| | No | 0.13 | 0.10 | 3 |
| Parents in HH | Yes | 0.10 | 0.13 | 3 |
| | No | 0.90 | 0.87 | 3 |
| Years since last live birth | No Children | 0.03 | 0.09 | 6 |
| | $\leq 2$ | 0.28 | 0.47 | 19 |
| | 2 - 4 | 0.15 | 0.23 | 8 |
| | 4 - 9 | 0.25 | 0.17 | 8 |
| | $> 9$ | 0.29 | 0.04 | 25 |
| Pregnant | Yes | 0.00 | 0.18 | 18 |
| | No | 1.00 | 0.82 | 18 |
| Ever used contraceptives | Yes | 0.77 | 0.71 | 6 |
| | No | 0.23 | 0.29 | 6 |
| Use of contraceptives | Yes | 0.59 | 0.42 | 17 |
| | No | 0.41 | 0.58 | 17 |

Table 5: Missing-value indicators (proportion observed).

| Covariate | $Z = C$ | $Z = T$ | Difference (in %) |
|---|---|---|---|
| Deprivation Index | 0.930 | 0.919 | 1.1 |
| Education level of HH head | 0.999 | 1.000 | 0.1 |
| Yrs of schooling of the HH head | 0.995 | 0.994 | 0.1 |
| Education level | 0.999 | 1.000 | 0.1 |
| Yrs of schooling | 0.997 | 0.995 | 0.2 |
| Activity last week | 0.998 | 1.000 | 0.8 |
| Age at first marriage | 0.985 | 0.993 | 0.7 |
| Islam | 0.996 | 0.997 | 0.1 |
| Parents in HH | 0.998 | 1.000 | 0.2 |
| Years since last live birth | 0.987 | 0.987 | 0.0 |
| Pregnant | 1.000 | 0.999 | 0.1 |
| Ever used contraceptives | 0.999 | 0.999 | 0.0 |
| Use of contraceptives | 0.998 | 0.997 | 0.1 |

standardized difference equal to 109%. Among categorical variables, we see that years since last live birth, pregnancy, and use of contraceptives are very different between women who experience a childbearing event and women who do not. The missing rates appear similar; there is no significant difference between the two groups in the observed proportions of covariates with some missing value. In addition, the observed proportions are quite high for each variables.

This unbalancing in observed background characteristics and pattern of missing covariates between women who have a child and women who do not implies that computed differences in wellbeing are highly likely to be confounded by these pretreatment variables and their missing-data pattern, a feature that needs to adjusted for. The ideal setting would be to compare a woman's level of wellbeing when experiencing a childbearing event to its counterfactual, which here would be the case when the same woman does not experience such an event. Such a comparison would enable us to single out the effect on wellbeing that is only attributable to the childbearing event. The problem of course is that for the same individual these two scenarios are mutually exclusive. In other words the counterfactual is indeed non observed, which clearly impedes such a comparison.

In order to address this problem we construct an approximation to the counterfactual using Propensity Score Matching methods described in section 3. Because the major goal of this article is to assess the impact of childbearing on wellbeing taking into account the presence of missing covariates when estimating propensity scores, we specifically focus on a comparison between the different approaches described in section 3.2 to estimating and using propensity scores with partially missing data.

In simple terms, the application of a quasi-experimental approach in our study can be outlined as follows. Women are divided into two types: those who had at least a live birth between 1993 and 1997 ($Z_i = T$) and those that did not ($Z_i = C$). Women are then matched by pairing units who undertook treatment (i.e., $Z_i = T$) with units of comparison (i.e., $Z_i = C$) that are similar in term of their observable characteristics prior to the childbearing event. When the relevant differences between treated and controls are captured by observable covariates, matching methods yield an unbiased estimate of the average impact of childbirth on treated. The matching is performed by means of the propensity score. Therefore, first we estimate propensity score for each individual, and then, provided the balancing property holds, which has to be tested, we proceed to compute the average effect of treatment on the treated.

The matching procedure based on the propensity score implies that all variables listed in Tables 2-4 and, according to the approach used, also the corresponding missing indicators listed in Tables 5, have to be balanced between treated and control units. Satisfying the balancing property is in our case a non-trivial exercise. This forced extensive use of interactions sometimes using higher order terms. In addition, the specification of the propensity score changes according to the approach used to handling the missing data problem and the more unbalanced the sample is, the greater need for interaction terms. In particular, when we consider the MI approach, for each imputed complete dataset, we might have to specify a different propensity score model. In all applied methods, the variables which are suspected to confound the fertility-poverty causal relationship are included in the estimation of the propensity score matching.

## 7 Results

This study started out with a real-world problem faced in most observational study: estimating and using propensity score with partially missing background data.

In our application, we have 4107 women. Of these women, 1083 (26.37%) had at least a live birth between 1993 and 1997, whereas 3024 (73.63%) had no live births. In addition, for 438 women (10.66%) - 122 treated and 316 control - some covariates had missing values.

We focused on three different possible solutions to this problem; specifically we compared the results obtained from the following methods: complete-case analysis, multiple imputation (with and without $Y$ in the imputation model), and pattern-mixture model approach. Table 6 shows the estimates of the average effects of childbearing on consumption expenditures (Rupiah in thousands) under these three different approaches of handling missing data.

As can be seen in Table 6, the results from different approaches of handling missing data show some differences between methods in terms of point estimate and its standard error. The two imputation models, one incorporating $Y$ and the other not incorporating $Y$, give very similar results. They yield values that appears to be slightly smaller than those resulting from the other approaches. On the contrary, the complete-case analysis gives the highest average treatment effect and the highest standard error. It seems clear that this approach is not the one we should use. As shown in many studies, complete case analysis can lead to biased and less efficient conclusions unless under Missing Completely At Random (MCAR) assumption (Rubin, 1976c; Little and Rubin, 1987). The MCAR assumption is fairly strong and in many cases implausible. It has testable implications and can be often rejected by the data. In our application, we find that the MCAR assumption is not plausible, in fact it is more reasonable to believe that the missing data mechanism is either Missing At Random (MAR) or nonignorable. For instance, if poorer women tend to have lower response rates, then the impact of an additional child from one period to the next will be stronger when considering the complete-case analysis. In addition, the exclusion of data from 10.66% participants reduce the power of statistical tests.

Concerning the multiple imputation approach and the model-based analysis, it is very hard to say which approach is better; they can be viewed as complementary to each other. Recall that we implemented the MI technique assuming that the missing data process was missing at random. This assumption could be questionable. On the other hand, the pattern-mixture approach does not make any assumption about the missing data mechanism, so it could appear more attractive. Theoretically, however, one could also implement the MI techniques in such a way as to satisfy the assumptions of the pattern-mixture approach.

In addition to the differences in assumptions, more broadly discussed in section 3.2, there are other differences between the three approaches, that we must consider in comparing them.

First, recall that propensity score models are generally chosen based on the balance they produce. This balance cannot be measured correctly using methods that consider only observations where all variables are observed, unless the appropriate assumption holds. Therefore model diagnostics may be misleading. The MI technique, with no added assumptions, can produce "completed" case diagnostics which can easily be combined across datasets. The pattern-mixture approach can be used to calculate expected values for each covariate.

Second, MI approach uses different models for imputation and propensity score. This allows the MI approach to incorporate model features in one model that might

Table 6: Average effects of childbearing on consumption expenditures (Rp in thousands) using different approaches to the missing covariates problem.

| Approach | Treated units | Control units | Effect | s.e. | t |
|---|---|---|---|---|---|
| Complete-data | 961 | 2387 | -29.990 | 13.615 | -2.203 |
| Pattern-mixture model | 1082 | 2670 | -28.563 | 13.527 | -2.112 |
| Multiple Imputation (without $Y$) | 1083 | 2625.375 | -26.305 | 13.014 | -2.021 |
| Multiple Imputation (with $Y$) | 1083 | 2628.5 | -26.087 | 13.012 | -2.005 |

be inappropriate for another. For example, we can include $Y$ in the imputation model to help predict missing covariate values, while inclusion of $Y$ in our propensity score model would be a strong violation of our assumptions.

Third, MI makes the choice of the propensity model easier. Missing data models such as the pattern-mixture model sometimes can be a bit tedious to fit and re-fit especially when there are many variables involved. Propensity score model fitting often involves iterations through many versions of the model using balance diagnostics to compare models. Using the MI technique the propensity score model fitting takes place on completed datasets. In our application, refitting logistic regressions in this scenario appears not nearly so cumbersome.

Finally, the MI approach allows for final analysis of the outcomes (such as covariance adjustment) which include covariates which are not fully observed. Often the causal estimand of interest is not simply a comparison of outcome means across treatment groups. For instance, treatment effects broken down by subgroups may be of interest. In addition, regression-adjusted results may be desired for increased precision. Such analyses may be difficult to perform if they involve missing covariate data; the MI technique easily handles any such analyses. Moreover, multiple imputation can be used as a tool for sensitivity analysis.

Despite these structural and philosophical differences, looking across the estimates in Table 6, we see that the various approaches of handling missing data lead to the same conclusion about the effect of having a child on wellbeing, that is, the estimated ATT effects suggest that having a child causes a non negligible reduction of consumption level, and all these effects appear to be significant at the 5% level according to a standard two-sided $t$-test. This would agree with the hypothesis that in developing countries, there is a strong positive effect of household size on poverty.

## 8 Concluding Remarks

In this paper, we analyze the causal relationship between fertility and individuals'welling in Indonesia using a sample of women drawn from the IFLS panel data, and propensity score methods. A complication is that for some women, some background variables have missing values. In such a case, it is not clear how the propensity score should be estimated. Here, we performed the data analysis using various strategies of handling this type of missing data. Specifically, we compare the following methods: complete-case analysis, multiple imputation, and a pattern mixture model based approach (Rosenbaum and Rubin, 1984).

As shown in many studies, complete case analysis can lead to biased estimate unless the MCAR assumption holds (Rubin, 1976c; Little and Rubin, 1987). The MCAR assumption, which is fairly strong and in many cases implausible, has several testable implication and can be often rejected. An attractive alternative to handling incomplete data is multiple imputation. In our application, we perform a MI analysis using a technique based on MICE method, and assuming that the missing data mechanism is Missing At Random (Rubin, 1976c; Little and Rubin, 1987). Finally, we consider the pattern-mixture approach proposed by Rosenbaum and Rubin (1984), which leads to estimate a "generalized" propensity score.

The results from different approaches of handling missing data show some differences between methods in terms of point estimate and its standard error; however all the three approaches provide some evidence that childbearing events have a negative impact on individuals'wellbeing, which gives more weight to this conclusion.

It is important to be aware that the analysis presented here has some shortcomings. First, we have focused on a binary treatment,"having or not at least a live birth between 1993 and 1997", but it could be of interest to assess if wellbeing associated with childbearing also varies by parity.

Also, we apply the pattern mixture approach for propensity score estimation with covariate missing data using only categorical variables. We could consider also continuous covariates. In such a case, Rosenbaum and Rubin (1984) suggested that in large enough samples, one can estimate the generalized propensity score by estimating a separate logit model using the subset of covariates fully observed for each pattern of missing data. However, this approach is not applicable when there are many patterns of missing covariates with only few individuals for each of treatment groups. In such situations, it would be more appropriate to use models similar to those in D'Agostino and Rubin (2000). Moreover, it would be also interesting to investigate results from different multiple imputation approaches, which rely on weaker assumptions than the MAR model (e.g., Hill, 2004).

Finally, it should be noted that our analysis captures only the total causal effect of a childbearing event. It could be interesting to assess if this effect is mediated by some intermediate variable, such as labor market behavior, mother's health status, father's activity and so on. Thus, the concept of "direct" and "indirect" causal effects comes to play. In the Rubin Causal Model potential outcomes framework, here adopted, the key organizing principle for addressing the topic of direct and indirect causal effects is based on the concept of principal stratification (Frangakis and Rubin, 2002; Mealli and Rubin, 2003). This perspective can be view as having its seeds in the instrumental variables method of estimation, as described within the context of the Rubin Causal Model in Angrist, Imbens, and Rubin (1996).

For having an idea about how this approach could be implemented in our study, suppose that we are interested in the causal effect of having at least a live birth between 1993 and 1997, $Z$, on the mother's health at time of second wave (1997), $S$, and on wellbeing, $Y$ three years after the childbearing events. In this case, for each subject $i$, there are two sets of potential outcomes: $\{S_i(C), Y_i(C)\}$, which would be observed if the woman does not experience childbearing events, and $\{S_i(T), Y_i(T)\}$ which would be observed if the woman experiences at least a childbearing event.

Assuming that we are within a cell defined by common values of observed covariates, the basic principal stratification with respect to the general posttreatment variable $S$ is the partition of subjects into sets such that all subjects in the same set have the same vector $(S(C), S(T))$, where $S(C)$ refers to the value of $S$ when assigned to treatment $C$ and $S(T)$ refers to the value of $S$ when assigned to treatment $T$.

Note that we cannot, in general, observe the principal stratum to which a subject belongs because we cannot directly observe both $S(C)$ and $S(T)$ for any subject. Principal strata are, by definition, not affected by treatment assignment, and can thus be used just as any other pretreatment variable to define subgroup of causal effects.

A comparison of causal effects of $Z$ on wellbeing for different principal strata defined by $(S(C), S(T))$ provides information on the extend to which a causal effect of $Z$ on wellbeing occurs together with a causal effect of $Z$ on the intermediate outcome, mother's health. A direct causal effect of childbearing events, after controlling for the mother's health, exists if there is a causal effect of childbearing events for women with $S(C) = S(T)$, i.e., women for whom the treatment $Z$ does not affect their health status. On the other hand, if there is no causal effect of childbearing events on wellbeing for these women, then there is no direct effect of childbearing events on wellbeing after controlling for mother's health status, because the causal effects of the treatment $Z$ on wellbeing exists only in presence of causal effects of $Z$ on the posttreatment variable, mother's health status.

# References

Aassve, A., Mencarini, L., and Mazzucco, S. (2003a) Poverty and household dynamics in the European Union: an application of the Difference-in-Difference estimator with matching. Department of Economics, University of Leicester, *Mimeo*.

Aassve, A., Burgess, S., and Propper, C. (2003b) Modeling poverty transitions as the outcome of employment, family unions, and childbearing decisions in the United Kingdom. Department of Economics, University of Leicester, *Mimeo*.

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.

Bane, M. J., and Ellwood D. (1986) Slipping into and out poverty: the dynamics of spells. *Journal of Human Resources*, **21**, 1–23.

Baulch, B. and Hoddinott J. (Eds.) (2000) Economic mobility and poverty dynamics in developing countries. *Journal of Development Studies*, **36(6)**.

Becker, S., and Ichino, A. (2002) Estimation of average treatment effects based on propensity scores. *The STATA Journal*.

Birdsall, N. A. C., Kelley and Sinding, S. W. (2001) *Population matter: demographic change, economic growth, and poverty in the developing world*. Oxford University Press, Part III. Fertility, Poverty, and the Family.

D'Agostino R. A. (1998) Propensity score methods for bias reduction in the comparison of a treatment to a nonrandomized control group. *Statistics in Medicine*, **17**, 225–281.

D'Agostino R. A., and Rubin, D. B. (2000) Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*,

**95**, 749–759.

Decorn, S., and Krishman, P. (2000) Vulnerability, seasonality and poverty in Ethiopia. *Journal of Development Studies*, **36(6)**.

Dehejia, R., and Wahba, S. (1999) Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, **94**, 448, 1053–1062.

Frangakis, C. E., and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

Gu, X. S., and Rosenbaum P. R . (1993) Comparison of multivariable matching methods; structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, **2**, 405–420.

Hill J. (2004) Reducing bias in treatment effect estimation in observational studies suffering from missing data. Working paper series, *School of International and Public Affairs*, Columbia University, NY.

Huff Stevens, A. (1999) Climbing out of poverty, falling back in: measuring the persistence of poverty over multiple spells. *Journal of Human Resources*, **34(3)**, 557–588.

International Financial Statistics Yearbook 2002, IMF

Jalan, J., and Ravallion M. (2000) Is transient poverty different? Evidence for rural China. *Journal of Development Studies*, **36(6)**.

Klepinger, D., Lundberg, S. and Plotnick, R. (1995) Instrumental selection: the case of teenage childbearing and women's educational attainment. *DP 1077-95*, 1995 29 pp.

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., and Rubin D. B. (1987) *Statistical analysis with missing data*, Wiley, New York (2nd edition in 2002).

Mcculloch, N., and Baulch, B. (2000) Stimulating the impact of policy upon chronic and transitory poverty in rural Pakistan. *Journal of Development Studies*, **36(6)**.

Mealli, F. and Rubin, D. B. (2003) Assumptions allowing the estimation of direct causal effects: Commentary on "Health, wealth, and wise? Test for direct causal paths between health and socioeconomic status" by Adams and al. *Journal of Econometrics*, **112**, 79–87.

Muffels, R. J. A. (2000) Dynamics of poverty and determinants of poverty transitions: result from the Dutch socioeconomic panels. in Rose D. (ed.), *Researching social and economic change. The use of household panel studies*, Routledge, New York.

Royston P. (2004) Multiple imputation of missing values *Stata Journal*, **4**, 227–241.

Rose D. (ed.) *Researching social and economic change. The use of household panel studies*, Routledge, New York.

Rosenbaum, P. T., and Rubin D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, **70(1)**, 41–55.

Rosenbaum, P. T., and Rubin D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516–524.

Rubin, D. B. (1976a) Matching to remove bias in observational studies. *Biometrics*, **29**, 185–203; Prints correction (1974), **30**, 728.

Rubin, D. B. (1976b) Multivariate matching methods that are equal percent bias reducing: some examples. *Biometrics*, **32**, 109–120.

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

Rubin, D. B. (1978) Multiple imputation in sample survey: a phenomenological Bayesian approach to nonresponse. *The Proceedings of Survey Research Methods Section of the American Statistical Association*, 20–34.

Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys.* Wiley, New York.

Rubin, D. B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, **91**, 473–489.

Rubin,D. B. (1997) "Estimating causal effect from large data sets using propensity scores." *Annals of Internal Medicine*,**127**, 757–763.

Rubin, D. B. (2003) Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, **57**, 3–18.

Rubin, D. B., and Thomas N. (1992a) Affinely invariant matching methods with ellipsoidal distribution. *The Annals of Statistics*, **20**, 1079–1093.

Rubin, D. B., and Thomas N. (1992b) Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biomatrika*, **79**, 797–809.

Rubin, D. B., and Thomas N. (1996) Matching using estimated propensity score; relating theory to practice. *Biomatrics*, **52**, 249–264.

Schoumaker, B., and Tabutin, D. (1999) *Relationship between poverty and fertility in Southern countries*. Knowledge, methodology and cases, WP.2, Department of Science of population and development, Université Catholique de Louvain.

Shen, Z. (2000) *Nested multiple imputation.* Ph.D. Thesis, Harvard University, Cambridge, MA.