



UNIVERSITÀ
DEGLI STUDI
FIRENZE



UNIVERSITÀ
DEGLI STUDI
DI PERUGIA



Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

**DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA
CURRICULUM IN STATISTICA
CICLO XXXV**

Sede amministrativa Università degli Studi di Firenze
Coordinatore Prof. Matteo Focardi

**Principal Stratification:
novel declinations and uses**

Settore Scientifico Disciplinare SECS-S/01

Dottorando:
Silvia Noirjean

Tutore
Prof. Fabrizia Mealli

Coordinatore
Prof. Matteo Focardi

Anni 2019/2022

Acknowledgements

There are several people I would like to thank for contributing to make this possible.

First, I would like to thank my supervisor, Fabrizia Mealli, for her guidance and advice. When I took her course on causal inference during the Master's program, I was immediately impressed by her exceptional ability to teach and convey her passion and commitment. I recently attended one of her lectures and, after 4 years, I had exactly the same feeling. I learned a lot from her, about causal inference and much more.

I would also like to thank Alessandra Mattei, another amazing Professor and woman. I am really grateful to her for the time she has devoted to me in the past few years. She was always there when I had a question and, no matter how silly, she always answered me promptly and with the extreme kindness that distinguishes her. Having the opportunity to work with her has been a blessing.

I also sincerely thank Ian White, who supervised me during my four-month visiting period at the MRC Clinical Trial Unit at UCL. The time under his supervision has been extremely beneficial for my research project and meant a lot to me personally.

Last but not least, I would like to thank the colleagues with whom I shared this experience. I did value every small talk, coffee, and sharing of thoughts and fears.

To Marco.

Contents

Acknowledgements	iii
Preface	xiii
1 Principal Stratification in the absence of interference	1
1.1 Setting and notation	1
1.2 Rubin Causal Model	2
1.3 The Principal Stratification framework	5
1.3.1 Super-population and finite-sample estimands	6
1.3.2 Applications	7
Non-compliance	7
Surrogate endpoints	8
Direct and indirect causal pathways	9
Missing outcome data	10
Truncation by death	10
1.4 Identifying assumptions	12
1.4.1 Assumptions restricting the number of principal strata	13
1.4.2 Assumptions on the equality of outcomes	16
1.4.3 Ranked Average Score/Stochastic Dominance	17
1.4.4 Assumptions that involve the existence of an instrument for the intermediate variable	18
1.4.5 Assumptions that rely on secondary outcomes	20
1.4.6 Assumptions that rely on pre-treatment auxiliary variables	20
Principal Ignorability assumptions	21
Auxiliary independence assumptions	22
1.4.7 Distributional assumptions	22
1.4.8 The role of the pre-treatment covariates	24
1.5 Bayesian model-based Principal Stratification	25

1.6	Sensitivity analyses	28
2	The Oyam Health Expenditures Study	31
2.1	Research aim and objectives	31
2.2	The intervention of Community Health Financing	33
2.2.1	Study design	34
2.2.2	Data	35
2.2.3	The outcome	37
2.3	Methodology	38
2.3.1	Basic notation and potential outcomes	38
2.3.2	Principal strata and definition of Principal Causal Effects . . .	39
2.3.3	Identification issues	40
2.3.4	Instrumental Variables approach	42
2.4	Analyses	43
2.4.1	Parametric models	43
2.4.2	Assuming and relaxing the Exclusion Restriction for Never Takers	47
2.4.3	Bayesian inference	48
2.5	Results	49
2.5.1	With Exclusion Restriction for Never Takers	49
2.5.2	Without Exclusion Restriction for Never Takers	52
2.6	Posterior predictive checks	55
2.7	Conclusions	56
3	Non-inferiority trials with non-adherence: a Simulation Study	59
3.1	Research aim and objectives	59
3.2	Methodology	62
3.2.1	Basic notation and potential outcomes	62
3.2.2	Principal strata	63
3.2.3	Definition and meaning of Principal Causal Effects	65
3.2.4	Relevance of Principal Causal Effects	66
3.2.5	Commonly invoked assumptions for identification: a critical discussion	67
3.2.6	Estimation strategies	69
3.3	Simulations	71
3.3.1	Planning	71

3.3.2	Results	75
3.4	Practical considerations	77
3.4.1	Description of the principal strata	77
3.4.2	Power	77
3.4.3	Weak predictors of adherence	79
3.4.4	Sensitivity to model misspecifications	81
3.5	Conclusions	82
4	Principal Stratification in the presence of network interference: an illustrative application to the Florentine Study	85
4.1	Research aim and objectives	85
4.2	Experimental design and data	89
4.2.1	Descriptive statistics	91
4.3	Methodology	93
4.3.1	Network and neighborhood treatment	93
4.3.2	Potential outcomes	94
4.3.3	Principal strata	97
4.3.4	Overall Principal Causal Effects	98
4.3.5	Disentangling Spillovers from overall Principal Causal Effects Through Mediation	100
4.3.6	Identification issues	103
4.4	Analyses	106
4.4.1	Parametric Models	106
4.4.2	Bayesian Inference	109
4.5	Results	112
4.5.1	Principal Strata	112
4.5.2	Overall Principal Causal Effects and Principal Natural Direct and Indirect Effects	113
4.5.3	Principal Controlled Direct and Spillover Effects	116
4.6	Posterior predictive checks	118
4.7	Conclusions	119
5	Final remarks	123
A	Appendix to the Oyam Health Expenditures Study	125
A.1	Assessing the validity of the Exclusion Restriction for Never Takers	125

A.2	Convergence checks	126
A.3	Additional figures	130
A.4	Supplementary analysis	130
A.4.1	The outcome	130
A.4.2	Methodology	131
A.4.3	Results	132
B	Appendix to the Simulation Study	135
B.1	Details of the calculations	135
B.1.1	ITT	135
B.1.2	PP	137
B.1.3	IPW	138
B.1.4	Principal scores	138
B.1.5	Bayesian model-based PS analysis	139
B.2	Convergence checks	141
C	Appendix to the Florentine Study	147
C.1	Convergence checks	147

List of Abbreviations

AACE	A lways A dherents C ausal E ffect
AT	A lways T akers
AUROC	A rea U nder the R eciever O perating C haracteristic
Bulk-ESS	B ulk- E ffective S ample S ize
C	C ompliers
CACE	C ompliers A verage C ausal E ffect
CDE	C ontrolled D irect E ffect
CED	C luster E ncouragement D esign
CHEE	C atastrophic H ealth E xpenditures E ffect
CHF	C ommunity H ealth F inancing
CONSORT	C ONSolidated S tandards O f R eporting T rials
CSE	C ontrolled S pillover E ffect
D	D efiers
DA	D ata A ugmentation
DGM	D ata G enerating M odel
EM	E xpectation M aximization
ER	E xclusion R estriction
FDA	F ood and D rug A dministration
HEE	H ealth E xpenditures E ffect
ICH	I nternational C ouncil of H armonsation
IPW	I nverse- P robability W eighting
ITT	I ntention- T o- T reat
IV	I nstrumental V ariables
LATE	L ate A verage T reatment E ffect
LMICs	L ow- and M iddle- I ncome C ountries
NACE	N on- A dherents C ausal E ffect
NDE	N atural D irect E ffect
NI	N on I nferiority
NIE	N atural I ndirect E ffect
NT	N ever T akers
OHES	O yam H ealth E xpenditures S tudy
PC	P resentation C ompliers
PCE	P rincipal C ausal E ffect
PI	P rincipal I gnorability
PPPV	P osterior P redictive P - V alue

PS	P roduct S trata S trati f ication
RC	R eward C ompliers
RCM	R ubin C ausal M odel
SE	S tandard E rror
SoC	S tandard-of- C are
SUTNVA	S table U nit T reatment on N eighborhood V alue A ssumption
SUTVA	S table U nit T reatment V alue A ssumption
UHC	U niversal H ealth C overage
ZHEE	Z ero H ealth E xpenditures E ffect
ZIP	Z ero-Inflated P oisson

Preface

An observational study resembles an experiment in that treated and control groups are compared to estimate the effects of a treatment; it differs in that the assignment of the treatment to units is not under experimental control. Treatment groups are either “self-selected” (e.g., smokers or non-smokers) or selected by some administrative decision (e.g., recipient or not recipient of financial aid) or external agent (e.g., having or not having a certain disease). As a result, they are typically substantially different. If the differences are in characteristics related to the outcome of interest, we are in the presence of *confounders*. Confounders are a serious issue in that they represent a source of variation in the outcome, other than the treatment, which is distributed differently among the compared populations. Thus, their presence can mask or confound the relation between the treatment and the outcome, complicating causal attribution or leading to potentially incorrect inferences. In light of this, it is intuitive that to estimate causal effects in observational studies, it is crucial to assume that the treatment assignment is *unconfounded*, that is, independent of the outcome given the observed covariates (Rubin, 1990). In other words, it is necessary to assume that there are no unobserved confounders of the relationship between treatment assignment and outcome. This assumption is what ensures that contrasts between the outcomes of units with the same values of the relevant observed pre-treatment variables, but undergoing different treatment levels, can be interpreted as credible causal effects estimates. Unfortunately, unconfoundedness of treatment assignment can be sometimes hard to believe. Careful planning of the study is paramount in making it more plausible. This usually involves listing the most relevant confounding variables, designing the study¹ to record them, and adopting some confounding adjustment techniques (such as stratification, matching, covariance adjustment) to reduce or eliminate the bias they may cause. However,

¹The design of an observational study consists of all activities that precede the examination or use of those outcome measures that will be the basis for the study’s conclusions (e.g., Rosenbaum, 2010)

this is not sufficient to completely remove the bias unless all confounders are observed (e.g., Rosenbaum, 1991). Some sensitivity analyses had been developed for the presence of unobserved confounding factors, the purpose of which is to assess the extent of bias (more precisely, the residual confounding effect of an unmeasured variable) that might alter the study conclusions (e.g., Copas & Li, 1997; Cornfield et al., 1959; Ding & VanderWeele, 2016; Imbens, 2003; Lin et al., 1998; Rosenbaum & Rubin, 1983; VanderWeele & Arah, 2011; VanderWeele & Ding, 2017).

In randomized experiments, on the contrary, randomization of treatment assignment prevents the existence of systematic differences between treated and control units if the sample is large enough (Fisher, 1936). In practice, sources of variation other than the treatment that might affect the outcome, whether observed or unobserved, are balanced out in expectation. Thus, outcome differences between treatment groups can be unambiguously attributed to the treatment only. This is one of the greatest benefits conferred by conducting a randomized experiment rather than an observational study. For this and other reasons, Cochran (1972) views observational studies as attempts to answer causal questions in settings “in which we would like to do an experiment but cannot” because it is impractical, unethical, or untimely. The concept of observational studies as an attempt to emulate randomized experiments was shared also by other statisticians and epidemiologists. Hernán (2021), for instance, referring to the specific case of clinical trials, affirmed that “causal inference from observational data can be viewed as an attempt to emulate a target trial” when an appropriate target trial does not exist.

Despite the advantage conferred by randomization, randomized experiments are far from being without complications. First and foremost, the fact that when receipt of the treatment requires individuals to take, or subject themselves to, a particular action, such as undergoing surgery or entering a job-training program, it is common that some of them do not comply with the assignment. Typically, these individuals differ systematically in characteristics associated with the outcome compared with those who choose/can comply. Hence, in randomized experiments with non-compliance, treatment assignment is unconfounded by design, but treatment receipt typically is not. In light of this, it is clear why Imbens and Rubin (1997) affirmed that they viewed randomized trials with non-compliance as a “a bridge between randomized trials and observational studies”.

Under the unconfoundedness of treatment assignment, it is possible and relatively

straightforward to unbiasedly estimate the causal effects of being assigned to the treatment. An Intention-To-Treat (ITT) analysis, in which units are analyzed according to their initial allocation ignoring information about their compliance behavior, allows such an evaluation. The problem is that these causal effects are usually not the causal effects of primary interest, which are, instead, the effects of the receipt of treatment. For example, if the experimental treatment was a new drug and the control was the standard-of-care (SoC), ITT would measure the effect of prescribing the experimental treatment (versus the SoC), rather than the effect of actually taking it. That is, it would measure the effect of the treatment policy, rather than drug efficacy (e.g., Nagelkerke et al., 2000). Assumptions that allow researchers to attribute these causal effects to the receipt of treatment – in the example, the uptake of the drug – are at the core of the Instrumental Variables (IV) approach. The latter is a set of methods developed in econometrics starting in the 1920s, and originally exploited to draw causal inferences in settings where the treatment cannot be credibly viewed as randomly assigned. It assumes the existence of a variable, called *instrument* or *instrumental variable*, which affects the exposure but not directly the outcome. Angrist et al. (1996) showed how this instrument acts as the initial random assignment in randomized experiments with non-compliance, contributing to the parallel between the latter and observational studies. With this in mind, treatment assignment, i.e., the instrument, would influence treatment receipt but not directly the outcome. This assumption is called *Exclusion Restriction* (ER). It is not implied by randomization, so its validity needs to be assessed on a case-by-case level. Concerns about the possible presence of direct effects of treatment assignment on the outcome are what motivates the use of placebos and double blinding in clinical trials. In double-blind, placebo-controlled studies, where neither the unit nor the researcher is aware of the assigned treatment, the assumption can be justified; however, it is often controversial in different settings, and usually stated on the basis of subject matter knowledge.

Principal Stratification (PS) allows, in some circumstances, relaxing the ER. The PS framework was proposed by Frangakis and Rubin (2002) to evaluate treatment effects while addressing post-assignment or post-treatment complications, such as non-compliance. In general, it assumes that units can be partitioned into latent subgroups, called *principal strata*. In case it is used to face the issue of non-compliance, principal stratum membership is based on the individual compliance behavior in response to the assigned and to the *counterfactual*, i.e., unassigned, treatment. The

impossibility of knowing how the subjects would have behaved if they had been assigned to the counterfactual treatment is what makes principal strata latent. The key advantage, on the other hand, is that although unconditionally receipt of treatment is confounded, within principal strata the receipt of treatment is unconfounded. This allows for the proper definition of stratum-specific causal effects, called *principal effects*.

Another major complication that can plague randomized experiments just like observational studies is interference between the units. In the presence of interference, the outcome for a given unit is unaffected by the treatment status of other units. Many studies are designed to isolate participants from one another so that the presence of interference can be reasonably ruled out. However, there are exceptions. First, because it is highly impractical in some contexts to impose the no-interference assumption through the experimental design; second, because the way in which subjects influence each other is of intrinsic interest, and is not a mere complication to be addressed in the analysis phase.

We use the PS framework for addressing post-assignment complications of different natures arising in experimental or quasi-experimental studies in the fields of development economics, medicine, and education. In the first two applications, there is non-compliance with, or non-adherence to, the treatment assignment, but the presence of interference between the units can be reasonably ruled out. The last application, however, is to a Cluster Encouragement Design (CED) in which there is non-compliance with treatment assignment and the units are also likely to influence each other. This motivated us to develop an innovative approach to the analysis of CEDs that also allows for this possibility to be properly accounted for. The latter blends concepts and tools from PS, mediation analysis, and causal inference with network data.

First, we conduct an impact evaluation of the “Oyam Health Expenditures Study” (OHES), a pilot intervention of Community Health Financing (CHF) implemented in the rural district of Oyam (Uganda), where households were assigned to attend sensitization sessions and then receive the offer to enroll in a CHF scheme. In this application, it is reasonable to assume that households do not interfere with each other. However, given the voluntary nature of enrollment in the program, non-compliance is present, meaning that not all the households that attended the sensitization sessions and received the offer to enroll in the CHF scheme actually

accepted it. Although the statistical literature is abundant with studies in which PS is used to analyze experimental data characterized by non-compliance, in the economics literature the use of the IV approach is still well-established. In this application, we show the consequences of violations of the ER on which the IV method relies by performing Bayesian model-based PS analyses by first assuming and then relaxing the ER for *Never Takers*, i.e., the principal stratum of those who would not enroll in the CHF scheme regardless of their assignment. By relaxing this assumption, it is possible to assess whether for this principal stratum there are causal effects of treatment assignment that do not pass through the actual enrollment into the CHF scheme (i.e., the treatment receipt). Although this possibility is ruled out when using the IV approach, it is of paramount importance in terms of policy-making. In fact, such causal effects would indicate that sensitization campaigns already have some effect, even if a CHF scheme is not implemented in practice; and, the implementation of sensitization campaigns would be much easier and less costly than that of CHF schemes.

As a second novel application, we propose the PS framework to conceptualize the problem of treatment non-adherence in non-inferiority trials (NI). NI trials are designed to test whether a new experimental treatment is not unacceptably worse, in terms of efficacy, than an active control already in use. The rationale beyond these trials is that the new treatment may not be more effective than currently available standard treatments, but it may offer important advantages such as an improved safety profile, greater ease of administration, better compliance, or reduced cost (for a review, see Schumi & Wittes, 2011). In clinical trials, in general, it is very common that some subjects do not adhere to the assigned treatment due to, for example, side effects, perceived lack of efficacy, or intake of rescue medication. In NI trials, treatment non-adherence makes standard measures of treatment benefit, such as the ITT principle, more likely to conclude the NI of the experimental treatment relative to the SoC (Food and Drug Administration, 2016). Although this is now well known in the literature, to date there is no general directive on the best strategy to handle the problem. We recognize that the ambiguity stems primarily from the lack of consensus on the estimand which is more relevant to targeting, which in turn depends on the question one wants to address. We believe that a clinical question of paramount scientific interest concerns the causal effect for the principal stratum of patients who would adhere to both the treatments being compared. Thus, in line with the principal stratum strategy listed in the ICH E9(R1) guideline (European

Medicines Agency, 2020), we provide a formal definition of the corresponding estimand, dwelling on its interpretation and relevance. We also perform simulation studies to provide a numerical evaluation of the impact of choosing to target an estimand over another, comparing the results obtained with different methods and showing that they often lead to opposite NI conclusions. In addition, these simulations also allow showing that a Bayesian model-based estimator of the principal stratum estimand we propose to target is almost unbiased and has nominal coverage probabilities. We also provide some important practical considerations on the advantages and limitations of using this estimator, in terms, for instance, of required assumptions and resulting power.

The third and final application of PS we present is on the “Florentine study”, a field experiment conducted in Florence (Italy), and aimed at assessing how appropriate incentives randomized to high school classes may enhance teens’ museum attendance. Specifically, students in these classes received one of three incremental encouragements to visit a specific museum in the city center, *Palazzo Vecchio*. The rationale was that if students made the proposed visit, they might be prompted to visit other museums in the future. Therefore, the actual visit to Palazzo Vecchio is not the outcome of interest, but rather the treatment. The outcome is the number of museums visited in the follow-up period. Non-compliance is present in that only a minority of the students performed the proposed visit to Palazzo Vecchio. Moreover, given the CED, we can expect that among students in the same class who interact with one another, the treatment actually received by one subject may affect other subjects’ outcomes, generating interference or spillover effects. The disentanglement of the different causal channels through which the intervention works, including spillovers, would provide an in-depth understanding of how the intervention works and allow for future improvements. On the one hand, we use PS to address the issue of non-compliance, thus defining and estimating stratum-specific causal effects. On the other hand, we exploit the distinguishing feature of this data: the availability of information about the friendship ties between the teens belonging to the same class. In particular, in line with a recent strand of literature on causal inference in the presence of networks (Del Prete et al., 2020; Forastiere, Airolidi, et al., 2021; Forastiere et al., 2018), we assume that interference is conveyed by these networks. Moreover, we postulate that a synthesis of the treatment of the units in the individual network may act as a mediator, so that part of the stratum-specific effects passes through the treatment of the neighbors, i.e., the visits to Palazzo Vecchio of friends. We

consider this causal channel as the source of spillover effects. Thus, we conduct a principal stratum mediation analysis by formally defining stratum-specific natural direct and indirect effects and controlled direct and spillover effects and using them to disentangle the different causal channels responsible for the global effect of the intervention.

Outline of the Thesis

The Thesis is organized as follows. Chapter 1 illustrates the PS framework and provides an extensive review of its common declinations and uses in contexts where it can be assumed that there is no interference between units. Chapter 2 reports the analyses of the OHES study. Chapter 3 describes our proposal to adopt a PS framework in the context of NI trials with non-adherence, also showing the simulation study results. Chapter 4 focuses on the problem of interference and explains our methodological proposal for the analysis of CEDs under network interference by using the Florentine study as an illustrative example. Chapter 5 concludes.

Chapter 1

Principal Stratification in the absence of interference

1.1 Setting and notation

Suppose to conduct a randomized or nonrandomized experiment to assess the causal effect of a treatment on a target population. Consider having a sample of n units, labeled $i = 1, 2, \dots, n$, drawn from this population. For each unit i , we observe a K -dimensional vector of pre-treatment variables, $\mathbf{X}_i$, the assigned treatment level, Z_i , the outcome of primary interest, Y_i , and a post-treatment variable, D_i , which is deemed to be useful in understanding the underlying causal mechanism and in explaining treatment effect heterogeneity. Frequently D_i is an intermediate variable, lying in the path between the treatment assignment and the outcome of interest.

In this Chapter, for ease of presentation, we refer to causal studies in which only two treatment levels are compared. In particular, $Z_i \in \mathcal{Z} \equiv \{0, 1\}$, with $Z_i = 1$ indicating that unit i is assigned or exposed to an active/new treatment, and $Z_i = 0$ indicating that it is assigned or exposed to the control/standard treatment. The intermediate variable and the outcome of primary interest, on the other hand, can be binary, categorical, continuous, or any type, $D_i \in \mathcal{D}$ and $Y_i \in \mathcal{Y}$.

We denote with $\mathbf{Z}$, $\mathbf{D}$ and $\mathbf{Y}$ the n -dimensional vectors of, respectively, the observed treatment assignments, values of the intermediate variable and outcomes for the whole sample, with i -th elements equal to Z_i , D_i and Y_i . We denote with $\mathbf{X}$ the $n \times K$ matrix of observed pre-treatment variables, with i -th row equal to $\mathbf{X}_i$.

Furthermore, throughout this Thesis, we indicate with $p(\cdot)$ and $p(\cdot|\cdot)$ the probability or probability density and conditional probability or conditional probability density,

respectively, depending on the context.

1.2 Rubin Causal Model

The notation and methodology we introduce from now on is rooted in what Holland (1986) named “Rubin Causal Model” (RCM). The RCM is a formal mathematical framework, introduced by Rubin in a series of papers (Rubin, 1974, 1975, 1976, 1977, 1978, 1979, 1980), generally credited for the clarity it brings to questions of causality. The RCM has two fundamental parts and a third optional one.

The first fundamental part of the RCM implies the definition of causal effects by using the concept of potential outcomes. Let $\mathbf{z}$ denote a n -dimensional vector of possible treatment assignments for the whole sample, with i -th element equal to $z_i : \mathbf{z} \in \{0, 1\}^n$. Differently from Z_i , which denotes the treatment level to which unit i is actually assigned, z_i denotes a treatment level to which unit i may, in principle, be assigned. For each unit i , we postulate the existence of two potential outcomes, $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$. These correspond, respectively, to the value of the intermediate variable and the outcome that would be observed if the vector of treatment assignments were $\mathbf{Z} = \mathbf{z}$. In this Chapter, to convey the main ideas, we refer to a setting where the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980) holds for $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$. We formalize this assumption as follows:

Assumption 1 (*SUTVA for $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$*).

No hidden variations of assignment: For each unit i , for all $\mathbf{z}, \mathbf{z}' : \mathbf{z} = \mathbf{z}'$

$$D_i(\mathbf{z}) = D_i(\mathbf{z}'), \quad Y_i(\mathbf{z}) = Y_i(\mathbf{z}').$$

No-interference: For each unit i , if $z_i = z'_i$, then

$$D_i(\mathbf{z}) = D_i(\mathbf{z}'), \quad Y_i(\mathbf{z}) = Y_i(\mathbf{z}').$$

In words, SUTVA for $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$ implies that no different forms or versions of each assignment level exist, and that the value of the intermediate variable and the outcome for a given unit do not vary with the treatment assignment of the other units. Therefore, $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$ can be written as $D_i(z_i)$ and $Y_i(z_i)$. The latter denote, respectively, the value of the intermediate variable and the outcome that would be observed for unit i if it were assigned to treatment $Z_i = z_i$. In the sequel,

we simplify the notation into $D_i(z) \equiv D_i(z_i)$ and $Y_i(z) \equiv Y_i(z_i)$. Also, we refer to the n -dimensional vectors whose i -th elements are, respectively, equal to $D_i(z)$ and $Y_i(z)$, as $\mathbf{D}(z)$ and $\mathbf{Y}(z)$.

According to the RCM, a causal effect of treatment assignment, Z , on the outcome Y is defined as a comparison between potential outcomes under alternative treatment levels, $Y_i(1)$ and $Y_i(0)$, for a common set of units, s , that is, a comparison between the ordered sets $\{Y_i(1) : i \in s\}$ and $\{Y_i(0) : i \in s\}$. For example, the *Average Causal Effect* (ACE) can be properly defined as the difference in the means of $Y_i(1)$ and $Y_i(0)$ for all units or for subgroups of units defined by values of covariates.

The “Fundamental Problem of Causal Inference” (Holland, 1986) is that only one value for each potential outcome can be observed, namely the one corresponding to the treatment actually assigned, while the potential outcome under the alternative treatment is necessarily missing. The latter is often referred to in the literature as *counterfactual* potential outcome. Specifically, $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$ and $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. Note that, since at least half of the potential outcomes are necessarily missing, causal inference under the potential outcomes framework is inherently a missing data problem.

The second fundamental part of the RCM concerns the definition of a model for the assignment mechanism. The assignment mechanism is the stochastic rule for assigning treatments to units and, therefore, is crucial in determining which potential outcomes are observed and which are missing. Specifically, it describes, as a function of the observed covariates and of all potential outcomes under study, the probability of any vector of assignments: $p(\mathbf{Z} | \mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X})$. We consider now three basic restrictions on assignment mechanisms.

Assumption 2 (*Probabilistic assignment*). For each unit i ,

$$0 < p(Z_i | \mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) < 1.$$

Assumption 3 (*Unconfounded assignment*).

$$p(\mathbf{Z} | \mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = p(\mathbf{Z} | \mathbf{X}).$$

Assumption 4 (*Individualistic assignment*). For some function $q(\cdot) \in [0, 1]$,

$$p(\mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = q(D_i(0), D_i(1), Y_i(0), Y_i(1), \mathbf{X}_i) \text{ for } i = 1, \dots, n,$$

and

$$p(\mathbf{Z}|\mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = c \cdot \prod_{i=1}^n q(D_i(0), D_i(1), Y_i(0), Y_i(1), \mathbf{X}_i)^{Z_i} (1 - q(D_i(0), D_i(1), Y_i(0), Y_i(1), \mathbf{X}_i))^{1-Z_i},$$

where c is the constant that ensures that the probabilities sum to unity. An assignment mechanism is probabilistic if the unit-level assignment probabilities are strictly between zero and one; it is unconfounded if it is free of dependence of any potential outcomes; and it is individualistic if treatment assignment of each unit i does not depend on the outcomes and assignments for other units.

A *classical randomized experiment* is a special case of randomized experiment in which the assignment mechanism is probabilistic, unconfounded, individualistic and is a known function of its arguments (Imbens & Rubin, 2016). In classical randomized experiments where pre-treatment covariates do not enter the assignment mechanism, the following holds by design:

Assumption 5 (*Randomization of treatment assignment*)

$$p(\mathbf{Z}|\mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = p(\mathbf{Z}).$$

An example is provided by *completely randomized experiments*, where n_t units are randomly selected to receive the treatment (with the remaining $n - n_t$ assigned to the control), and each unit has the same assignment probability. In this case, the assignment mechanism equals

$$p(\mathbf{Z}|\mathbf{D}(0), \mathbf{D}(1), \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \begin{cases} 1/\binom{n}{n_t} & \text{if } \sum_{i=1}^n Z_i = n_t, \\ 0 & \text{otherwise.} \end{cases} \quad (1.1)$$

Without a model for how treatments are assigned to individuals, formal causal inference, at least using probabilistic statements, is impossible. Yet, we do not need to know the assignment mechanism, but it is mandatory to posit one in order to make any statistical claims about causal effects.

Finally, the third optional part of the RCM is an extension to include model-based inference. In particular, this extension allows for model-based Bayesian posterior predictive inference on causal effects.

1.3 The Principal Stratification framework

PS was proposed by Frangakis and Rubin (2002) as a general framework to evaluate treatment effects while adjusting for post-treatment complications. This approach, deeply rooted in the RCM, has been successfully employed to handle problems of non-compliance (e.g., Forastiere et al., 2016; Frangakis et al., 2002; Hirano et al., 2000; Imbens & Rubin, 1997), surrogate endpoints (e.g., Gilbert & Hudgens, 2008; Zigler & Belin, 2012), direct and indirect causal pathways (e.g., Elliott et al., 2010; Gallop et al., 2009), loss at follow-up or missing outcome data (e.g., Mattei et al., 2014), truncation by death (e.g., Mattei & Mealli, 2007; Rubin, 2006; Zhang et al., 2009), but also combinations of these and other complications (e.g., Frumento et al., 2012; Li et al., 2022). In the setting we consider in this first Chapter, the treatment assignment is a two-valued variable, $Z_i \in \{0, 1\}$. Then, the basic principal stratification P_0 with respect to post-treatment variable D is the partition of the units $i = 1, \dots, n$ such that all units within any set of P_0 have the same vector $(D_i(0), D_i(1))$. The resulting subgroups of units are called (basic) *principal strata*. Formally, let $G_i = (D_i(0), D_i(1))$, then the basic principal strata are given by the possible values of G_i :

$$(d_0, d_1) \equiv \{i : D_i(0) = d_0, D_i(1) = d_1\} \text{ with } d_0, d_1 \in \mathcal{D}.$$

We relabel the possible values of G_i as d_0d_1 . When $\mathcal{D}$ is a finite set, we have $|\mathcal{D}|^{|\mathcal{Z}|}$ possible principal strata. For example, if $\mathcal{D} \equiv \{0, 1\}$, the possible principal strata are $2^2 = 4$, and they can be denoted as $G_i \in \{11, 01, 10, 00\}$. Despite we are referring, for ease of presentation, to the case of two treatment levels only, i.e., $|\mathcal{Z}| = 2$, it is worth noting that the number of possible principal strata grows exponentially as the cardinality of $\mathcal{Z}$ increases. Also, it grows rapidly as the cardinality of $\mathcal{D}$ increases. In the limiting case in which $\mathcal{D} \equiv \mathbf{R}$, the basic principal stratification with respect to D consists of a continuum of principal strata. In such circumstances, the basic principal strata are usually aggregated into meaningful groups. A partition of the units whose sets are unions of sets in the basic principal stratification P_0 is the formal definition of a principal stratification P with respect to D .

Since the value of the ordered pair $(D_i(0), D_i(1))$ is, by definition, not affected by treatment assignment, principal stratum G_i is also not affected by treatment assignment for any principal stratification P . As a result, conditioning on principal strata leads to well-defined causal effects, named *Principal Causal Effects* (PCEs). Formally, a PCE can be defined as a comparison between the ordered sets

$$\{Y_i(0) : G_i = g\} \quad \text{and} \quad \{Y_i(1) : G_i = g\}, \quad (1.2)$$

where g is a principal stratum in the principal stratification P .

The ACE is a *global* ITT effect of treatment assignment, Z , on the outcome Y for the whole sample. PCEs are, instead, *local* ITT effects, meaning that they measure the causal effect of Z on Y for specific subgroups, i.e., the principal strata. These effects are either *associative* or *dissociative* with effects on the post-treatment variable D . Formally, Frangakis and Rubin (2002) defined a dissociative effect as a comparison between the ordered sets in (1.3), and an associative effect as comparison between the ordered sets in (1.4).

$$\{Y_i(0) : D_i(0) = D_i(1)\} \quad \text{and} \quad \{Y_i(1) : D_i(0) = D_i(1)\}. \quad (1.3)$$

$$\{Y_i(0) : D_i(0) \neq D_i(1)\} \quad \text{and} \quad \{Y_i(1) : D_i(0) \neq D_i(1)\}. \quad (1.4)$$

In other words, PCEs for a stratum of the form $G_i = d_0 d_1$ are dissociative if $d_0 = d_1$, and associative if $d_0 \neq d_1$.

1.3.1 Super-population and finite-sample estimands

In this Thesis, we specifically target average PCEs defined as the difference in the means of $Y_i(0)$ and $Y_i(1)$ for subgroups of units defined by the same value of G_i . Formally, we define the super-population conditional average PCE, for units belonging to a principal stratum g and with observed covariate values $\mathbf{x}$, as

$$\vartheta_{sp}^{(g)}(\mathbf{x}) \equiv \mathbb{E}[Y_i(1) - Y_i(0) | G_i = g, \mathbf{X}_i = \mathbf{x}]. \quad (1.5)$$

This causal estimand is useful to investigate treatment effect heterogeneity with respect to observed covariates. Also, we formally define the (marginal) super-population average PCE, for units belonging to a principal stratum g , as

$$\vartheta_{sp}^{(g)} \equiv \mathbb{E}[Y_i(1) - Y_i(0) | G_i = g] = \mathbb{E}[\vartheta_{sp}^{(g)}(\mathbf{X}_i)]. \quad (1.6)$$

Super-population refers to the infinite population from which the sample units were drawn. The expectations in Equation (1.6) are with respect to the distribution generated by random sampling from the super-population. In some of the works presented in this Thesis, we also refer to finite-sample average PCEs. We define the finite-sample average PCE for a principal stratum g as

$$\vartheta_{fs}^{(g)} \equiv \frac{\sum_{i:G_i=g} Y_i(1) - Y_i(0)}{n_g}, \quad (1.7)$$

where n_g is the number of sample units belonging to the principal stratum g .

The ACE can be re-obtained as a weighted average of the PCEs for the principal strata. Formally, the super-population and finite-sample ACEs are, respectively,

$$\vartheta_{ACE,sp} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E} \left[\sum_g \vartheta_{sp}^{(g)} \cdot \pi_i^{(g)} \right] \text{ and} \quad (1.8)$$

$$\vartheta_{ACE,fs} = \frac{\sum_{i=1}^n Y_i(1) - Y_i(0)}{n} = \sum_g \vartheta_{fs}^{(g)} \cdot \frac{n_g}{n}, \quad (1.9)$$

where for the super-population the weights are given by the conditional probability of belonging to the principal stratum g , i.e. $\pi_i^{(g)} \equiv p(G_i = g)$, and for the finite-sample by the sample proportion of units belonging to the principal stratum g , i.e., n_g/n .

Generally, the choice to target a super-population or finite-sample causal estimand is driven by the scientific question rather than statistical considerations. In what follows, we omit to specify the subscripts sp and fs when what is being said applies to both the super-population and the finite-sample versions of the estimands.

1.3.2 Applications

PS has been successfully applied to many scientific problems where intermediate variables are usually at stake. In this Section, we provide an overview of the declinations of the approach in these different contexts.

Non-compliance

In randomized studies characterized by non-compliance, D is typically a binary indicator for treatment receipt; specifically, $D_i = 1$ indicates that unit i received

the assigned treatment and $D_i = 0$ that it did not. Thus, the PS with respect to D partitions the sample into four principal strata based on the joint potential values of treatment receipt under each level of the assignment. These latent subgroups were named by Rubin (1978) as shown in Table 1.1. Fundamentally, each unit i belongs

TABLE 1.1: Principal strata and their labels

$D_i(0)$	$D_i(1)$	<i>Principal Stratum</i>	<i>Stratum Label</i>
1	1	$G = 11$	<i>Always Takers</i>
0	1	$G = 01$	<i>Compliers</i>
1	0	$G = 10$	<i>Defiers</i>
0	0	$G = 00$	<i>Never Takers</i>

to one of the following groups: *Always Takers* (AT), *Compliers* (C), *Defiers* (D) and *Never Takers* (NT). *Always Takers* and *Never Takers* are units that, respectively, would receive, and would not receive the treatment regardless of the assignment. *Compliers* are units that would only receive the treatment if it were assigned to them. Finally, *Defiers* are units that do the opposite of what has been assigned to them: they would not receive the treatment if it were assigned to them, but they would receive it if it were not. The presence of Defiers is often ruled out invoking the monotonicity of the compliance status.

Effects for those whose treatment receipt is affected by the assignment are associative, and thus may actually derive from treatment receipt, though not solely from it. Effects for those whose treatment receipt is unaffected by the assignment are, instead, dissociative, and thus cannot derive from treatment receipt (Section 1.3). Consequently, the PCE of interest in randomized studies with non-compliance is usually the average causal effect for Compliers, $\vartheta^{(01)}$, also named *Complier Average Causal Effect* (CACE), or, synonymously, *Local Average Causal Effect* (LATE). Indeed, unlike the PCEs for NT and AT, the CACE is an associative effect, which therefore can provide information on the effect of the actual treatment (treatment assignment plus treatment receipt). We will discuss the identifying assumptions for this causal estimand in Section 1.4.

Surrogate endpoints

In order to assess the effects of a treatment on the outcome, some studies would require such long follow-up as to be practically unfeasible (e.g., studies whose outcome

is survival time). Nevertheless, if there exists a variable, measurable early in the follow-up, that is known to be linked to the effect of the treatment on survival, then that variable can help to understand the effect of the treatment on the outcome. This variable is called *surrogate*. The use of surrogates reduces trial duration (and thereby costs) and ensures faster patient access to new therapies (see Burzykowski et al., 2005, for a textbook discussion).

Frangakis and Rubin (2002) exploited the PS framework, and in particular, the PCEs, to propose a criterion to define good surrogates. They referred to a surrogate that satisfies this criterion as a *principal surrogate*. The requirement for D to be a principal surrogate is that, for all fixed $d : D_i(0) = D_i(1) = d$, the comparison between the ordered sets in (1.3) results in equality. This implies that $\vartheta^{(00)}$ and $\vartheta^{(11)}$, which are dissociative effects, are zero. This satisfies the property of *causal necessity*, according to which an effect of the treatment on the outcome can be present only when an effect of the treatment on the surrogate has occurred. Gilbert and Hudgens (2008) modified this definition of a principal surrogate to also possess the property of *causal sufficiency*. The latter requires that the comparison between the ordered sets in (1.4) does not result in equality. This means that associative effects must be different from zero. In other words, whenever the treatment does change the surrogate, it must influence the outcome as well.

In practice, it is difficult for the property of causal necessity to be perfectly met, which means that dissociative effects are rarely zero. Yet, large values of associative effects and small values of dissociative effects correspond to good surrogates from a PS perspective.

Direct and indirect causal pathways

The PS framework also proved useful when D can be conceptualized as a mediator (e.g., Mealli & Mattei, 2012; VanderWeele, 2008). In a mediation framework, the total effect (e.g., the ACE) can be decomposed into direct and indirect effects. Direct effects are the effects of Z on Y that do not act through D ; indirect effects are, instead, the effects of Z on Y that pass through D . To gather information on direct effects, dissociative effects are especially useful. If they are zero, it means that there is no evidence of direct effects of treatment assignment after controlling for the mediator. Yet, this does not mean that there are no direct effects whatsoever. Associative effects, in fact, generally blend direct and indirect effects.

Missing outcome data

In many empirical studies, outcome values are missing for some subjects, often due to nonresponse. To provide valid inference, one needs to posit a mechanism leading to these missing data. A possibility is to assume *missing completely at random* (MCAR), which means that the probability of nonresponse is constant across units. Under this assumption, the missing data problem can be solved simply by restricting the sample to respondents only. Another possibility is to assume *missing at random* (MAR), a weaker assumption according to which the probability of nonresponse depends on the observed values (including the observed covariates and the parameters governing the missing data mechanism) but not on the missing values. The assumptions of MCAR and MAR are convenient because they do not require explicitly modeling the missing mechanism (Little & Rubin, 2019). However, sometimes missing values carry on information about the probability of nonresponse. In this case, MCAR and MAR are not valid. The missing mechanism is called *nonignorable*, and requires to be explicitly modeled in the analysis.

PS has been used to cope with this complication (e.g., Mattei et al., 2014). In the presence of nonignorable missing outcome values, the PS approach stratifies the sample into different principal strata based on the joint values of the potential response indicator under each treatment condition. The resulting strata are the following: $G = 11$ comprises those who would respond both under treatment and under control; $G = 01$ comprises those who would respond under treatment but not under control; $G = 10$ comprises those who would not respond under treatment but would respond under control; and $G = 00$ those who would never respond regardless of treatment assignment. The estimand of interest is $\vartheta^{(11)}$, i.e. the causal effect on stratum $G = 11$. In fact, the latter is the only stratum in which $Y(1)$ is observed for units assigned to the treatment and $Y(0)$ is observed for units assigned to the control. In the other three strata, the outcome for respondents in at most one treatment arm is observed.

Truncation by death

In studies where the outcome of interest is truncated by death, D is often the survival status; specifically, $D_i = 1$ indicates that unit i survived until the endpoint and $D_i = 0$ that it did not. In these studies, the PS with respect to D partitions the sample into four principal strata based on the joint potential survival statuses.

These latent subgroups are usually named as shown in Table 1.2. Each unit i

TABLE 1.2: Principal strata and their labels

$D_i(0)$	$D_i(1)$	Principal Stratum	Stratum Label
1	1	$G = 11$	<i>Always Survivors</i>
0	1	$G = 01$	<i>Compliant Survivors</i>
1	0	$G = 10$	<i>Defiant Survivors</i>
0	0	$G = 00$	<i>Never Survivors</i>

belongs to one of the following groups: *Always Survivors*, *Compliant Survivors*, *Defiant Survivors* and *Never Survivors*. *Always Survivors* and *Never Survivors* are units who, respectively, would be alive, and would be dead at the endpoint regardless of the treatment assignment. *Compliant Survivors* are units that would be alive at the endpoint only if assigned to the active treatment. Finally, *Defiant Survivors* are units that would not survive to the endpoint if assigned to the active treatment but would survive if assigned to the control treatment.

This conceptualization of the problem of truncation by death shares some similarities with the way the problem of missing outcomes data is tackled from a PS perspective. First, the fact that in both cases for all members of the strata other than $G = 11$ it is possible to observe the outcome at most in one treatment arm. Yet, the problem is conceptually different. Suppose, for example, that Y is the quality of life one year after the treatment; for patients who die after being treated, this outcome is not only missing but also undefined. Therefore, in the context of studies where the outcome of interest is truncated by death, $v^{(11)}$ is the only well-defined and scientifically meaningful PCE. This estimand was named *Survivor Average Causal Effect* (SACE) by Rubin (2006). The second reason why the issues of loss at follow-up or missing outcome data and truncation by death resemble each other is that identification issues in estimating the PCE for the stratum of the *Always Respondents* ($G = 11$ with D being the response indicator) are analogous to those encountered in estimating the PCE for the stratum of *Always Survivors* ($G = 11$ with D being survival status) (Mattei et al., 2014). We discuss identification issues arising from the latent nature of the principal strata in the next Section.

To conclude, it is worth mentioning that in applications of PS to studies where the outcome of interest is truncated by death, D might also denote survival time, rather than survival status. In that case, aspects regarding identification and estimation

are more complex, but the interest is still in the treatment effect among the sub-population that would have survived in either treatment arm.

1.4 Identifying assumptions

The Unconfoundedness of treatment assignment (Assumption 3) allows the identification of the ACE. By this, we mean that a unique consistent estimate of the ACE can be found from the observed data. Specifically, it can be found by comparing the outcome of treated and control units that have the same value of covariates. In the super-population:

$$\begin{aligned}\vartheta_{ACE,sp}(\mathbf{x}) &= \mathbb{E}[Y_i(1) - Y_i(0) | \mathbf{X}_i = \mathbf{x}] \\ &= \mathbb{E}[Y_i(1) | Z_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i(0) | Z_i = 0, \mathbf{X}_i = \mathbf{x}] \\ &= \mathbb{E}[Y_i | Z_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i | Z_i = 0, \mathbf{X}_i = \mathbf{x}],\end{aligned}$$

where $\vartheta_{ACE,sp} = \mathbb{E}[\vartheta_{ACE,sp}(\mathbf{x})]$. If strata memberships were known, the Unconfoundedness of treatment assignment would also allow the PCEs to be identified. Specifically, their identification can be achieved by comparing the outcome of units that are members of the same stratum and have the same value of the covariates. Unfortunately, we generally do not know which principal stratum the units belong to. The observed subgroups of units defined by the assigned treatment and the value of the intermediate variable, in fact, host mixtures of principal strata. Formally, let $\mathcal{O} = (z, d) \equiv \{i : Z_i = z, D_i = d\}$ for $z, d \in \{0, 1\}$. The observed subgroups of units are given by the possible values of $\mathcal{O}$, which we relabel as zd . In the case of binary D , $\mathcal{O} \in \{11, 01, 10, 00\}$; we report the possible principal strata for each observed subgroup in Table 1.3. From this Table, it is evident that each observed subgroup is

TABLE 1.3: Observed subgroups and underlying principal strata

Z_i	D_i	$\mathcal{O}$	G_i
0	0	00	$\{00, 01\}$
0	1	01	$\{11, 10\}$
1	0	10	$\{00, 10\}$
1	1	11	$\{11, 01\}$

a mixture of multiple latent principal strata. Thus, the introduction of some explicit

assumptions is necessary to restrict the possible patterns of principal strata for the observed subgroups, and thus facilitate the identification and estimation of PCEs.

In recent years, various identifying assumptions have been proposed. Some of them yield point-identify the PCEs, others only partial identification. Partial identification of the PCEs means that the observed data do not allow us to infer their true value, but rather a set of possible values consistent with the observed data and strictly included in the parameter space. This set is called the “identified set”, and is, essentially, a region of identification of the target estimand.

We provide an overview of the possible identifying assumptions that can be made, aggregating them into the following meaningful categories: assumptions restricting the number of principal strata; assumptions on the equality of outcomes; Ranked Average Score/Stochastic Dominance assumptions; assumptions that involve the existence of an instrument for the intermediate variable; assumptions that rely on secondary outcomes; assumptions that rely on pre-treatment auxiliary variables; and, finally, distributional assumptions.

1.4.1 Assumptions restricting the number of principal strata

To restrict the number of principal strata, monotonicity assumptions can be made.

In the case of binary D , *Monotonicity* and *Strong Monotonicity* are often invoked to unmask principal stratum membership of some units. We provide a formal definition of both assumptions below.

Assumption 6 (*Strong Monotonicity*). For each unit i ,

$$D_i(z) = c \text{ for } z \in \{0, 1\}, \text{ where } c \text{ is a constant.}$$

In randomized studies with non-compliance, Strong Monotonicity is usually stated for $z = 0$ and $c = 0$, implying that only two principal strata exist: $G = 00$ and $G = 01$ (see Table 1.1). In fact, it is quite common that access to the treatment is denied to those who have not been assigned to it; in that case, the assumption is automatically fulfilled. Under this assumption, $\mathcal{O} = 10$ is made only of units belonging to the $G = 00$ stratum, and $\mathcal{O} = 11$ is made only of units belonging to the $G = 01$ stratum. This enables the identification of the conditional distribution of the principal strata given the covariates. Specifically, it implies the following

results:

$$\begin{aligned} p(G_i = 00 | \mathbf{X}_i = \mathbf{x}) &= p(D_i(1) = 0 | \mathbf{X}_i = \mathbf{x}) \stackrel{(1)}{=} p(D_i = 0 | Z_i = 1, \mathbf{X}_i = \mathbf{x}), \\ p(G_i = 01 | \mathbf{X}_i = \mathbf{x}) &= 1 - p(G_i = 00 | \mathbf{X}_i = \mathbf{x}), \end{aligned}$$

where the equalities denoted with $\stackrel{(1)}{=}$ follows SUTVA (Assumption 1) and Unconfoundedness of treatment assignment (Assumption 3). Furthermore, Strong Monotonicity for $z = 0$ and $c = 0$ enables the identification of the distributions of $Y(1)$ for $G = 00$ and of $Y(1)$ for $G = 01$ within cells defined by the values of the covariates. In terms of expectations, the following relations apply:

$$\begin{aligned} \mathbb{E}[Y_i(1) | G_i = 00, \mathbf{X}_i = \mathbf{x}] &= \mathbb{E}[Y_i(1) | D_i(1) = 0, \mathbf{X}_i = \mathbf{x}] \\ &\stackrel{(1)}{=} \mathbb{E}[Y_i | Z_i = 1, D_i = 0, \mathbf{X}_i = \mathbf{x}], \\ \mathbb{E}[Y_i(1) | G_i = 01, \mathbf{X}_i = \mathbf{x}] &= \mathbb{E}[Y_i(1) | D_i(1) = 1, \mathbf{X}_i = \mathbf{x}] \\ &\stackrel{(1)}{=} \mathbb{E}[Y_i | Z_i = 1, D_i = 1, \mathbf{X}_i = \mathbf{x}]. \end{aligned}$$

However, we cannot identify the conditional distribution of $Y(0)$ given the principal stratum and covariates without introducing further assumptions. Indeed, we do not know if those assigned to $Z = 0$ belong to the stratum $G = 00$ or $G = 01$. We only know that

$$\begin{aligned} \mathbb{E}[Y_i(0) | \mathbf{X}_i = \mathbf{x}] &= (\mathbb{E}[Y_i(0) | G_i = 00, \mathbf{X}_i = \mathbf{x}] \cdot p(G_i = 00 | \mathbf{X}_i = \mathbf{x}) + \\ &\quad (\mathbb{E}[Y_i(0) | G_i = 01, \mathbf{X}_i = \mathbf{x}] \cdot p(G_i = 01 | \mathbf{X}_i = \mathbf{x})), \end{aligned}$$

which can be identified from $\mathbb{E}[Y_i | Z_i = 0, \mathbf{X}_i = \mathbf{x}]$.

Assumption 7 (*Monotonicity*). For each unit i ,

$$D_i(z) \geq D_i(z') \quad \text{for } z, z' \in \{0, 1\} \text{ with } z \neq z'.$$

Monotonicity essentially allows the effect of the treatment on the intermediate variable to be only in one direction. In the case of randomized studies with non-compliance, Monotonicity is usually stated for $z = 1$ and $z' = 0$, ruling out the presence of Defiers (see Table 1.1); therefore, it is also called *No-Defiers* assumption. Under this assumption, $\mathcal{O} = 01$ is made only of units belonging to the stratum $G = 11$, and $\mathcal{O} = 10$ is made only of units belonging to the stratum $G = 00$. This

enables the identification of the conditional distribution of the principal strata given the covariates. Specifically, it implies the following results:

$$\begin{aligned}
p(G_i = 10 | \mathbf{X}_i = \mathbf{x}) &= 0, \\
p(G_i = 11 | \mathbf{X}_i = \mathbf{x}) &= p(D_i(0) = 1 | \mathbf{X}_i = \mathbf{x}) \stackrel{(1)}{=} p(D_i = 1 | Z_i = 0, \mathbf{X}_i = \mathbf{x}), \\
p(G_i = 00 | \mathbf{X}_i = \mathbf{x}) &= p(D_i(1) = 0 | \mathbf{X}_i = \mathbf{x}) \stackrel{(1)}{=} p(D_i = 0 | Z_i = 1, \mathbf{X}_i = \mathbf{x}), \\
p(G_i = 01 | \mathbf{X}_i = \mathbf{x}) &= 1 - p(G_i = 11 | \mathbf{X}_i = \mathbf{x}) - p(G_i = 00 | \mathbf{X}_i = \mathbf{x}).
\end{aligned}$$

Furthermore, Monotonicity for $z = 1$ and $z' = 0$ enables the identification of the distribution of $Y(0)$ for $G = 11$ and of $Y(1)$ for $G = 00$ within cells defined by the values of the covariates. In terms of expectations, the following relations apply:

$$\begin{aligned}
\mathbb{E}[Y_i(0) | G_i = 11, \mathbf{X}_i = \mathbf{x}] &= \mathbb{E}[Y_i(0) | D_i(0) = 1, \mathbf{X}_i = \mathbf{x}] \\
&\stackrel{(1)}{=} \mathbb{E}[Y_i | Z_i = 0, D_i = 1, \mathbf{X}_i = \mathbf{x}], \\
\mathbb{E}[Y_i(1) | G_i = 00, \mathbf{X}_i = \mathbf{x}] &= \mathbb{E}[Y_i(1) | D_i(1) = 0, \mathbf{X}_i = \mathbf{x}] \\
&\stackrel{(1)}{=} \mathbb{E}[Y_i | Z_i = 1, D_i = 0, \mathbf{X}_i = \mathbf{x}].
\end{aligned}$$

In the case of continuous D , the *Equipercntile Equating* assumption can be invoked (Efron & Feldman, 1991; Jin & Rubin, 2008). Formally,

Assumption 8 (*Equipercntile Equating*). For each unit i ,

$$D_i(1) = m(D_i(0)) = F_{D(1)}^{-1}\{F_{D(0)}[D_i(0)]\},$$

where $F_{D(z)}(\cdot)$ is the cumulative distribution function of $D(z)$ with $z = 0, 1$. Under this assumption, $D_i(1)$ determines $D_i(0)$ based on $m(\cdot)$. Nonetheless, this one-to-one mapping function denies the possibility that two units, with the same value of $D(z)$, may have different values of the counterfactual intermediate variable, i.e., $D(1 - z)$.

As we have shown in the case of Strong Monotonicity, neither Monotonicity nor the Equipercntile Equating assumption, even if coupled with Unconfoundedness of treatment assignment, is sufficient to achieve the point identification of $\mathbb{E}[Y_i(0) | G_i = g, \mathbf{X}_i = \mathbf{x}]$. Therefore, the PCEs cannot be identified without imposing additional assumptions.

1.4.2 Assumptions on the equality of outcomes

When appropriate, assumptions on the equality of outcomes under the different assignment levels can be made for one or more principal strata. Such assumptions are called Exclusion Restrictions (ERs). Formally, the stochastic ER for a principal stratum g can be formulated as:

Assumption 9 (*Stochastic ER for a principal stratum g*)

$$p(Y_i(1)|G_i = g, \mathbf{X}_i = \mathbf{x}) = p(Y_i(0)|G_i = g, \mathbf{X}_i = \mathbf{x}).$$

The assumption states the equality of the conditional distributions of the potential outcomes given the covariates for members of principal stratum g . In randomized studies with non-compliance, the Assumption is often stated for Always Takers ($G = 11$) and Never Takers ($G = 00$). As a result, $\mathbb{E}[Y_i(0)|G_i = g, \mathbf{X}_i = \mathbf{x}] = \mathbb{E}[Y_i(1)|G_i = g, \mathbf{X}_i = \mathbf{x}]$ for $g \in \{11, 00\}$, which in turn imply that the super-population average PCEs for these two principal strata, $\vartheta_{sp}^{(11)}$ and $\vartheta_{sp}^{(00)}$, are zero. Alternatively, a non-stochastic version of Assumption 9 can be stated. We call it weak ER and formalize it as follows:

Assumption 10 (*Weak ER for a principal stratum g*). For all i such that $G_i = g$, $Y_i(1) = Y_i(0)$.

The assumption states the equality of the individual potential outcomes for members of principal stratum g . Therefore, under this assumption, $\vartheta_{fs}^{(g)} = 0$.

By assuming the Unconfoundedness of treatment assignment (Assumption 3), the absence of Defiers (Assumption 7 for $z = 1$ and $z' = 0$), and the ER for non-Compliers, the CACE can be point identified¹. Specifically, the conditional CACE can be point identified from Equation (1.10).

$$\begin{aligned} \vartheta_{sp}^{(11)}(\mathbf{x}) &= \mathbb{E}[Y_i(1) - Y_i(0)|G_i = 01, \mathbf{X}_i = \mathbf{x}] = \frac{\mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]}{p(D_i(1) - D_i(0) = 1|\mathbf{X}_i = \mathbf{x})} = \\ &= \frac{\mathbb{E}[Y_i|Z_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i|Z_i = 0, \mathbf{X}_i = \mathbf{x}]}{p(D_i = 1|Z_i = 1, \mathbf{X}_i = \mathbf{x}) - p(D_i = 1|Z_i = 0, \mathbf{X}_i = \mathbf{x})}. \end{aligned} \quad (1.10)$$

¹Also, the expected values in the numerator of Equation (1.10) must be finite, and $p(D_i = 1|Z_i = 1, \mathbf{X}_i = \mathbf{x}) \neq p(D_i = 1|Z_i = 0, \mathbf{X}_i = \mathbf{x})$ so that the denominator is non-zero (Imbens & Angrist, 1994).

PCEs for non-Compliers, i.e., Always Takers and Never Takers in absence of Defiers, are dissociative (Section 1.3). Thus, an ER for them implies that there are no effects of being assigned to the treatment on the outcome after controlling for treatment receipt. This assumption is often reasonable. Yet, in applications of PS to post-treatment complications other than non-compliance, dissociative effects are not only likely to exist but are often the subject of scientific interest. For example, as we have already mentioned, in studies where the outcome of interest is truncated by death, the SACE, i.e. $\vartheta^{(11)}$, is the causal estimand of interest. In such circumstances, the ER for the principal strata $G = 11$ and $G = 00$ cannot be imposed.

1.4.3 Ranked Average Score/Stochastic Dominance

Zhang and Rubin (2003) derived large-sample bounds for the SACE, without imposing any identifying assumption or parametric model. In addition, they showed that Monotonicity (Assumption 7 for $z = 1$ and $z' = 0$) and a *Ranked Average Score* assumption, whether specified individually or together, can be used to sharpen these bounds. Generally, Ranked Average Score assumptions concern the expected value of a potential outcome across two principal strata, but within either the treatment or the control group. Specifically, the Ranked Average Score assumption that Zhang and Rubin (2003) made is the following:

Assumption 11 (*Ranked average score*)

$$\begin{aligned}\mathbb{E}[Y_i(1)|D_i(0) = 1, D_i(1) = 1] &\geq \mathbb{E}[Y_i(1)|D_i(0) = 1, D_i(1) = 0], \\ \mathbb{E}[Y_i(0)|D_i(0) = 1, D_i(1) = 1] &\geq \mathbb{E}[Y_i(0)|D_i(0) = 0, D_i(1) = 1].\end{aligned}$$

The Assumption, in this formulation, means that when assigned to the treatment, the outcome for the $G = 11$ group is no less than that for the $G = 10$ group, and when assigned to the control, the outcome for the $G = 11$ group is no less than that for the $G = 01$ group.

An alternative to the assumption of Ranked Average Score is that of *Stochastic Dominance*. A random variable A is said to be stochastically dominant over another random variable B , if for any real number t , $p(A \leq t) \leq p(B \leq t)$. Stochastic Dominance assumptions are made on the *distributions* of a potential outcome across two principal strata, within either the treatment or the control group (e.g., Mattei &

Mealli, 2007; Rubin, 2006). When D is binary, a Stochastic Dominance assumption that is sometimes made is the following:

Assumption 12 (*Stochastic dominance*)

$$\begin{aligned} p(Y_i(1)|D_i(0) = 1, D_i(1) = 1) &\geq p(Y_i(1)|D_i(0) = 1, D_i(1) = 0), \\ p(Y_i(0)|D_i(0) = 1, D_i(1) = 1) &\geq p(Y_i(0)|D_i(0) = 0, D_i(1) = 1). \end{aligned}$$

The Assumption, in this formulation, states that when assigned to the treatment, the score for the $G = 11$ group is stochastically dominant over the score for the $G = 10$ group, and when assigned to the control, the score for the $G = 11$ group is stochastically dominant over the score for the $G = 01$ group. This Assumption is stronger than Assumption 11, and leads to the same bounds for the average causal effect for the $G = 11$ group (Zhang & Rubin, 2003). Imai (2008) exploited it to derive an alternative proof that the bounds proposed by Zhang and Rubin (2003) are sharp (i.e., the shortest bounds possible without additional assumptions or auxiliary information available), and to derive sharp bounds on the quantile treatment effect for the $G = 11$ group.

Unfortunately, the large-sample bounds in Zhang and Rubin (2003) are typically too wide for practical use. Several ways to sharpen them had been proposed in the literature; generally speaking, the proposed strategies involve the use of additional information, for instance, on pre-treatment auxiliary variables, secondary outcomes, or instruments for the intermediate variable. We present some of these strategies in the following Sections.

1.4.4 Assumptions that involve the existence of an instrument for the intermediate variable

The literature includes works showing how an instrument for the intermediate variable, when it exists, can be used to sharpen the bounds of PCEs. Among these, two major contributions are those of Mattei and Mealli (2011) and Mattei et al. (2014), which we briefly describe below.

Mattei and Mealli (2011) developed an augmented design that proves useful to sharpen the bounds of dissociative effects, which they named *principal strata direct effects* (PSDEs) to explicitly link them to the mediation framework. This new

design involves the augmentation of the standard randomized experiment with an *instrument* for the mediating variable. Let us denote the treatment with $Z_i \in \{0, 1\}$, the mediator with $D_i \in \{0, 1\}$, the instrument for the mediator with $W_i \in \{0, 1\}$, and the outcome with $Y_i \in \mathcal{Y}$. W_i is an instrument in that it is assumed to influence the mediator (and consequently the outcome) but not directly the outcome. Specifically, W_i is viewed as an encouragement to exhibit a certain value of D_i . Because W_i essentially constitutes “an additional treatment”, the authors defined potential outcomes of the form $D_i(Z_i = z, W_i = w)$ and $Y_i(Z_i = z, W_i = w)$ with $z, w \in \{0, 1\}$, and built a principal stratification based on the joint potential values of $D_i(z, w)$. For 8 out of the 16 resulting principal strata, PCEs are dissociative and are therefore the target of the investigation.

The assumptions their approach relies on are: (1) SUTVA; (2) randomization of both Z and W ; (3) stochastic ER with respect to W ; (4) non-zero average causal effect of W on D ; (5) monotonicity of D with respect to W ; (6) and monotonicity of D with respect to Z . (3) rules out the presence of direct effects of W on Y ; together with (4), it is what denotes W as an instrument. (5) and (6) instead limit, respectively, the direction of the effect of W on D and that of Z on D . Under these assumptions, the large-sample bounds that can be derived for PSDEs are tighter than those derived in standard randomized experiments.

Mattei et al. (2014) proposed to adopt a similar approach when the outcome is missing on a subset of units and dependence of non-response on the outcome cannot be ruled out (even after conditioning on observed covariates). In this case, they proposed to augment the design with an instrument for non-response, and define principal strata based on the non-response behavior in all possible combinations of treatment and instrument. They focus on global ITT effects and local ITT effects for the stratum of those who would always respond. The results show that the use of an instrument for non-response greatly helps the identification and estimation of these causal estimands when deriving bounds and also when performing a Bayesian model-based PS analysis; in fact, in the former case, the bounds are tighter, and in the latter case, the estimates are more precise. Also, a stronger instrument is shown to provide greater benefits.

1.4.5 Assumptions that rely on secondary outcomes

Mealli and Pacini (2013) showed that existing large-sample bounds on PCEs for subpopulations defined by compliance behavior (i.e., Always Takers, Never Takers and Compliers under the No-Defiers assumption) can be tightened by exploiting the joint distribution of the outcome of primary interest and a secondary outcome, for which the ER more plausibly holds. Specifically, they showed that the Stochastic ER (Assumption 9) on a secondary outcome for AT and NT reduces the admissible values of the joint distribution of the two outcomes conditional on compliance status and, in turn, the admissible values of the marginal distribution of the primary outcome conditional on compliance status. This eventually reduces the widths of large-sample bounds on PCEs. Also, the authors showed that the stronger the association of the secondary outcome with the primary outcome and the compliance status, the more the restriction helps tighten the bounds.

Finally, Mealli and Pacini (2013) showed, as a special case, that a covariate not affected by treatment assignment may serve the same purpose. In fact, when using a covariate, the independence between that covariate and treatment assignment (within strata) plays the same role as the ER on a secondary outcome. We explore this aspect in the next Section.

1.4.6 Assumptions that rely on pre-treatment auxiliary variables

Several studies leveraged pre-treatment auxiliary variables to improve inference on PCEs. Jiang and Ding (2021) have recently established a general theory for the identification and estimation of the PCEs with a pre-treatment auxiliary variable satisfying some conditional independence assumptions. They divide the latter into two categories. The first category is that of *Principal Ignorability* (PI) assumptions, which assume that the outcome is independent of the principal strata given the auxiliary variable. The other category assumes conditional independence between the potential outcomes and the auxiliary variable within principal strata. We refer to this conditional independence as *Auxiliary Independence*.

In what follows, we denote the value of the auxiliary variable for each unit i as $A_i \in \mathcal{A}$.

Principal Ignorability assumptions

PI assumptions are homogeneity assumptions that can be used to extrapolate the missing information across principal strata. In some circumstances, they allow to reach the point identification of the PCEs. Two different formulations of PI are *Strong PI* and *Weak PI* (Feller et al., 2017; Jo & Stuart, 2009).

Assumption 13 (*Strong PI*). For $z \in \{0, 1\}$, $Y_i(z) \perp\!\!\!\perp (D_i(0), D_i(1)) | A_i$.

For clarity of exposition, we also express this assumption in terms of mean independence²:

$$\mathbb{E}[Y_i(z)|G_i = g, A_i = a] = \mathbb{E}[Y_i(z)|A_i = a] \stackrel{(1)}{=} \mathbb{E}[Y_i|Z_i = z, A_i = a].$$

Strong PI states that, given the auxiliary variable, the potential outcomes do not depend on the principal stratum; in other words, the partition into principal strata is not relevant to the treatment effects. To make the assumption more plausible, researchers often include all the observed pre-treatment covariates in A , so that $A \equiv \mathbf{X}$.

For ease of presentation, let us now refer to the case D is binary and Monotonicity (Assumption 7 for $z = 1$ and $z' = 0$) is also assumed, ruling out the presence of units belonging to the $G = 10$ stratum. Strong PI results in $\vartheta^{(11)} = \vartheta^{(00)} = \vartheta^{(01)} = \vartheta_{ACE}$; thus, it is clearly a strong conditional independence assumption. It has, however, testable implications since under Monotonicity the principal stratum membership of some units is revealed. In the case we are referring to, for example, we directly observe the average outcome for members of the $G = 11$ stratum assigned to control, and for members of the $G = 00$ stratum assigned to treatment. Therefore, we could exploit this and assume Weak PI instead of Strong PI. Specifically, Weak PI could be formulated as follows:

Assumption 14 (*Weak PI*). For $z \in \{0, 1\}$, $Y_i(z) \perp\!\!\!\perp D_i(1 - z) | A_i, D_i(z) = z$.

²The formal identification results only require mean independence (rather than full stochastic independence).

Written in terms of mean independence we have that, for $d \in \{0, 1\}$,

$$\begin{aligned} \mathbb{E}[Y_i(1)|D_i(0) = d, D_i(1) = 1, A_i = a] &= \mathbb{E}[Y_i(1)|D_i(1) = 1, A_i = a] \\ &\stackrel{(1)}{=} \mathbb{E}[Y_i|Z_i = 1, D_i = 1, A_i = a], \end{aligned} \quad (1.11)$$

$$\begin{aligned} \mathbb{E}[Y_i(0)|D_i(0) = 0, D_i(1) = d, A_i = a] &= \mathbb{E}[Y_i(0)|D_i(0) = 0, A_i = a] \\ &\stackrel{(1)}{=} \mathbb{E}[Y_i|Z_i = 0, D_i = 0, A_i = a]. \end{aligned} \quad (1.12)$$

Equations (1.11) and (1.12) constitute Weak PI. Equation (1.11) states that, given the auxiliary variable, the same outcome under the treatment is expected for a member of $G = 11$ and a member of $G = 01$. Equation (1.12) states that, given the auxiliary variable, the same outcome under the control is expected for a member of $G = 00$ and a member of $G = 01$.

Finally, it is worth mentioning that, to reach the point identification of the PCEs, PI assumptions can also be mixed and matched with ERs, as shown in Feller et al. (2017).

Auxiliary independence assumptions

Assumption 15 (*Auxiliary Independence*). For $z \in \{0, 1\}$, $Y_i(z) \perp\!\!\!\perp A_i | G_i$.

Auxiliary Independence states that, for units in the same principal stratum, the distribution of potential outcomes is identical regardless of the value of the auxiliary variable. Pre-treatment covariates can be included to make this assumption more plausible, i.e., $Y_i(z) \perp\!\!\!\perp A_i | (G_i, \mathbf{X}_i)$; we omit them for simplicity.

Jiang et al. (2016) provided formal identification results for PCEs with an auxiliary variable in the case of binary D ; Jiang and Ding (2021) provided them also in the case of continuous D . For the sake of brevity, we do not report these results but defer to the respective papers.

1.4.7 Distributional assumptions

Distributional information, when available, can be exploited to achieve point identification of the PCEs. In fact, if distributional assumptions, such as the assumption of normal outcome distributions within principal strata, are made, then a standard mixture model analysis can be performed. This allows deriving unique maximum likelihood point estimates of the causal parameters of interest by using the

Expectation-Maximization (EM) algorithm (e.g., Frumento et al., 2012; Zhang et al., 2009).

To explain the basic ideas, we now refer to the case of studies where the outcome of interest is truncated by death, and thus the interest is in the SACE. Define P_{zd} as the sample proportion of units belonging to $\mathcal{O} = zd$, and the probability of belonging to a principal stratum g as $\pi^{(g)} \equiv p(G_i = g)$. When the treatment is randomized, in large samples we have

$$\begin{aligned}\pi^{(11)} + \pi^{(01)} &= \frac{\sum_i Z_i D_i}{\sum_i Z_i} \equiv P_{11}, \\ \pi^{(10)} + \pi^{(00)} &= \frac{\sum_i Z_i (1 - D_i)}{\sum_i Z_i} = 1 - P_{11}, \\ \pi^{(11)} + \pi^{(10)} &= \frac{\sum_i (1 - Z_i) D_i}{\sum_i (1 - Z_i)} \equiv P_{01}, \\ \pi^{(01)} + \pi^{(00)} &= \frac{\sum_i (1 - Z_i) (1 - D_i)}{\sum_i (1 - Z_i)} = 1 - P_{01}.\end{aligned}\tag{1.13}$$

If the distributions for the potential outcomes given the principal strata are Normal, then $Y_i(1)$ for the subgroup $\mathcal{O} = 11$ follows a mixture of two Normal distributions: one for $Y_i(1)$ given $G_i = 11$ and the other for $Y_i(1)$ given $G_i = 01$. Using standard normal mixture model analysis, we can estimate the proportions of the two components, and therefore $\pi^{(11)}/\pi^{(01)}$. Similarly, $Y_i(0)$ for the subgroup $\mathcal{O} = 01$ follows a mixture of two Normal distributions: one for $Y_i(0)$ given $G_i = 11$ and the other for $Y_i(0)$ given $G_i = 10$; so, we can estimate $\pi^{(11)}/\pi^{(10)}$. By combining the estimates of $\pi^{(11)}/\pi^{(01)}$ and $\pi^{(11)}/\pi^{(10)}$ with the equations in (1.13), we can uniquely estimate the π 's. Then, we can exploit them and, except in the special case when $\pi^{(11)} = \pi^{(01)}$ or $\pi^{(11)} = \pi^{(10)}$, we can uniquely estimate $\mathbb{E}[Y_i(1)|G_i = 11]$ from the component with mixture proportion $\pi^{(11)}/(\pi^{(11)} + \pi^{(01)})$ in the distribution of $Y_i(1)$ for the $\mathcal{O} = 11$ group, and $\mathbb{E}[Y_i(0)|G_i = 11]$ from the component with mixture proportion $\pi^{(11)}/(\pi^{(11)} + \pi^{(10)})$ in the distribution of $Y_i(0)$ for the $\mathcal{O} = 01$ group.

Alternatively, it is possible to exploit parametric models within a Bayesian framework. In general, Bayesian statistics is an inferential approach where any available background knowledge about the parameters in a model is updated in light of the observed data. The background knowledge is expressed through a prior distribution while the information in the data is given by the likelihood function; the two are combined through the Bayes' theorem to obtain the posterior distribution (for

a review, van de Schoot et al., 2021). In a Bayesian framework, we have that, as the sample size goes to infinity, the marginal posterior distribution of the partially identified causal effects will converge to a nondegenerate distribution with support equal to the identified set (Gustafson, 2010).

A Bayesian model-based PS, such as originally proposed by Imbens and Rubin (1997), requires the specification of two sets of models: one for the principal strata membership given the covariates and one for the potential outcomes conditional on the covariates and the principal stratum. This usually leads to weakly identified causal effects; “identified” in the sense of having a proper posterior distribution, but “weak” in the sense that this distribution will have regions of flatness around its maximum – in frequentist terms, there are no unique maximum likelihood estimates (Hirano et al., 2000). However, from a Bayesian perspective, even when the likelihood is rather flat, if the prior is proper, the posterior will also be proper. Also, the shape of the posterior distribution usually helps to distinguish some values in the identification region that are more plausible than others.

To sharpen inference, one could provide prior information or make one or more identifying assumptions presented in this Section. In the latter case, the Bayesian approach is especially useful because it allows for relaxing all of the assumptions that might be made, thus providing an intuitive and straightforward way to perform sensitivity analyses. Also, if information on multivariate outcomes is available, Mattei et al. (2013) showed that it can be used to further sharpen inferences for weakly identified PS models.

This is the approach we follow in the works presented in this Thesis. Section 1.5 will provide the details.

1.4.8 The role of the pre-treatment covariates

In randomized studies, pre-treatment covariates are typically recorded. In a completely randomized experiment, although these variables do not enter the treatment assignment mechanism, they can be used to improve the efficiency of estimation for many reasons. First, they improve the prediction of the principal stratum membership and missing potential outcomes, which in turn improves the precision of the causal estimates (e.g., Gilbert & Hudgens, 2008; Hirano et al., 2000). Second, covariates allow the specification of the identifying assumptions as conditional, rather

than marginal, and this generally makes the assumptions more plausible. Also, additional restrictions can be imposed on coefficients of the covariates based on plausible hypotheses within or among groups with the same values of the covariates. For example, the same coefficient across strata can be imposed for one or more covariates (Frangakis, 2006). Finally, the use of covariates allows more plausible generalization to a population that differs in some ways from the one under study.

1.5 Bayesian model-based Principal Stratification

Let us now describe the Bayesian model-based PS approach in detail. Throughout this thesis, we analyze experimental or quasi-experimental studies in which the intermediate variable D is binary, and some form of monotonicity holds by design or can be reasonably assumed. Therefore, the following description is based on this assumption; in particular, we refer to the case where it is the presence of the $G = 10$ stratum that is ruled out.

Let $\boldsymbol{\psi}$ be the parameter vector, including the parameters of the distribution of the principal stratum membership given the covariates and of the conditional distributions of the potential outcome given the covariates and the principal stratum. Assume a prior distribution $p(\boldsymbol{\psi})$. Let $\pi_i(\mathbf{x}, \boldsymbol{\psi})^{(g)} \equiv p(G_i = g \mid \mathbf{X}_i = \mathbf{x}, \boldsymbol{\psi})$ for $g \in \{11, 01, 00\}$ (the G -model) and $f_i(\mathbf{x}, \boldsymbol{\psi})^{(z,g)} \equiv p(Y_i(z) \mid G_i = g, \mathbf{X}_i = \mathbf{x}, \boldsymbol{\psi})$ for $z \in \{0, 1\}$ and $g \in \{11, 01, 00\}$ (the Y -model). More compactly, let $\pi_i^{(g)} \equiv \pi_i(\mathbf{x}, \boldsymbol{\psi})^{(g)}$ and $f_i^{(z,g)} \equiv f_i(\mathbf{x}, \boldsymbol{\psi})^{(z,g)}$. Given the latent nature of the principal strata, the observed data likelihood, namely $\mathcal{L}_{obs}(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$, results in the following finite mixture model likelihood:

$$\begin{aligned} \mathcal{L}_{obs}(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) = & \prod_{i: Z_i=0, D_i=1} \left[\pi_i^{(11)} \cdot f_i^{(0,11)} \right] \times \prod_{i: Z_i=1, D_i=0} \left[\pi_i^{(00)} \cdot f_i^{(0,00)} \right] \\ & \times \prod_{i: Z_i=0, D_i=0} \left[\pi_i^{(01)} \cdot f_i^{(0,01)} + \pi_i^{(00)} \cdot f_i^{(0,00)} \right] \\ & \times \prod_{i: Z_i=1, D_i=1} \left[\pi_i^{(01)} \cdot f_i^{(1,01)} + \pi_i^{(11)} \cdot f_i^{(1,11)} \right]. \quad (1.14) \end{aligned}$$

Let us exploit the superscript *miss*, so that $\mathbf{D}^{miss}$ and $\mathbf{Y}^{miss}$ are the n -dimensional

vectors of the counterfactual potential outcomes, whose i -th elements are respectively equal to $D_i^{miss} \equiv D_i(1 - Z_i)$ and $Y_i^{miss} \equiv Y_i(1 - Z_i)$. The derivation of the likelihood function in Equation (1.14) requires to integrate out $\mathbf{D}^{miss}$ and $\mathbf{Y}^{miss}$ from $p(\mathbf{Y}^{miss}, \mathbf{Y}, \mathbf{D}^{miss}, \mathbf{D} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\psi})$:

$$\begin{aligned} \mathcal{L}_{obs}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) &\equiv p(\mathbf{Y}, \mathbf{D} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\psi}) \\ &= \int \int p(\mathbf{Y}^{miss}, \mathbf{Y}, \mathbf{D}^{miss}, \mathbf{D} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\psi}) d\mathbf{Y}^{miss} d\mathbf{D}^{miss} \end{aligned} \quad (1.15)$$

$p(\mathbf{Y}^{miss}, \mathbf{Y}, \mathbf{D}^{miss}, \mathbf{D} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\psi})$ can be easily obtained in three steps. First, given that we specify models for $p(\mathbf{G} | \mathbf{X}, \boldsymbol{\psi})$, $p(\mathbf{Y}(0) | \mathbf{G}, \mathbf{X}, \boldsymbol{\psi})$ and $p(\mathbf{Y}(1) | \mathbf{G}, \mathbf{X}, \boldsymbol{\psi})$, we can infer the joint distribution $p(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1) | \mathbf{X}, \boldsymbol{\psi})$. Second, given Unconfoundedness of treatment assignment (Assumption 3), we have that

$$p(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1) | \mathbf{X}, \boldsymbol{\psi}) = p(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1) | \mathbf{X}, \mathbf{Z}, \boldsymbol{\psi})$$

Finally, under SUTVA (Assumption 1), the distribution of the observed potential outcomes can be expressed in terms of the distribution of the observed values:

$$p(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{D}(0), \mathbf{D}(1) | \mathbf{X}, \mathbf{Z}, \boldsymbol{\psi}) \equiv p(\mathbf{Y}^{miss}, \mathbf{Y}, \mathbf{D}^{miss}, \mathbf{D} | \mathbf{X}, \mathbf{Z}, \boldsymbol{\psi})$$

Unfortunately, the integration over $\mathbf{D}^{miss}$ in Equation (1.15) is not as straightforward. As a result, the posterior distribution of $\boldsymbol{\psi}$, namely,

$$p(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) \propto p(\boldsymbol{\psi}) \times \mathcal{L}_{obs}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$$

is analytically intractable.

A solution consists in targeting, instead of $p(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$, the joint posterior distribution of the parameters and the latent variable $\mathbf{D}^{miss}$, namely $p(\boldsymbol{\psi}, \mathbf{D}^{miss} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$, or alternatively $p(\boldsymbol{\psi}, \mathbf{G} | \mathbf{Z}, \mathbf{Y}, \mathbf{X})$. To obtain $p(\boldsymbol{\psi}, \mathbf{D}^{miss} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$, we implement a two-stage Gibbs-sampling. The first stage of this algorithm consists of a Data Augmentation (DA) step, in which $\mathbf{D}^{miss}$ is sampled from its posterior predictive distribution; this informs us about the principal strata memberships. Then, the second stage consists of assessing the posterior distribution of parameters based on the complete-data likelihood, that is, the likelihood conditional on the imputed strata memberships. The latter, which we denote with $\mathcal{L}_{comp}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{G}, \mathbf{Y}, \mathbf{X})$, can

be factorized into six components, each one depending on a priori independent sub-vector of the full parameter vector:

$$\begin{aligned} \mathcal{L}_{comp}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{G}, \mathbf{Y}, \mathbf{X}) = & \prod_{i: Z_i=0, G_i=11} \left[\pi_i^{(11)} \cdot f_i^{(0,11)} \right] \times \prod_{i: Z_i=1, G_i=11} \left[\pi_i^{(11)} \cdot f_i^{(1,11)} \right] \times \\ & \prod_{i: Z_i=0, G_i=00} \left[\pi_i^{(00)} \cdot f_i^{(0,00)} \right] \times \prod_{i: Z_i=1, G_i=00} \left[\pi_i^{(00)} \cdot f_i^{(1,00)} \right] \times \\ & \prod_{i: Z_i=0, G_i=01} \left[\pi_i^{(01)} \cdot f_i^{(0,01)} \right] \times \prod_{i: Z_i=1, G_i=01} \left[\pi_i^{(01)} \cdot f_i^{(1,01)} \right]. \end{aligned}$$

By exploiting $\mathcal{L}_{comp}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{G}, \mathbf{Y}, \mathbf{X})$, it is possible to obtain $p(\boldsymbol{\psi}, \mathbf{D}^{miss} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$, given that:

$$p(\boldsymbol{\psi}, \mathbf{D}^{miss} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) \propto p(\boldsymbol{\psi}) \times \mathcal{L}_{comp}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{G}, \mathbf{Y}, \mathbf{X}) \quad (1.16)$$

In other words, the complete-data likelihood allows the full conditional distributions $p(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}^{miss}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$ and $p(\mathbf{D}^{miss} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\psi})$ to be analytically tractable. By iteratively sampling from these two distributions, it is possible to obtain the joint posterior distribution of $(\boldsymbol{\psi}, \mathbf{D}^{miss})$, which in turn provides the marginal posterior distribution of the parameters $p(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$.

Then, a final step within the MCMC is needed to compute super-population and finite-sample PCEs. At each iteration k of the algorithm ($k = 1, \dots, K$), we use the conditional distribution of the potential outcomes given the covariates and the principal stratum memberships imputed at that iteration $\mathbf{G}^k$, evaluated at the parameter values $\boldsymbol{\psi}^k$. We denote the mean of this distribution as $m_y(\mathbf{x}, \boldsymbol{\psi}^k)^{(z,g)} \equiv \mathbb{E}[Y_i(z) | G_i^k = g, \mathbf{X}_i = \mathbf{x}, \boldsymbol{\psi}^k]$ for $z \in \{0, 1\}$ and $g \in \{11, 01, 00\}$. For the super-population PCEs, we also need the conditional distribution of the principal strata given the covariates, evaluated at the parameter values $\boldsymbol{\psi}^k$. We denote it as $\pi_i(\mathbf{x}, \boldsymbol{\psi}^k)^{(g)} \equiv p(G_i^k = g | \mathbf{X}_i = \mathbf{x}, \boldsymbol{\psi}^k)$ for $g \in \{11, 01, 00\}$. An estimate of the super-population PCEs at iteration k is thus obtained by averaging the model-based outcome probabilities over the empirical distribution of the covariates:

$$\widehat{v}_{Bayes_sp}^{(g),k} = \frac{\sum_{i=1}^n \pi_i(\mathbf{x}, \boldsymbol{\psi}^k)^{(g)} \times \left(m_y(\mathbf{x}, \boldsymbol{\psi}^k)^{(1,g)} - m_y(\mathbf{x}, \boldsymbol{\psi}^k)^{(0,g)} \right)}{\sum_{i=1}^n \pi_i(\mathbf{x}, \boldsymbol{\psi}^k)^{(g)}} \quad (1.17)$$

for $g \in \{11, 01, 00\}$. An estimate of the finite-sample PCEs at iteration k , instead, can be obtained by imputing the value of the counterfactual potential outcome, $Y_i^{miss,k}$, for every subject from its posterior predictive distribution. This can be accomplished as follows:

$$Y_i^{miss,k} \sim p(Y_i^{miss} \mid G_i^k = g, \mathbf{X}_i, \boldsymbol{\psi}^k). \quad (1.18)$$

Then, the observed and missing information is used to derive the posterior distribution of the finite-sample PCEs. Specifically,

$$\hat{\vartheta}_{Bayes_fs}^{(g),k} = \frac{\sum_{i:G_i^k=g} (2Z_i - 1)(Y_i - Y_i^{miss,k})}{\sum_{i=1}^n \mathbb{1}(G_i^k = g)}. \quad (1.19)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, taking on value one if its argument is true and zero otherwise. Thus, the denominator of Equation (1.19) is the total number of subjects imputed as members of the principal stratum g at iteration k . After deriving the posterior distributions of the super-population and finite-sample PCEs, a point estimate and an interval estimate can be obtained by computing, respectively, the average over the K iterations and the credible intervals.

1.6 Sensitivity analyses

When adopting a fully parametric approach, sensitivity analyses aimed at detecting any model misspecification must be performed. In a Bayesian framework, the goodness-of-fit of the models can be naturally evaluated by performing posterior predictive checks (Rubin, 1984). This practice is well-established in Bayesian model-based PS analyses (e.g., Barnard et al., 2003; Forastiere, Lattarulo, et al., 2021; Mattei et al., 2013). A posterior predictive check involves a comparison between the observed data and replications of them generated from their posterior predictive distribution. Intuitively, if the models are specified appropriately, the latter should generate data similar to the observed ones. In addition to making a visual inspection, we can formally test the null hypothesis that a chosen observed statistic comes from the model we actually specified. A formal posterior predictive check involves first the choice of a discrepancy measure, which is a function of the observed data and the parameter vector $\boldsymbol{\psi}$, and second the computation of a Bayesian p -value. A widely used Bayesian p -value is the so-called *posterior predictive p -value* (PPPV)

(Gelman et al., 1996; Meng, 1994), which is, in this context, the probability over the posterior predictive distribution of the principal strata and the parameter vector ψ that a given discrepancy measure in a new study data, drawn with the same ψ as in the observed study, would be as or more extreme than its realized value in the observed study. An estimate of the PPPV can be derived by first obtaining the frequency distribution of the discrepancy measure in K replications of the data, i.e., the frequency distribution of Δ^{rep} , and then calculating the percentage of draws in which Δ^{rep} exceeds the realized value of the discrepancy measure in the observed study, i.e., Δ^{obs} . Formally, the PPPV of a generic discrepancy measure is calculated as $\frac{1}{K} \sum_{k=1}^K \mathbb{1}\{\Delta^{rep} \geq \Delta^{obs}\}$. Looking at the p -value thus estimated, we assess whether the model can preserve features of the data relevant to the calculation of the discrepancy measure. If its value is extreme (close to 0 or 1), it means that the prior distribution and the likelihood cannot adequately replicate the discrepancy measure to which the p -value refers.

Let ψ^k ($k = 1, \dots, K$) denote the draw of the parameters vector from its posterior distribution at iteration k . We simulate $K = 3000$ replicated data sets as follows. First, we generate the principal stratum membership of every subject in the sample from the hypothesized principal strata model given ψ^k , i.e., $G_i^{rep,k}$. Thus, it is possible to simulate the outcome $Y_i^{rep,k}$ from its posterior predictive distribution, keeping the encouragement Z_i fixed at its observed value. We repeat the whole procedure K times and thus we obtain K replicated data sets, $(G_i^{rep,k}, Y_i^{rep,k}, Z_i)$. Once we have these replicated data sets, we can use them to compute all the discrepancy measures of interest.

Let us indicate with $\mathcal{D}_{z,g}^{study} = \{i : Z_i = z \text{ and } G_i^{study} = g\}$ the group of units belonging to $G_i^{study} = g$ assigned to $Z_i = z$ in the study data, where $study = obs$ denotes the observed data and $study = rep$ a replication of them obtained from their posterior predictive distribution as described above. Let $|\mathcal{D}_{z,g}^{study}|$ indicate the cardinality of $\mathcal{D}_{z,g}^{study}$, i.e., the number of units in this group. Finally, let $\bar{Y}_{z,g}^{study}$ and $s_{z,g}^{2,study}$ denote the mean and the variance of the outcome variable Y^{study} for units in the $\mathcal{D}_{z,g}^{study}$ group. In the works presented in this Thesis, we adopt the following discrepancy measures:

$$SI_{z,z',g}^{study} = |\bar{Y}_{z,g}^{study} - \bar{Y}_{z',g}^{study}|, \quad NO_{z,z',g}^{study} = \sqrt{\frac{s_{z,g}^{2,study}}{|\mathcal{D}_{z,g}^{study}|} + \frac{s_{z',g}^{2,study}}{|\mathcal{D}_{z',g}^{study}|}}, \quad \text{and}$$

$$SN_{z,z',g}^{study} = \frac{SI_{z,z',g}^{study}}{NO_{z,z',g}^{study}} \quad \text{for } z, z' \in \{0, 1\}, \text{ with } z > z', z' \neq z, \text{ and for all } g.$$

The measures of signal, $SI_{z,z',g}^{study}$, noise, $NO_{z,z',g}^{study}$, and signal to noise, $SN_{z,z',g}^{study}$, were proposed by Barnard et al. (2003) and adopted by Mattei et al. (2013) in a Bayesian model-based PS analysis. They aim at assessing whether the model can preserve the corresponding features. We conclude that the model fits the data well if the estimated PPPVs for these discrepancy measures are all far from zero and one.

Chapter 2

The Oyam Health Expenditures Study

2.1 Research aim and objectives

In Low- and Middle-Income Countries (LMICs), direct payments often dominate health care financing and represent a serious obstacle to the achievement of Universal Health Coverage (UHC). Out-of-pocket expenditures tend to exacerbate the severity of poverty mainly among the poorest and rural populations who lack adequate financial protection (Kruk et al., [2009](#); Leive & Xu, [2008](#); Wagstaff et al., [2011](#)). Interventions of Community Health Financing (CHF) involve the introduction of a prepayment financing mechanism aimed, on the one hand, at enhancing access to health services and, on the other, at protecting individuals from financial catastrophe due to health-related costs. Informal and rural sectors represent the main target of these interventions, where the community is actively involved in the design and management of the financing mechanism (Jütting, [2004](#)).

Several studies analyze the impact of CHF schemes on health expenditures, providing evidence that this model often presents the limitation of not including the most disadvantaged and vulnerable because they lack the financial capacity to pay the initial participation fee. Although the financial contribution provided by the scheme is thus not extended to these vulnerable individuals, the comprehension of principles of prepayment for health financing could make them rethink their health-seeking behaviors and savings practices even if they decide not to enroll. This means that there could be either beneficial or detrimental effects of CHF schemes even on those who opt out: investigating this possibility is paramount to advance the debate on effective policies for UHC in LMICs.

The case study is a pilot intervention of CHF implemented in the rural district of Oyam (Uganda). Throughout the thesis, we refer to this study as the “Oyam Health Expenditures Study” (OHES). In the OHES, households were offered to enroll in a CHF scheme upon being assigned to sensitization sessions. In this study, the participation in the CHF scheme is the treatment of interest, while the sensitization and offer to enroll are an encouragement to take the treatment. The fact that enrollment was on a voluntary basis gave rise to noncompliance, in the sense that there were households who, after being encouraged, accepted the offer to join the scheme, as well as households who declined despite the encouragement.

Nannini et al. (2021) have already conducted an impact evaluation of this pilot intervention using the IV method to address the issue of non-compliance. This method allows, under a set of assumptions, to identify causal effects of the intervention for the subset of households whose enrollment is affected by the instrument (i.e., the initial sensitization and the offer to enroll in the scheme). These causal effects are attributed solely to the actual participation in the scheme for households that would join the scheme if encouraged to, namely those that, in a PS framework, are called *Compliers*. However, as we mentioned above, investigating the presence of effects that are not channeled by the actual participation in the scheme but instead derive from the sensitization sessions would provide important insights for policy-making in the field of public health. It is here, therefore, that we make our contribution. To address the issue of non-compliance, instead of using IV methods, we exploit the PS framework. As already explained in Section 1.3, the approach accounts for the presence of heterogeneous causal effects, which may be associative or dissociative depending on the specific principal stratum to which they refer. In the OHES, associative effects are the effects for those whose enrollment is affected by the encouragement and thus may derive from (though not only from) the participation in the scheme; dissociative effects are the effects for those whose enrollment is unaffected by the encouragement and thus cannot derive from participating in the scheme. Thus, the present work aims to advance the analysis by identifying the distinctive characteristics of these subgroups, and by acknowledging the possibility that both associative and dissociative causal effects are present, meaning that there could be effects of attending sensitization sessions and receiving the offer to join the CHF scheme whether or not that offer is accepted.

2.2 The intervention of Community Health Financing

The intervention of CHF was run in Oyam by the international non-governmental organization “Doctors with Africa CUAMM”. The financial protection scheme has been implemented in two subcounties (Ngai and Myene), randomly selected out of a total of twelve in the Oyam district. Hereafter, we refer to these two subcounties as the intervention area.

In the intervention area, 359 community groups are present. A community group is a savings group or a mutual-help association whose purpose is to help people facing financial hardship. These groups include nearly 75–80% of the population (Biggeri et al., 2018). The understanding and acceptance of the proposed prepayment financing mechanism can be facilitated by educating these existing informal groups to use the scheme (Chemin, 2018; Mladovsky et al., 2014; Sommerfeld et al., 2002). Therefore, the intervention targeted 42 community groups¹, randomly selected out of 359, including a total of 2137 households.

Targeted households first attended activities of sensitization aimed at raising awareness on the impoverishing effects of illness and on the importance of timely preparedness for unexpected health expenses; then, they received the offer to join the scheme and decided whether to enroll or not on a voluntary basis. Although the intervention is explained in detail elsewhere (Nannini et al., 2021), some important points are here reported. Those households who accepted to enroll in the scheme contribute a fixed amount (5000 UGX per member per year) to create and maintain a community fund. In case of health needs, each enrolled member can borrow money from that fund and, after receiving the necessary medical service (inpatient or outpatient care), has four months to repay with zero interest. The maximum amount that can be borrowed each time is 150000 UGX, corresponding to the 85th percentile of the distribution of households’ health expenditure observed in a previous study (Biggeri et al., 2018). At the end of the first year of intervention, households can decide whether to renew their membership for the following year.

¹42 community groups were chosen in order to reach 10304 individuals, accounting for nearly 20% of the population in the intervention area.

2.2.1 Study design

Data on the pilot intervention were collected through a panel households survey conducted in the intervention area and in a third subcounty of the district (Iceme). The latter was included because interviewing households resident in a subcounty where the intervention was not implemented limits the possibility of interference effects. Hereafter, we refer to this third subcounty as the control area. This area was chosen because it is similar to the two subcounties in the intervention area in terms of socioeconomic profile and distance from the main road and health facilities. To determine the sample of households that were to be interviewed, a two-stage sampling design was applied.

In the first stage, 63 villages were selected, including 62 in the intervention area and one in the control area. The 62 villages in the intervention area were chosen because they are homogeneous in terms of population density and distance from health facilities, and the community groups targeted by the intervention resided there². Then, a village with the same demographic and geographical characteristics was selected in the control area.

In the second stage, among residents of these 63 villages, 282 households were randomly chosen to be survey participants, and were interviewed at the baseline (shortly before the roll-out of the intervention), in January 2019, and during the second round of the longitudinal survey, in January 2020. These households are distributed in the encouragement groups as shown in Table 2.1.

TABLE 2.1: Number of households in the sample by encouragement group and area

	<i>No Unenrolled</i>	<i>Sensitization and offer</i>		<i>Overall</i>
		<i>Yes Enrolled</i>	<i>Unenrolled</i>	
Number of households in the sample from the intervention area	26	135	34	195
Number of households in the sample from the control area	87	0	0	87
Number of households in the sample	113	135	34	282

²The vast majority of residents in the 62 selected villages in the intervention area were targeted by the intervention.

The sample of 282 households includes: 243 households that are part of at least one community group and were therefore eligible for the intervention³, and 39 households who are not part of any community group and were therefore non-eligible. Given that households that are part of at least one community group are likely to differ from households that are not part of any community group with respect to observable but also unobservable characteristics, we do not include in the analyses the latter subgroup. Thus, our final sample is composed of 243 households, who are distributed in the encouragement groups as shown in Table 2.2. All the causal contrasts that will be the subject of our investigation are referred to this finite sample.

TABLE 2.2: Number of households in the sample that are part of at least one community group by encouragement group and area

	<i>No</i> <i>Unenrolled</i>	<i>Sensitization and offer</i>		<i>Overall</i>
		<i>Yes</i> <i>Enrolled</i>	<i>Unenrolled</i>	
Number of households in the sample from the intervention area	24	135	34	193
Number of households in the sample from the control area	50	0	0	50
Number of households in the sample	74	135	34	243

2.2.2 Data

The survey questionnaire includes questions on a wide range of household and individual characteristics (e.g., demographic and socioeconomic factors, occurrence of negative shocks, and health-related events and perceptions). Retrospective questions consider a recall period of either one month or one year, depending on the issue. For example, questions about the household's expenses are referred to the previous month to prevent significant recall bias, whereas questions on whether the household had taken out a loan are referred to the previous year.

In Table 2.3, we present some summary statistics for the 243 households in the final sample (at the baseline survey). In particular, for each covariate, the table shows (i) the sample average values by encouragement group, which we denote by $\bar{X}_0$ for those unencouraged and by $\bar{X}_1$ for those encouraged; (ii) the sample variance

³As explained in Section 2.2, the intervention was assigned to community groups.

by encouragement group, that we denote by s_0^2 and s_1^2 ; and (iii) the normalized difference in average covariate values, i.e., $\Delta_X = (\bar{X}_1 - \bar{X}_0)/\sqrt{(s_1^2 + s_0^2)/2}$.

TABLE 2.3: Summary statistics for the 243 households in the final sample (at the baseline survey)

	<i>Sensitization and offer</i>				Δ_X
	<i>No (74 households)</i>		<i>Yes (169 households)</i>		
	$\overline{X}_0$	s_0^2	$\overline{X}_1$	s_1^2	
<i>Referred to the previous month</i>					
Share of health costs over the total monthly expenditures	0.19	0.04	0.18	0.05	-0.03
Savings (UGX)	80513	123300527212	126455	327440332840	0.10
Total monthly expenditures (UGX)	607589	704534492484	358877	362729197467	-0.34
Total monthly expenditures pro capite (UGX)	108222	20691332680	69757	14713488538	-0.29
No food to eat in the household (1/0)	0.35	0.23	0.23	0.18	-0.27
<i>Referred to the previous year</i>					
Loan taken (1/0)	0.49	0.25	0.49	0.25	-0.00
Serious illness occurred (1/0)	0.39	0.24	0.42	0.25	0.06
Number of shocks ^a	3.70	2.76	3.27	2.91	-0.26
<i>Referred to the present</i>					
Female household head (1/0)	0.19	0.16	0.11	0.10	-0.21
Single, widow or divorced household head (1/0)	0.22	0.17	0.15	0.13	-0.18
Age of the household head	44.49	206.03	41.85	208.73	-0.18
Literacy level of the household head (1-3) ^b	2.64	0.56	2.75	0.42	0.16
Household size	5.54	4.33	5.65	5.17	0.05
Children aged 5 or less (1/0)	0.57	0.25	0.66	0.23	0.18
Subsistence farming as main source of income (1/0)	0.72	0.21	0.89	0.10	0.46
Would pay 10000 UGX/year per member for health insurance	0.26	0.19	0.76	0.18	1.15
Satisfaction for health status of household members (1-4) ^c	2.01	0.64	2.22	0.72	0.25
Concerns regarding health (1/0)	0.49	0.25	0.47	0.25	-0.03

^aShocks included: serious illness or accident of household member/s, death of household member/s, drought, irregular rains, floods, unusually high level of crop pests & disease, unusually high costs of agricultural inputs, unusually low prices for agricultural output, theft of money/valuables/non-agricultural assets, theft of agricultural assets/output (crop or livestock), conflict/violence, fire.

^bLevels of literacy of the household head: 1 if unable to read and write, 2 if able to read only, 3 able to read and write.

^cLevels of satisfaction for the health status of household members: 1 if not at all satisfied, 2 if only a little satisfied, 3 if rather satisfied, 4 if completely satisfied.

From this Table, we can observe that, as it is often the case in small-scale field experiments, there are considerable imbalances in the covariate distributions. Notably, while the share of health costs out of total monthly expenditures is roughly the same, on average, for the two encouragement groups, the total monthly expenditures for households in the control group are significantly higher. These households tend also to be less satisfied with the health status of their members, but nevertheless they are much less willing to pay 10000 UGX/year per member to have a health insurance (5000 UGX is the amount required to enroll into the CHF scheme). They also experienced a considerably higher number of shocks in the previous year. Moreover, they do not have food to eat in the household more frequently, probably due to a lack of economic resources. Putting all the pieces together, households that are not encouraged to join the CHF scheme seem to be slightly more disadvantaged.

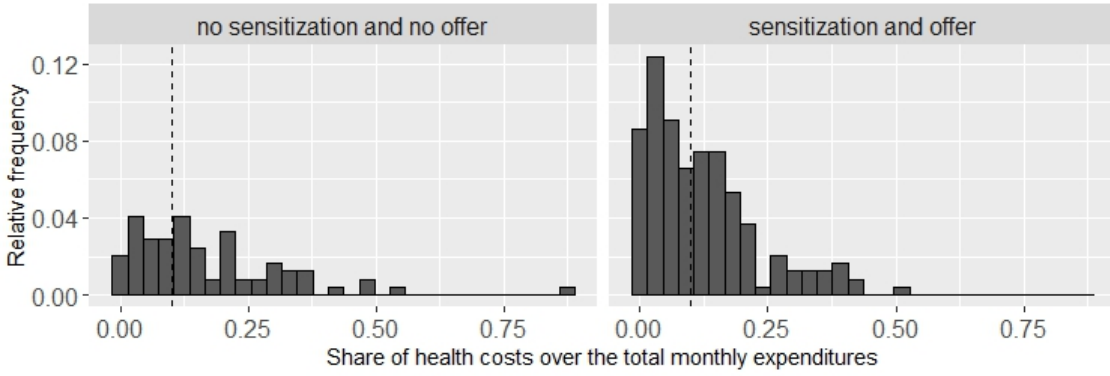
2.2.3 The outcome

As extensively outlined in Section 2.1, the main purpose of introducing interest-free loans in the context of rural Uganda was to increase the utilization of health services by the households but without the financial hardship that this would have normally caused them. In addition, this utilization of health services is expected to be more timely given that the economic barrier that delays or prevents the seeking of treatment has been lowered through the implementation of the CHF scheme. This is especially important because the timely seeking of medical consultation or treatment prevents diseases from worsening and thus improves prognosis and reduces the occurrence of catastrophic health expenditures. However, we would like to point out that, although it is desirable that, in general, healthcare costs fall, it is not necessarily a success if a household spends nothing on medical products/services. Indeed, in that case, it may be that the household did not actually need them, but it may also be that they could not afford them at all. For these reasons, we decided to analyze the share of health costs over the total monthly expenditures, and to define three causal estimands as contrasts of different features of the distribution of this outcome under two treatment conditions. With the first estimand, we aim at measuring causal effects on the share of health costs over the total monthly expenditures; with the second, causal effects on the incidence of catastrophic health expenditures; and, with the third, causal effects on zero health expenditures. The money borrowed from the common fund does not figure as part of these monthly

medical expenses, but the money being paid back with zero interest to the common fund does. Also, health expenditures are considered catastrophic in case they amount to at least 10% of households' monthly expenditures (Wagstaff et al., 2018).

The histograms of the empirical distribution of the share of health costs over the total monthly expenditures by assignment group are shown in Figure 2.1. From this

FIGURE 2.1: Histogram of the empirical distribution of the outcome by encouragement group. The dashed line represents the cut-off beyond which health expenditures are considered catastrophic



plot, it is evident that the distribution of the outcome for households that received sensitization, regardless of whether they enrolled in CHF scheme or not, is more shrunk towards zero, which means that the share of health expenditures tends to be lower for those who attended the sensitization sessions.

2.3 Methodology

2.3.1 Basic notation and potential outcomes

Our sample consists of $n = 243$ households, indexed by $i = 1, \dots, n$. For each household i , let us denote the observed encouragement assignment, i.e., the assignment to the sensitization and offer to enroll, by $Z_i \in \mathcal{Z} \equiv \{0, 1\}$. $Z_i = 1$ indicates that the household was assigned to the encouragement, $Z_i = 0$ that it was not. Let us also denote the treatment by $D_i \in \mathcal{D} \equiv \{0, 1\}$, where $D_i = 1$ indicates that the household enrolled in the CHF scheme, and $D_i = 0$ that it did not. As for the outcome, we indicate with $Y_i \in \mathcal{Y} \equiv [0, 1)$ the share of health costs out of the total monthly expenditures. Finally, for each household i , we observe a vector of K pre-treatment covariates, $\mathbf{X}_i$.

We denote with $\mathbf{Z}$, $\mathbf{D}$ and $\mathbf{Y}$ the n -dimensional vectors of, respectively, the observed encouragement, treatment and outcome for the whole sample, with i -th elements equal to Z_i , D_i and Y_i . Also, we denote with $\mathbf{X}$ the $n \times K$ matrix with i -th row equal to $\mathbf{X}_i$.

Let us now introduce the potential outcomes for the relevant post-encouragement variables. Let $\mathbf{z}$ be a n -dimensional vector of encouragement assignments, with i -th element equal to $z_i : \mathbf{z} \in \{0, 1\}^n$. Then, $D_i(\mathbf{z})$ is the potential enrollment in the CHF scheme given the encouragement vector $\mathbf{z}$, and $Y_i(\mathbf{z})$ is the share of health expenditures out of the total monthly expenditures given the same encouragement vector.

We assume that SUTVA (Assumption 1) holds for the potential outcomes $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$. In words, the assumption means that no different forms or versions of the encouragement exist and that the treatment and the outcome of a given household do not vary with the encouragement of the other households. Thus, under SUTVA, the potential outcomes can be written as $D_i(\mathbf{z}) = D_i(z_i) \equiv D_i(z)$ and $Y_i(\mathbf{z}) = Y_i(z_i) \equiv Y_i(z)$. Also, the following relations hold: $D_i = D_i(Z_i)$ and $Y_i = Y_i(Z_i)$.

2.3.2 Principal strata and definition of Principal Causal Effects

To deal with the issue of non-compliance, we use the PS framework. Accordingly, we classify households into (basic) principal strata defined by their compliance behavior. Because encouragement assignment is a two-valued variable, $Z_i \in \{0, 1\}$, the basic PS with respect to D_i classifies households into basic principal strata based on the vector $(D_i(0), D_i(1))$, i.e., based on their potential enrollment in the CHF program under each encouragement level. Note that, since the possibility of enrolling in the program was given only to households who were assigned to it, non-compliance is *one-sided*: for each household $i = 1, \dots, n$, $D_i(0) = 0$, while $D_i(1) \in \{0, 1\}$. Formally, let $G_i = (D_i(0), D_i(1))$, then the basic principal strata are given by the possible values of G_i :

$$(0, d_1) \equiv \{i : D_i(0) = 0, D_i(1) = d_1\} \text{ with } d_1 \in \{0, 1\}.$$

We relabel the possible values of G_i as $0d_1$, so $G_i \in \{01, 00\}$. Thus, one-sided non-compliance limits the total number of possible principal strata to two, namely Compliers and Never Takers (Table 2.4). The former are those who, if encouraged,

TABLE 2.4: Principal strata and their labels

$D_i(0)$	$D_i(1)$	<i>Principal Stratum</i>	<i>Stratum Label</i>
0	1	$G = 01$	<i>Compliers</i>
0	0	$G = 00$	<i>Never Takers</i>

would decide to enroll. The latter are those who would not enroll in the program even if encouraged.

Our impact evaluation targets three different finite-sample PCEs, namely, that on the share of health costs over the total monthly expenditures, that on the incidence of catastrophic health expenditures, and that on the incidence of zero health expenditures. We name the first *Health Expenditures Effect* (HEE), the second *Catastrophic Health Expenditures Effect* (CHEE), and the third *Zero Health Expenditures Effect* (ZHEE). We formally define them as

$$HEE^{(g)} = \frac{\sum_{i:G_i=g} Y_i(1) - Y_i(0)}{n_g}, \quad (2.1)$$

$$CHEE^{(g)} = \frac{\sum_{i:G_i=g} [\mathbb{1}(Y_i(1) \geq 0.10) - \mathbb{1}(Y_i(0) \geq 0.10)]}{n_g}, \quad (2.2)$$

$$ZHEE^{(g)} = \frac{\sum_{i:G_i=g} [\mathbb{1}(Y_i(1) = 0) - \mathbb{1}(Y_i(0) = 0)]}{n_g}, \quad (2.3)$$

$HEE^{(g)}$ is the average difference between the health costs over the total monthly expenditures in the two assignment conditions for households belonging to the stratum g . $CHEE^{(g)}$ is given by an analogous comparison between the proportions of catastrophic health expenditures, and $ZHEE^{(g)}$ between the proportions of zero health expenditures.

2.3.3 Identification issues

By definition, those who attended the sensitization sessions and then joined the program are necessarily Compliers, because a Never Taker would never join the program regardless of the assignment. Conversely, those who attended the sensitization

sessions and then refused to join the program are necessarily NT. What is unknown is the principal stratum membership of those households that did not attend the sensitization sessions; in fact, we cannot know whether they would have joined the program if they had received sensitization and been offered to enroll. In Table 2.5 we summarize the possible principal strata for each observed subgroup defined by the values of Z_i and D_i .

TABLE 2.5: Some descriptive statistics for the observed subgroups of households and underlying principal strata

Z_i	D_i	No.	% $Z_i = z$	$Mean(Y_i)$	$p(Y_i > 0.10)$	G_i
0	0	74	100%	0.17	0.62	{00,01}
1	0	34	20.1%	0.11	0.44	00
1	1	135	79.9%	0.12	0.48	01

The latent nature of the principal strata makes it necessary to introduce some assumptions to identify and estimate PCEs. Since non-compliance is one-sided, Strong Monotonicity (Assumption 6 for $z = 0$ and $c = 0$) holds by design. Therefore, to point identify the PCEs for Compliers, the following assumptions are sufficient: unconfoundedness of encouragement assignment and ER for NT.

The unconfoundedness of the encouragement assignment assumption (Assumption 3) would hold by design in a (block) randomized experiment. In this study, it is plausible because the control area has been chosen because it was similar to the intervention area in terms of characteristics that may reasonably affect the outcome variables (more details were given in Section 2.2.1). As a result, we can reasonably rely on the validity of Assumption 3, and the conditioning on a rich set of covariates makes it more plausible. This set, in fact, contains many pieces of information about household and household head characteristics, the health status of household members, as well as proxies for the financial status of the household (we will provide the full list of covariates included in $\mathbf{X}_i$ in the Section 2.4.1).

We also assume the ER for NT. Specifically, we impose the Stochastic ER for NT (Assumption 9 for $g = 00$) by assuming prior equality of some models' parameters (details will be provided in Section 2.4.2). In the OHES, ERs amount to assuming that attending the sensitization sessions and receiving the offer to participate in the program do not have any direct influence on the outcome. However, depending on

the principal stratum for which these assumptions are stated, there could be effects that pass through the actual participation. For Never Takers, causal effects are dissociative, meaning that they could only derive from the encouragement; thus, their presence is ruled out when the ER for NT is assumed. In contrast, causal (associative) effects can exist for Compliers and can be attributable to an effect of participating in the scheme as well as to a direct effect of the encouragement.

Often ERs are controversial, and their plausibility needs to be assessed at a case-by-case level. In the application considered here, on one hand, receiving sensitization does not provide the households with additional tools to improve health financing and outcomes of financial protection; on the other hand, however, participating in sensitization sessions may lead to a change in health-seeking behavior after awareness-raising. In view of this, the ER for NT might not hold. For this reason, we also obtain the results without assuming it, thus allowing for the possibility that dissociative effects may exist.

2.3.4 Instrumental Variables approach

Applied economists typically analyze data from experimental studies with non-compliance adopting an IV approach (for a review, see Angrist & Pischke, 2009). The latter was developed in the 1920s to draw causal conclusions in contexts where the treatment cannot be deemed as randomly assigned, even conditional on covariates. If the effects are heterogeneous, Imbens and Angrist (1994) provided a set of assumptions that allow identifying the causal effect for Compliers. The first is the fact that the so-called instrumental variable (here, the encouragement assignment) is either randomized or at least as good as if it was randomized. In our application, the encouragement is assumed to be unconfounded (Assumption 3), therefore this first requirement would be fulfilled. The second is the fact that the instrumental variable influences the treatment but does not affect the outcome directly. This is clearly an ER, which also applies to Compliers, not only to NT. Therefore, in addition to stating that there are no effects for NT, this amounts to attributing the effect for Compliers to the participation in the CHF scheme only, and not to the sensitization received. It is worth noting, however, that this assumption only concerns the interpretation of PCEs for Compliers, while their identification or estimation does not change compared to when the ER is stated for NT only (e.g., Mealli & Mattei, 2012; Mealli & Rubin, 2002). The third assumption on which the IV methods rely is

the Monotonicity of compliance (Assumption 7 for $z = 1$ and $z' = 0$). This assumption holds by design because it is implied by Strong Monotonicity (Assumption 6 for $z = 0$ and $c = 0$). Finally, the last condition needed for using an IV method is that the instrument must be substantially correlated with the intermediate variable. This last condition can be verified from the data: here, it is satisfied because of 169 households who attended sensitization sessions, 135 participated in the CHF program (79.88%).

When all the underlying assumptions are satisfied, the IV estimand is equivalent to the average causal effect for Compliers (Angrist et al., 1996; Imbens & Angrist, 1994). However, when the ER does not hold, the IV estimand is equal to the average causal effect for Compliers plus a bias term. The size of the latter depends on the proportion of NT: the more NT, the greater the bias is (Angrist et al., 1996). In any case, even in the presence of a few NT, it is important to know whether there are causal effects of the sensitization for them. This would, in fact, have paramount practical implications for policy-making. First, delivering sensitization sessions is much easier and less costly than implementing CHF schemes; hence, hypothetical future awareness-raising campaigns would easily reach many more people. Moreover, even the poorest and most vulnerable, who would not have the financial capacity to pay the initial participation fee required to join a CHF scheme, could benefit from these campaigns; therefore, it would be important to think carefully about how to design them to maximize their effectiveness.

2.4 Analyses

To deal with the partial identification of the PCEs, we conduct a Bayesian model-based PS analysis. To this end, we specify a model for the principal stratum membership given the covariates, and a model for the potential outcomes conditional on the covariates and the principal stratum.

2.4.1 Parametric models

We model the principal stratum membership using a logit model:

$$\log \left(\frac{p(G_i = 01 \mid \mathbf{W}_i)}{1 - p(G_i = 01 \mid \mathbf{W}_i)} \right) = \alpha + \mathbf{W}_i' \boldsymbol{\alpha}_W$$

where α is the intercept and $\boldsymbol{\alpha}_W = (\alpha_{W_1}, \dots, \alpha_{W_T})'$ is a vector whose elements are the remaining regression parameters. $\mathbf{W}_i$ is, instead, a vector of T pre-treatment covariates, partially different from $\mathbf{X}_i$, entering the principal stratum model. Specifically, $\mathbf{W}_i$ includes: the share of health expenditures out of the total monthly expenditures and the total monthly expenditures prior to the intervention; whether the household head is female, her/his age and her/his literacy level; whether there are household members aged 5 or less; whether the main source of income is subsistence farming; whether a serious illness occurred in the household in the previous year and whether the respondent is scared about the cost of health expenditures; the number of shocks that the household experienced in the previous year; and, finally, its willingness to pay 10000 UGX/year per member for health insurance. A different model could have been chosen, for instance, a probit model. However, since the model is specified as a flexible function of the covariates, we expect to obtain the same results.

As for the outcome, a common model choice for outcomes that are continuous but restricted to the interval $(0,1)$, such as rates or proportions, is the Beta distribution (e.g., Ferrari & Cribari-Neto, 2004; Kieschnick & McCullough, 2003). The latter has the advantage of being very flexible since its density can have quite different shapes depending on the values of the two parameters that index the distribution. However, it includes neither zero nor one in its support. In our empirical application, some households reported that they had no health expenditure at all in the previous month; thus, $Y_i \in [0,1) \forall i$. Therefore, the beta distribution is not suitable to model our outcome. In similar cases, in which rates or proportions data include a non-negligible number of zeros and/or ones, a more appropriate choice is the zero and/or one inflated beta distribution (Ospina & Ferrari, 2010, 2012). Indeed, the latter is a mixed continuous–discrete distribution defined on the unit interval with probability mass at zero and/or one. For this reason, for the potential outcomes, we assume a Zero-Inflated Beta distribution composed of two parts: a binary part, responsible for the zeros, and a continuous part, that is responsible for the outcome values that are greater than zero⁴.

⁴Despite Ospina and Ferrari (2010, 2012) called this model Zero-Inflated Beta distribution, the latter is more similar to a Hurdle model than a Zero-Inflated model. In fact, since the support of the Beta distribution does not include zero, the zeros are not *inflated*, but rather are given only by the binary part of the model. The two mixture components (the binary and the continuous parts) are conditionally independent, which means that the likelihood factorizes into the product of the likelihood functions for each process, which is usually computationally convenient. However, in this study, since we do not know the principal stratum membership for the households in the

$$p(Y_i(z) = y \mid G_i = g, \mathbf{X}_i) = \begin{cases} \omega_i^{(z,g)} & \text{if } y = 0 \\ (1 - \omega_i^{(z,g)}) \times \text{Beta}(y; \mu_i^{(z,g)}, \phi_i^{(z,g)}) & \text{if } y \in (0, 1) \end{cases}$$

Here, $\omega_i^{(z,g)}$ is the probability of a given household i of having zero health expenditures under the assignment condition z and given that it belongs to the principal stratum g . $\text{Beta}(y; \mu_i^{(z,g)}, \phi_i^{(z,g)})$ is instead the density of a beta distribution with parameters $\mu_i^{(z,g)} \in (0, 1)$ and $\phi_i^{(z,g)} > 0$, that has the form

$$\frac{\Gamma(\phi_i^{(z,g)})}{\Gamma(\mu_i^{(z,g)} \phi_i^{(z,g)}) \cdot \Gamma((1 - \mu_i^{(z,g)}) \phi_i^{(z,g)})} \times y^{\mu_i^{(z,g)} \phi_i^{(z,g)} - 1} (1 - y)^{(1 - \mu_i^{(z,g)}) \phi_i^{(z,g)} - 1}$$

for $z \in \{0, 1\}$ and $g \in \{01, 00\}$. $\mu_i^{(z,g)}$ is the mean of the latter distribution, and $\phi_i^{(z,g)}$ is responsible for its precision: for fixed $\mu_i^{(z,g)}$, the larger the value of $\phi_i^{(z,g)}$, the smaller the variance. This parameterization of the beta distribution is not the usual one, but it is convenient here mainly because it allows modeling its mean, i.e., the mean of the continuous part of the Zero-Inflated Beta distribution, and its precision parameter as a function of the covariates directly (Ospina & Ferrari, 2012). In order to have a more parsimonious model, we assume that $\phi_i^{(z,g)}$ depends only on the assignment condition and the principal stratum membership, and not on covariates: thus, $\phi_i^{(z,g)} = \phi^{(z,g)} \forall i$. On the contrary, we assume that $\omega_i^{(z,g)}$ and $\mu_i^{(z,g)}$ are defined as:

$$h_1(\omega_i^{(z,g)}) = l_1(\mathbf{X}_i, \boldsymbol{\rho}^{(z,g)}) \quad (2.4)$$

$$h_2(\mu_i^{(z,g)}) = l_2(\mathbf{X}_i, \boldsymbol{\beta}^{(z,g)}) \quad (2.5)$$

where $\boldsymbol{\rho}^{(z,g)} = (\rho^{(z,g)}, \boldsymbol{\rho}_X^{(z,g)})' = (\rho^{(z,g)}, \rho_{X_1}^{(z,g)}, \dots, \rho_{X_K}^{(z,g)})'$ and $\boldsymbol{\beta}^{(z,g)} = (\beta^{(z,g)}, \boldsymbol{\beta}_X^{(z,g)})' = (\beta^{(z,g)}, \beta_{X_1}^{(z,g)}, \dots, \beta_{X_K}^{(z,g)})'$ are vectors of unknown regression parameters. $h_1(\cdot)$ and $h_2(\cdot)$ are the link functions. $l_1(\cdot, \cdot)$ and $l_2(\cdot, \cdot)$ are, instead, linear or nonlinear twice continuously differentiable functions in the second argument. Here, we assume that the covariates are linear in the model parameters and specify a logit link function,

control group, the likelihood still contains a mixture of subjects belonging to different strata (see Section 2.4.3).

resulting in:

$$\begin{aligned}\log\left(\frac{\omega_i^{(z,g)}}{1 - \omega_i^{(z,g)}}\right) &= \rho^{(z,g)} + \mathbf{X}_i' \boldsymbol{\rho}_X^{(z,g)} \\ \log\left(\frac{\mu_i^{(z,g)}}{1 - \mu_i^{(z,g)}}\right) &= \beta^{(z,g)} + \mathbf{X}_i' \boldsymbol{\beta}_X^{(z,g)}.\end{aligned}$$

We assume, for the sake of parsimony, that the effect of the covariates on the outcome is the same for all the households: $\boldsymbol{\rho}_X^{(z,g)} \equiv \boldsymbol{\rho}_X$ and $\boldsymbol{\beta}_X^{(z,g)} \equiv \boldsymbol{\beta}_X$. To explain the meaning of this assumption, suppose that $\mathbf{X}_i$ contains information about the household size. The assumption would imply that the effect of the size of the household on the share of health costs is the same irrespective of the encouragement assignment and compliance status of that household. Its validity could be tested by relaxing it by letting $\boldsymbol{\rho}_X$ and $\boldsymbol{\beta}_X$ free to vary depending on (z, g) . If the corresponding posterior distributions are similar across (z, g) , it means that the assumption could be safely assumed. However, doing so would substantially increase the number of parameters in the model. Indeed, even in the case where the predictors in $\mathbf{X}_i$ are only three, instead of $3 \times 2 = 6$ coefficients ($\boldsymbol{\rho}_X = (\rho_{X_1}, \rho_{X_2}, \rho_{X_3})'$ and $\boldsymbol{\beta}_X = (\beta_{X_1}, \beta_{X_2}, \beta_{X_3})'$), we would have $3 \times 2 \times 4 = 24$ ($\boldsymbol{\rho}_X^{(z,g)} = (\rho_{X_1}^{(z,g)}, \rho_{X_2}^{(z,g)}, \rho_{X_3}^{(z,g)})'$ and $\boldsymbol{\beta}_X^{(z,g)} = (\beta_{X_1}^{(z,g)}, \beta_{X_2}^{(z,g)}, \beta_{X_3}^{(z,g)})'$ for $z \in \{0, 1\}$ and $g \in \{01, 00\}$). Unfortunately, the actual sample size is not appropriate to perform such a sensitivity analysis.

Note that the intercepts $\rho^{(z,g)}$ and $\beta^{(z,g)}$ are allowed to vary depending on the pair (z, g) . Since $\boldsymbol{\rho}_X$ is constant over (z, g) , $\rho^{(z,g)}$ is what allows $\omega^{(z,g)}$ to vary depending on (z, g) . In other words, $\rho^{(z,g)}$ is what allows the mass at zero to be different depending on (z, g) . The same applies to $\beta^{(z,g)}$, which is, instead, responsible for the mean of the beta distribution, which can potentially differ depending on (z, g) .

Finally, it would be possible to include different covariates in the binary and in the continuous part of the zero-inflated beta model. Yet, we choose the same set of covariates, $\mathbf{X}_i$, for both parts, since we believe that the predictors of zero medical expenses and the amount of non-zero medical expenses are plausibly the same. Specifically, $\mathbf{X}_i$ includes: the share of health expenditures out of the total monthly expenditures prior to the intervention; the total monthly expenditures and the amount of monthly savings; whether the household head is female; whether in the household there are members aged 5 or less and, more generally, the size of the household; whether the main source of income is subsistence farming; whether it

happened the previous month that there was no food to eat in the house because of the lack of resources to get the food; and, finally, the level of satisfaction with the health status of household members. This set of covariates is partially different from the one that enters the principal strata model, $\mathbf{W}_i$. In fact, we have observed variables that we reasonably believe are predictors of compliance and also of the outcome, as well as variables that are related to only one of them. For example, the willingness to pay for a health financing scheme is considered in the literature as a specific proxy of the ability to pay for it (Adebayo et al., 2015), thus affecting the enrolment in CHF.

We assume that the parameters are a priori independent and specify weakly informative prior distributions for them. Specifically, for the parameters of the principal strata model, α and α_W , we assume Normal prior distributions with mean 0 and standard deviation equal to 5. Also for the parameters in the potential outcomes model we assume Normal prior distributions centered at zero. However, for the intercepts of the binary and the continuous part of the model, $\rho^{(z,g)}$ and $\beta^{(z,g)}$, we set the prior standard deviation to 2, while for the regression coefficients of the model, ρ_X and β_X we set it to 5. This diversification of the standard deviations is motivated by stability reasons. Finally, $\phi^{(z,g)} \sim \text{Gamma}(0.25, 0.25)$, where the hyperparameters are chosen to make the prior variance for the potential outcomes similar to the observed outcome variance.

2.4.2 Assuming and relaxing the Exclusion Restriction for Never Takers

As anticipated in Section 2.3.3, we first obtain the results under the ER for NT. We assume it by imposing the following constraints on the models' parameters:

$$\rho^{(1,00)} = \rho^{(0,00)}, \quad (2.6)$$

$$\beta^{(1,00)} = \beta^{(0,00)}, \quad (2.7)$$

$$\phi^{(1,00)} = \phi^{(0,00)}. \quad (2.8)$$

A sensitivity analysis for this assumption can be performed by relaxing it and then comparing the posterior distributions of the estimated potential outcomes for Never Takers. The most straightforward way to do this is not to impose the constraints in Equations (2.6), (2.7) and (2.8) and examine how the posterior distributions of

estimated finite-sample PCEs change (e.g., Imbens & Rubin, 1997; Mattei et al., 2013). In fact, when assuming the ER for Never Takers, PCEs for this stratum are forced to be equal to zero. When relaxing it, they have a nondegenerate posterior distribution, which could be centered around zero, thus lending credibility to the ER, or not. This is the sensitivity analysis we perform later in the paper. In Appendix A.1, we describe an equivalent way of performing this sensitivity analysis.

2.4.3 Bayesian inference

Let us denote the full parameters vector as $\boldsymbol{\psi} \in \{\alpha, \boldsymbol{\alpha}_W, \rho^{(0,00)}, \rho^{(1,00)}, \rho^{(0,01)}, \rho^{(1,01)}, \boldsymbol{\rho}_X, \beta^{(0,00)}, \beta^{(1,00)}, \beta^{(0,01)}, \beta^{(1,01)}, \boldsymbol{\beta}_X, \phi^{(0,00)}, \phi^{(1,00)}, \phi^{(0,01)}, \phi^{(1,01)}\}$. Let $\pi_i^{(g)} \equiv p(G_i = g | \mathbf{W}_i, \boldsymbol{\psi})$ for $g \in \{01, 00\}$ and $f_i(y)^{(z,g)} \equiv p(Y_i(z) = y | G_i = g, \mathbf{X}_i, \boldsymbol{\psi})$ for $z \in \{0, 1\}$ and $g \in \{01, 00\}$. The observed data likelihood results in the following finite mixture model likelihood:

$$\begin{aligned} \mathcal{L}_{obs}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{W}, \mathbf{X}) = & \prod_{i: Z_i=0} \left[\pi_i^{(00)} \cdot f_i(Y_i)^{(0,00)} + \pi_i^{(01)} \cdot f_i(Y_i)^{(0,01)} \right] \times \\ & \prod_{i: Z_i=1, D_i=0} \left[\pi_i^{(00)} \cdot f_i(Y_i)^{(1,00)} \right] \times \prod_{i: Z_i=1, D_i=1} \left[\pi_i^{(01)} \cdot f_i(Y_i)^{(1,01)} \right]. \end{aligned}$$

Let $p(\boldsymbol{\psi})$ denote the joint prior distribution of $\boldsymbol{\psi}$. The posterior distribution,

$$p(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{W}, \mathbf{X}) \propto p(\boldsymbol{\psi}) \times \mathcal{L}_{obs}(\boldsymbol{\psi} | \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{W}, \mathbf{X})$$

is analytically intractable due to the complex mixture structure of the likelihood function. As explained in Section 1.5, a common strategy to address this issue is to perform a Gibbs Sampling that first samples the principal strata memberships and then assesses the posterior distribution of the parameters, and thus the posterior of the causal estimands, given the augmented data. Here, we exploited Stan (Stan Development Team, 2014), a free and open-source C++ software that performs Hamiltonian Monte Carlo using the no-U-turn sampler to get approximate simulations from the posterior distribution (Hoffman, Gelman, et al., 2014). We use three independent chains from different starting values. We run each chain for 2000 iterations, discarding the first half and saving the other half. The Markov chains mix well (see Appendix A.2 for convergence checks). We combine the three chains and use the remaining 3000 iterations to draw inference.

2.5 Results

2.5.1 With Exclusion Restriction for Never Takers

Principal strata Let us now evaluate the results relative to the principal strata size and characteristics when ER for Never Takers is stated. Regarding their size, the estimated posterior probabilities of principal strata membership are reported in Table 2.6. The vast majority of the sample is made of Compliers (posterior mean

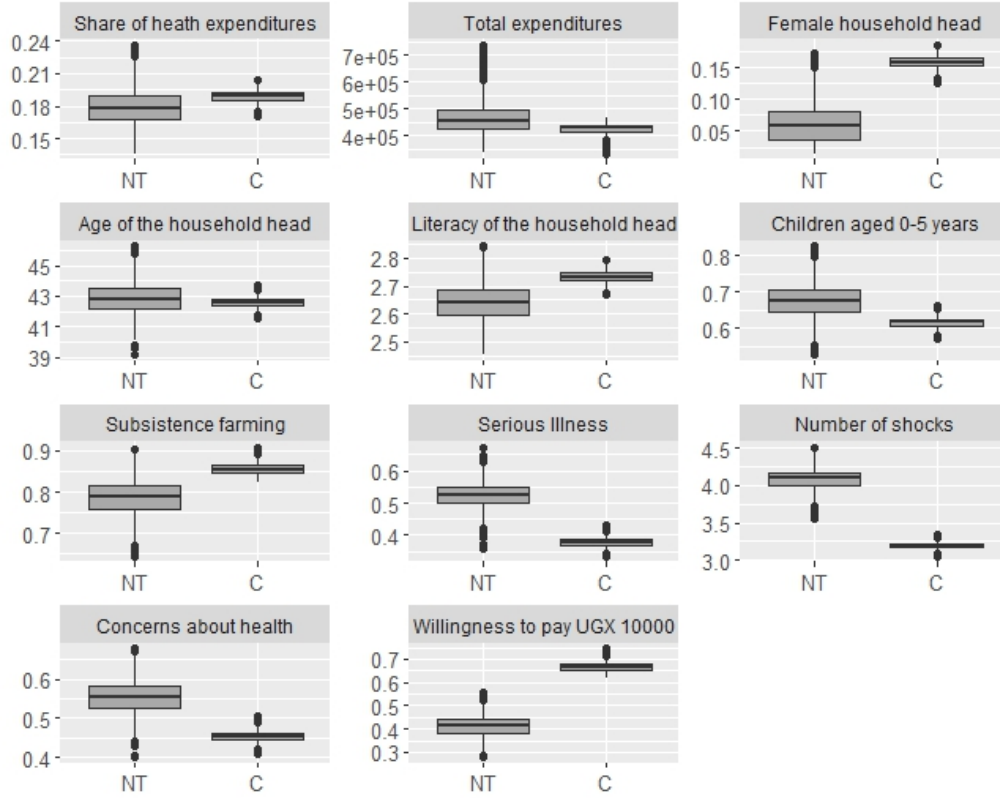
TABLE 2.6: Estimated posterior probabilities of principal strata membership when ER for Never Takers is assumed

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>
Proportion of Compliers	0.768	0.022	0.724	0.807
Proportion of Never Takers	0.232	0.022	0.193	0.276

76.8%, 95% credible interval 72.4%–80.7%). This means that the vast majority of the households would accept the offer to enroll in the CHF scheme after attending sensitization sessions. Examining the characteristics of the households that, instead, would opt out can provide us with insight into the reasons behind their lack of compliance.

We summarize the distributions of the characteristics of the principal strata in Figure 2.2. From this figure, it is evident that Compliers and Never Takers are somewhat different with regards to almost every aspect considered. First and foremost, Never Takers have a lower share of health expenditures but, at the same time, higher total expenditures. Therefore, one could suppose that these households would not accept to enroll in the CHF scheme because they do not need it. However, to a closer look, it is evident that this is not the case. Indeed, Never Takers more often have children aged 5 or less, which usually implies more healthcare costs, and also a higher incidence of serious illness among household’s members. It is noteworthy that Never Takers faced more difficulties than Compliers in the year prior to the intervention, since they experienced a remarkably higher number of shocks. These shocks include not only the occurrence of serious illness but also the death of household members, as well as the occurrence of conflicts, thefts, fires, droughts, floods, and other episodes that may partially explain the higher total expenditures and that are likely to have a major impact on the financial well-being of the family. Moreover, Never Takers declared to have concerns about the affordability of healthcare services more often

FIGURE 2.2: Boxplots summarizing the distributions of the covariates in the principal strata model by principal stratum; “NT” stands for Never Takers and “C” for Compliers



than Compliers. In addition, they are less willing to pay 10000 UGX/year per household member in order to have a health insurance; in the light of what we have said so far, this is reasonably due to the amount of money required, rather than to a lack of willingness to have a health insurance. Given the complete picture, it does not seem plausible that Never Takers did not need or did not want a health insurance, but rather that they could not afford it. Finally, we found further differences between Never Takers and Compliers with respect to the literacy level and gender of the household head. In particular, the former confirms the lower socio-economic status of Never Takers, while the latter indicates that female-headed households are more likely to be Compliers.

Estimated Principal Causal Effects Since we have assumed the ER for Never Takers, which implies the absence of any treatment effect for this subgroup, we now focus on the estimated treatment effects for the subgroup of Compliers. The

estimated finite-sample mean potential outcomes and PCEs for the latter principal stratum are reported in Tables 2.7 and 2.8, respectively.

TABLE 2.7: Summary statistics of the posterior distributions of the estimated potential outcomes, and their relevant transformations, for Compliers when ER for Never Takers is assumed

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>
<i>For Compliers</i>				
$\overline{Y(1)}$	0.122	0.005	0.113	0.132
$\overline{Y(0)}$	0.167	0.023	0.128	0.224
$\mathbb{1}(Y(1) \geq 0.10)^a$	0.486	0.022	0.444	0.529
$\mathbb{1}(Y(0) \geq 0.10)$	0.629	0.061	0.503	0.746
$\mathbb{1}(Y(1) = 0)^b$	0.086	0.015	0.060	0.119
$\mathbb{1}(Y(0) = 0)$	0.033	0.026	0.000	0.098

^ai.e., the average proportion of catastrophic health expenditures if encouraged.

^bi.e., the average proportion of zero health expenditures if encouraged.

TABLE 2.8: Summary statistics of the posterior distributions of the estimated PCEs for Compliers when ER for Never Takers is assumed

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>	<i>p($\cdot < 0$)</i>
$HEE^{(01)}$	-0.045	0.024	-0.102	-0.004	0.985
$CHEE^{(01)}$	-0.143	0.067	-0.269	0.000	0.975
$ZHEE^{(01)}$	0.053	0.032	-0.022	0.104	0.056

The estimated share of health costs over the total monthly expenditures is fairly high for Compliers in the absence of intervention (posterior mean 16.7%, 95% credible interval 12.8% – 22.4%); however, it is significantly reduced by attending the sensitization sessions and enrolling into the CHF scheme (posterior mean 12.2%, 95% credible interval 11.3% – 13.2%). The intervention is shown to have a considerable positive impact on catastrophic health expenditures as well; indeed, their average incidence is 62.9% (95% credible interval 50.3% – 74.6%) in the absence of intervention, but it reduces to 48.6% (95% credible interval 44.4% – 52.9%) when the intervention is offered and undertaken. Consistently, in the latter case there is also a noteworthy increase in the incidence of zero monthly healthcare expenditures, which reaches 8.6%, on average (95% credible interval 6.0% – 11.9%); indeed, in the absence of the intervention, this incidence is, on average, 3.3% (95% credible interval 0.0% – 9.8%), that is, it amounts to less than half.

In light of these findings, the intervention has accomplished its goal of reducing health care costs for Compliers. The estimated effects are the result of attending the sensitization sessions and participating in the CHF scheme as a whole. In fact, on the one hand, sensitization was aimed at raising awareness on the importance of using health care services promptly when needed; on the other hand, the prompt access to health care services was facilitated by the possibility of using the common fund to cover the costs at the immediate time of need. Avoiding waiting for worsening health conditions before seeking medical consultation or treatment avoids having to pay for expensive emergency services. Moreover, the greater incidence of zero monthly health costs resulting from the intervention indicates that households could have further changed their health-seeking behaviors realizing the importance of using free prevention services instead of waiting for health problems to occur.

2.5.2 Without Exclusion Restriction for Never Takers

Principal strata Let us now evaluate the results relative to the principal strata size and characteristics when ER for Never Takers is not assumed. The posterior distributions of the proportion of Compliers and Never Takers almost entirely overlap with those obtained when assuming it. We report some summary statistics in Table 2.9. Not only the size of the principal strata but also their characterization remains

TABLE 2.9: Estimated posterior probabilities of principal strata membership when ER for Never Takers is not assumed

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>
Proportion of Compliers	0.788	0.022	0.741	0.827
Proportion of Never Takers	0.212	0.022	0.173	0.259

essentially the same. We report the boxplots summarizing the distributions of the covariates by principal stratum in Appendix A.3.

Estimated Principal Causal Effects Since we do not assume, at this stage, the ER for Never Takers, we allow for the presence of encouragement effects for this subgroup as well as for that of Compliers. For both principal strata, we report some summary statistics of the estimated finite-sample mean potential outcomes in Table 2.10. From this Table it is evident that the share of health costs over the total monthly expenditures is extremely high for Never Takers when they do not attend

TABLE 2.10: Summary statistics of the posterior distributions of the estimated potential outcomes, and their relevant transformations, within each principal stratum when ER for Never Takers is not assumed

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>
<i>For Compliers</i>				
$\overline{Y(1)}$	0.122	0.005	0.112	0.133
$\overline{Y(0)}$	0.155	0.017	0.122	0.189
$\mathbb{1}(Y(1) \geq 0.10)$	0.485	0.022	0.444	0.530
$\mathbb{1}(Y(0) \geq 0.10)$	0.607	0.061	0.481	0.717
$\mathbb{1}(Y(1) = 0)$	0.088	0.016	0.061	0.122
$\mathbb{1}(Y(0) = 0)$	0.024	0.020	0.000	0.076
<i>For Never-Takers</i>				
$\overline{Y(1)}$	0.113	0.013	0.092	0.142
$\overline{Y(0)}$	0.242	0.061	0.135	0.379
$\mathbb{1}(Y(1) \geq 0.10)$	0.430	0.046	0.346	0.520
$\mathbb{1}(Y(0) \geq 0.10)$	0.609	0.128	0.333	0.836
$\mathbb{1}(Y(1) = 0)$	0.066	0.027	0.035	0.133
$\mathbb{1}(Y(0) = 0)$	0.187	0.103	0.035	0.432

sensitization sessions; specifically, it amounts to 24.2% of the total monthly expenditures, on average (95% posterior credible interval 13.5%–37.9%). This mean is much higher than the threshold above which monthly health expenditures are considered “catastrophic”, (i.e., 10% of the total monthly expenditures). It follows that also the incidence of catastrophic health expenditures is extremely high for this principal stratum in the absence of sensitization: in fact, in that case, the estimated percentage of Never Takers that have catastrophic health expenditures is, on average, 60.9% (95% posterior credible interval 33.3%–83.6%). At the same time, however, the estimated incidence of zero health expenditures among these households is high as well: specifically, on average, 18.7% of Never Takers is estimated to have zero monthly health expenditures when not receiving sensitization (95% posterior credible interval 3.5%–43.2%). Given the overall characteristics of this stratum discussed in Section 2.5.1, it seems plausible that most of these households could not afford any health expenditures.

On the other hand, the estimates reported in Table 2.10 for Compliers when they do not attend sensitization sessions and thereby do not participate in the scheme are

similar to those obtained under the ER for Never Takers. Hence, the disadvantage we found, based on pre-intervention characteristics, of Never Takers compared to Compliers persists or is even worsened at follow-up if no appropriate intervention is implemented.

The resulting estimates of the finite-sample PCEs are reported in Table 2.11. From

TABLE 2.11: Summary statistics of the posterior distributions of the estimated PCEs when ER for Never Takers is not assumed

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>	<i>p</i> ($\cdot < 0$)
$HEE^{(01)}$	-0.033	0.018	-0.069	0.002	0.966
$CHEE^{(01)}$	-0.122	0.066	-0.246	0.016	0.955
$ZHEE^{(01)}$	0.064	0.027	0.000	0.110	0.021
$HEE^{(00)}$	-0.129	0.063	-0.269	-0.018	0.991
$CHEE^{(00)}$	-0.180	0.138	-0.431	0.113	0.882
$ZHEE^{(00)}$	-0.121	0.108	-0.370	0.043	0.873

this Table, it is evident that beneficial effects of the intervention exist for both the principal strata, and that they are even bigger for Never Takers than for Compliers. In fact, we estimate a significant reduction of the share of health costs over the total monthly expenditures that averages -3.3% (95% credible interval -6.9% – 0.2%) for Compliers, while it reaches an average of -12.9% (95% credible interval -26.9% – -1.8%) for Never Takers. This effect might seem counterintuitive because Never Takers do not actually participate in the program, while Compliers do. However, the different characteristics of Never Takers and Compliers may help explain the finding. First, Never Takers have much higher health care costs in the absence of intervention (see Table 2.10), and thus significantly more room for improvement than Compliers. Moreover, the disadvantages Never Takers face, based on their observed characteristics, probably make them more sensitive and responsive to the principles taught during the sensitization sessions. Of course, this does not mean that participating in the program is detrimental and that it would be better to attend only the sensitization sessions; in fact, Compliers seem to benefit from the program. The results suggest that, for some individuals, a well-designed sensitization may already produce some beneficial results.

This is reflected in the incidence of catastrophic health expenditures; in particular, we estimate that, on average, 12.2% fewer Compliers and 18.0% fewer Never Takers

have catastrophic health expenditures when assigned to the encouragement. About the former effect, $CHEE^{(01)}$, the 95% credible interval covers zero, although the posterior probability of this effect being negative is very high (0.955). The same is true for the latter effect, $CHEE^{(00)}$, which has a posterior probability of being negative of 0.882.

Finally, the intervention also has an impact on the incidence of zero monthly health expenditures. On the one hand, we estimate that 6.4% more Compliers have zero monthly health expenditures. As previously argued in Section 2.5.1, this effect could be due to the more widespread use of free preventive services. On the other hand, we estimate that 12.1% fewer Never Takers have zero health expenditures, which is equivalent to saying that 12.1% more Never Takers spend some money over the course of a month for medical or health services. Given the observed characteristics of Never Takers, we believe that most of them were facing financial hardship and therefore would not have spent any money for apparently postponable health problems if not encouraged to join the program; however, if encouraged, they are persuaded to seek for medical consultation or treatment. In fact, during the sensitization sessions, these households were instructed on the importance of using the necessary health care services without stalling because this would have improved the health status of household members and this, in turn, would have prevented them from facing additional health costs in the future. The 95% credible interval for the former effect, i.e. $ZHEE^{(01)}$, does not include the zero, while that for the latter effect, i.e. $ZHEE^{(00)}$, includes the zero but it is almost entirely negative.

2.6 Posterior predictive checks

We evaluated the goodness-of-fit of the models, when we did not assume the ER for Never Takers, by conducting posterior predictive checks (Rubin, 1984). For this purpose, we chose to refer to the discrepancy measures defined in Section 1.6, and to estimate the corresponding Bayesian p -values. The results are reported in Table 2.12. All the estimated PPPVs for both Never Takers and Compliers are far from zero and one, suggesting a good model fit.

TABLE 2.12: Estimated PPPVs for Never Takers and Compliers

Discrepancy Measure	Never Takers ($G_i^{study} = 00$)	Compliers ($G_i^{study} = 01$)
$SI_{1,0,g}^{study}$	0.491	0.549
$NO_{1,0,g}^{study}$	0.477	0.588
$SN_{1,0,g}^{study}$	0.511	0.545

2.7 Conclusions

Although the statistical literature is abundant with studies in which PS is used to analyze experimental data characterized by non-compliance, in the economics literature the use of the IV approach is still well-established. This work shows the consequences of violations of the ER on which the IV method relies in the impact evaluation of a pilot intervention of CHF implemented in rural Uganda. This sensitivity analysis is conducted by using a Bayesian model-based PS first assuming an ER and then relaxing it. Results show that this intervention has positive effects not only on Compliers, i.e. households that would join the scheme upon receiving an initial sensitization and offer to enroll, but also on Never Takers, i.e. households that would refuse to join the scheme despite the initial sensitization and offer to enroll. The effects for Compliers are associative, meaning that they may derive from the initial sensitization and offer to enroll but also from participating in the scheme, while those for Never Takers are dissociative, meaning that they can only derive from the initial sensitization and offer to enroll. The presence of the latter is ruled out when the ER is assumed; however, our analysis provides evidence that they do exist.

More specifically, the analyses yield the following results. For Compliers, the intervention is shown to be strongly effective in reducing the share of health costs over the total monthly expenditures and the incidence of catastrophic health expenditures as well. Thus, the scheme improves financial protection for Compliers allowing them to promptly use health care services by delaying the economic burden of health expenses. Moreover, it increases the incidence of zero health expenditures, i.e., it increases the number of Compliers having zero monthly health costs. This might indicate a change in health-seeking behaviors leading to an increase in the

utilization of preventive services free of charge and a decrease in that of expensive emergency services.

For Never Takers, the strongest causal effect of the intervention is in reducing the share of health costs over the total monthly expenditures and, to a weaker extent, in decreasing both the incidence of catastrophic and zero health expenditures. Hence, the intervention reduces the often excessively high health costs for some households while inducing others to spend something for health that they would not have spent otherwise. Therefore, we believe that Never Takers are of two different types. Some do not join the scheme because they cannot afford the initial amount required to participate (5000 UGX/year per member). However, by attending the sensitization sessions, they learned the importance of trying to save a little money to be able to deal with health problems that might occur, but also to spend on health today to avoid worse health problems and higher expenses in the future. Others are already facing financial hardship due to health problems that occurred in the previous year and thus do not join the scheme to save the initial money required to participate. They do, however, have many health-related expenses with and without the intervention but spend, in general, less because the sensitization makes them understand how to manage their finances more effectively.

We conduct a supplementary analysis (reported in Appendix [A.4](#)) aimed at assessing if this effect on health costs was accompanied by an effect on the amount of monthly savings. The results indicate some positive effects for Compliers, but not for Never Takers. The lack of a clear effect for the latter group might be attributable to the fact that sensitization makes some households have fewer health costs and thereby more savings, but also makes others spend some money on health that they would not have spent otherwise, thereby preventing their savings from increasing substantially.

These analyses have two important implications in terms of policy making. The first is the fact that those that would benefit most from a health insurance do not seem to have the resources needed to take it out. The second is the fact that activities of sensitization alone can be extremely effective to reduce vulnerability against health costs for all population groups (not only enrolled members). The implementation of awareness-raising campaigns has thus potential to significantly contribute to enlarging health coverage and reducing impoverishment for poor and vulnerable population groups who receive information and change their behaviors for health financing.

Chapter 3

Non-inferiority trials with non-adherence: a Simulation Study

3.1 Research aim and objectives

NI trials test whether an experimental treatment is not unacceptably worse, in terms of efficacy, than an active control already in use. More precisely, the goal of these studies is to assess, with a given degree of confidence, whether the difference in efficacy between the two treatments is less than a certain pre-specified amount, called the NI margin. The rationale beyond these trials is that the new treatment may not be more effective than currently available standard treatments, but it may offer important advantages such as an improved safety profile, greater ease of administration, better compliance, or reduced cost (Schumi and Wittes, [2011](#) for a review). Nowadays, highly efficacious SoC treatments have been established for many diseases. This makes it unethical to conduct placebo-controlled trials without any active comparator, and, in any case, makes it unlikely to conclude the superiority of the new treatment over the SoC (Mauri & D'Agostino Sr, [2017](#)). In light of this, it is clear why conducting NI studies has become increasingly widespread (Suda et al., [2011](#)).

Despite their appeal, NI trials present substantial methodological challenges. One of the main ones stems from the lack of a univocal strategy to handle, in the analysis phase, subjects who did not adhere to the randomized treatment. Treatment non-adherence is conceptually different from non-compliance, as is mostly considered as a binary indicator for whether a patient was not taking the randomized treatment as prescribed in the protocol to the endpoint. In the context of clinical trials, non-compliance is, instead, often referred to as how many doses the patient missed while

still continuing the treatment.

Clearly, non-adherence can result from various post-randomization complications, such as side effects of the treatment, its lack of efficacy, taking rescue medication, forgetfulness, or administrative reasons. Yet, standard measures of treatment benefit, which have been used for many years now, do not adequately account for these events or at all. Among them is the ITT principle, whereby patients are analyzed as part of the treatment group to which they were randomized, ignoring information on treatment adherence. In NI trials, an ITT analysis provides an estimate of the causal effect of prescribing an experimental treatment instead of the SoC, not of actually taking it (Nagelkerke et al., 2000). Furthermore, protocol violations typically dilute treatment differences, making treatment arms appear similar. In superiority trials, this results in a conservative estimate of the difference in efficacy between the two treatments, i.e., the ITT estimator is biased towards the null hypothesis of no difference between treatments. Conversely, in NI trials, it can be anti-conservative, making it more likely to conclude NI of the experimental treatment compared to the SoC (Food and Drug Administration, 2016).

An alternative conventional approach is to perform a Per-Protocol (PP) analysis, which is essentially an ITT analysis on the subset of patients who adhered to the treatment as specified in the study protocol (the so-called “PP population”). The problem with such an approach, however, is that often it does not compare like with like: the adherents in one treatment arm might systematically differ, in terms of prognostic factors, compared with adherents in the other arm. In that case, the PP estimator provides biased treatment effect estimates which, in NI trials, may result in false conclusions of NI (Matilde Sanchez & Chen, 2006).

To date, there is no general directive on what is the best method to analyze NI trials in the presence of non-adherence. The Food and Drug Administration (FDA) discouraged the use of the ITT principle but did not recommend any alternative (Food and Drug Administration, 2016). The Consolidated Standards of Reporting Trials (CONSORT) group recommended referring to both the ITT and the PP populations as a robustness check of the NI conclusion (Piaggio et al., 2012). However, the agreement between these results certainly does not guarantee that the conclusion is meaningful. More sophisticated methods, that attempt to account for the impact of non-adherence, such as IV methods and Inverse-Probability Weighting (IPW), are

available, but their use in NI trials remains extremely infrequent (see Dodd et al., 2022 for a review).

In any case, what drives a preference for one method of analysis over another seems to be the need to obtain a conservative estimate of the difference in efficacy between the two treatments being compared. We believe, however, that the adoption of a certain estimation method rather than another should be motivated by the appropriateness and relevance of the underlying estimand for the questions one wants to address.

Depending on the question being investigated in a trial, very different conclusions can be drawn. Therefore, the estimand of interest has to be explicitly reported, and its meaning has to be clear to all stakeholders so that fully informed decisions about medical interventions can be made (e.g., Akacha et al., 2017; Cro et al., 2022). After considering the perspectives and interests of patients, physicians, regulators, and payers, Akacha et al. (2017) developed a framework consisting of three fundamental questions that should guide the choice of the estimand: what is the proportion of patients who stop taking the randomized treatment due to (i) safety and (ii) lack of efficacy, and (iii) what is the efficacy and safety profile of the treatment when patients are actually able to take it as prescribed for its intended duration. An answer to the first two questions can be easily found by collecting information on why patients discontinue randomized treatment. The last question, on the other hand, is much more complicated to answer. To this end, since excluding patients who discontinue randomized treatment from the analysis may result in the treatment groups no longer being comparable, a causal inference framework must be adopted to define the estimand and construct a corresponding estimator.

In the E9(R1) guideline, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) Steering Committee listed various strategies to define an estimand that responds to the need to address intercurrent events, each reflecting a different clinical question of interest (European Medicines Agency, 2020). One of these is the so-called *principal stratum strategy*, which refers to the use of PS (Frangakis & Rubin, 2002). Principal strata can, indeed, be defined based on the joint potential values of treatment adherence. Then, it would be of major scientific interest to investigate the causal effect for the principal stratum of patients who would be able to adhere to both the treatments being compared. We call patients in this principal stratum *Always Adherents*. Since, by definition, Always Adherents would fully adhere to both treatments, a difference in

the outcomes that would follow constitutes a casual estimand which is net of the effect of treatment adherence – it can solely be interpreted as a different efficacy of the two treatments. Note that this principal stratum estimand is fundamentally different from the PP estimand because is defined on a common set of units (i.e., the principal stratum members); it is, therefore, a properly defined causal effect according to the RCM. Plus, it corresponds to question (iii) posed by Akacha et al. (2017). The same authors mention, in fact, the possibility of using PS, but also affirm that experience with such approaches is very limited in clinical drug development and that formal training would be necessary.

The focus on the effect for Always Adherents does not imply that the rest of the population is not of interest. The PS framework allows for the definition of causal effects specific for any principal stratum, i.e., also for those who would adhere only to one treatment, or to any of them. The recognition that the effects can be heterogeneous for different (latent) sub-populations is a distinct feature of this approach, since no other estimand in the ICH E9(R1) guidance (European Medicines Agency, 2020) contemplates this possibility.

We propose to adopt, in NI trials with non-adherence, a PS framework to define an estimand that we believe answers a clinical question of fundamental importance, namely, what is the causal effect for Always Adherents. We shed light on the meaning and relevance of this causal effect, and also provide the tools to correctly interpret causal effects for other principal strata. Also, we focus on some of the assumptions presented in Section 1.4, and discuss their plausibility in the context of NI trials by identifying situations in which they are more or less likely to hold. Finally, we enrich the discussion on the choice of the estimand by presenting the results of simulations studies in which the results obtained with methods targeting different estimands are compared in terms of performance measures and, in particular, NI conclusions.

3.2 Methodology

3.2.1 Basic notation and potential outcomes

We again adopt the notation introduced in Section 1.1, making some necessary specifications. The setting consists of a NI trial conducted on a sample of $i = 1, \dots, n$ patients. For each patient i , we observe a K -dimensional vector of pre-treatment variables, $\mathbf{X}_i$, the randomized treatment, Z_i , treatment adherence, D_i , and the

outcome of interest, Y_i . $Z_i \in \mathcal{Z} \equiv \{0, 1\}$, with $Z_i = 1$ indicating that patient i is randomized to the experimental treatment and $Z_i = 0$ that he or she is randomized to the SoC. $D_i \in \mathcal{D} \equiv \{0, 1\}$, with $D_i = 1$ if patient i adheres to the randomized treatment and $D_i = 0$ if he or she deviates in any way from the protocol before the endpoint. Finally, $Y_i \in \mathcal{Y} \equiv \{0, 1\}$, with $Y_i = 1$ denoting an adverse health outcome and $Y_i = 0$ a positive health outcome. We consider a binary outcome for the purposes of the simulations, although it might be categorical, continuous, or any type.

We denote with $\mathbf{Z}$, $\mathbf{D}$ and $\mathbf{Y}$ the n -dimensional vectors of, respectively, the treatment assignments, treatment adherences and outcomes for the whole sample, with i -th elements equal to Z_i , D_i and Y_i . We also denote with $\mathbf{X}$ the $n \times K$ matrix of observed pre-treatment variables, with i -th row equal to $\mathbf{X}_i$.

Finally, let $\mathbf{z}$ be a n -dimensional vector with i -th element equal to z_i : $\mathbf{z} \in \{0, 1\}^n$. The potential outcomes for the relevant post-assignment variables are $D_i(\mathbf{z})$ and $Y_i(\mathbf{z})$. Under SUTVA (Assumption 1), they can be written as $D_i(z_i) \equiv D_i(z)$ and $Y_i(z_i) \equiv Y_i(z)$. $D_i(z)$ is the potential adherence of patient i to treatment z ; $Y_i(z)$ is his or her outcome if randomized to treatment z . Furthermore, under SUTVA, $D_i = D_i(Z_i)$ and $Y_i = Y(Z_i)$.

3.2.2 Principal strata

To deal with the issue of non-adherence, we use the PS framework. Accordingly, we classify patients into (basic) principal strata defined by their potential adherence to the treatment conditions being compared. Specifically, the basic PS with respect to D_i classifies patients into basic principal strata based on the vector $(D_i(0), D_i(1))$. Here, we refer to a simplified scenario in which non-adherence is one-sided, meaning that it occurs in only one treatment arm, in particular in the control arm. Thus, for each patient i , $D_i(0) \in \{0, 1\}$, whereas $D_i(1) = 1$. As a result, the PS with respect to D_i classifies patients into two basic principal strata. Formally, let $G_i = (D_i(0), D_i(1))$, then the basic principal strata are given by the possible values of G_i :

$$(d_0, 1) := \{i : D_i(0) = d_0, D_i(1) = 1\} \text{ with } d_0 \in \{0, 1\}.$$

We relabel the possible values of G_i as d_01 , so $G_i \in \{11, 01\}$. These two basic principal strata are named as shown in Table 3.1. Fundamentally, each patient i

TABLE 3.1: Principal strata and their labels

$D_i(0)$	$D_i(1)$	<i>Principal Stratum</i>	<i>Stratum Label</i>
1	1	$G = 11$	<i>Always Adherents</i>
0	1	$G = 01$	<i>Non-Adherents</i>

belongs either to the stratum of *Always Adherents* (AA), i.e., he or she would fully adhere to the assigned treatment, regardless of which one it is; or to the stratum of *Non-Adherents* (NA), i.e., he or she would adhere to the experimental treatment entirely, but not to the SoC.

There are important extensions that go beyond this framework. More complex settings could be envisioned, such as those in which there are patients who would not adhere to either treatment. In that case, the stratum of Always Adherents would remain unchanged, but multiple types of Non-Adherents could be defined. Specifically, in addition to the strata depicted in Table 3.1, there would be an additional subpopulation comprising patients who would not adhere to either treatment (i.e., those with $D_i(0) = 0$ and $D_i(1) = 0$). The larger number of strata complicates the setting, but can still be accommodated by the methodology we are proposing. To keep the explanation as simple as possible, and to avoid losing focus on the subpopulation of interest, i.e., that of Always Adherents, we do not consider this possibility here. Another simplification is to consider adherence as “all or nothing”, whereas it may be partial. Jin and Rubin (2008), for example, proposed to use PS to account for “extended partial compliance”. In their framework, strata are defined based on the value of four variables: the potential proportion of the treatment (i.e., active and placebo) taken under each treatment assignment. The result is an infinite number of strata, and thus an infinite number of stratum-specific effects. Since we are interested only in the treatment effect for the stratum of those who would take either treatment as prescribed in the protocol for the entire trial duration, such an extension would unnecessarily complicate the setting. Finally, non-adherence could be time-dependent. Extension to non-fixed time strata is possible, although not straightforward – only a few examples exist in the literature. Among them, Mattei et al. (2020) applied PS to a longitudinal setting, in which patients randomized to control are allowed to switch to treatment, and the strata are defined by the potential switching time.

3.2.3 Definition and meaning of Principal Causal Effects

Typically, clinical trials are based on classical randomized experiments, thus treatment assignment is unconfounded by definition (Section 1.2). Yet, treatment adherence typically is not. Adopting PS and targeting stratum-specific causal effects solves this issue. We have already defined super-population and finite-sample average PCEs in Section 1.3.1. In this work, we specifically target super-population average PCEs marginalized over the distribution of the covariates:

$$\vartheta_{sp}^{(g)} \equiv \mathbb{E}[Y_i(1) - Y_i(0) | G_i = g], \quad (3.1)$$

for $g \in \{11, 01\}$. Hereafter, we simplify the notation by omitting the subscript: $\vartheta^{(g)} \equiv \vartheta_{sp}^{(g)}$. We name $\vartheta^{(11)}$ *Always Adherents Causal Effect* (AACE), and $\vartheta^{(01)}$ *Non-Adherents Causal Effect* (NACE).

For Non-Adherents, treatment adherence is affected by treatment assignment, because these patients would adhere to the experimental treatment but not to the SoC. In other words, PCEs for Non-Adherents are associative: they are gross effects that may be due both to an intrinsic difference between the alternative treatment regimes and to a different adherence to them. Suppose, for example, the expected values of the potential outcomes under the two alternative treatment regimes are equal for Non-Adherents, and that, consequently, the NACE is zero. Suppose, also, that these patients do not take the SoC at all if randomized to it, or that they stop taking it after a few doses, without taking any other treatment. Under these circumstances, the experimental treatment would be highly ineffective, because when taken as prescribed, it is equal to not taking any treatment or taking only a few doses of the SoC. Thus, the treatments would be far from equivalent. The zero NACE would possibly be the result of the different efficacy of the two treatments but also the different adherence to them. These effects are blended together and their disentanglement is not feasible¹.

For Always Adherents, treatment adherence is not affected by treatment assignment because they adhere to the treatment to which they are randomized, regardless of which one it is. PCEs for Always Adherents are, therefore, dissociative. Following the example above, an AACE equal to zero could not derive from a different

¹Actually, these effects could be disentangled by assuming an ER for Non-Adherents, which is equivalent to the ER for Compliers that we discussed in Section 2.3.4. However, such an assumption cannot be made in clinical trials evaluating different treatment regimes.

adherence to the two treatments: it must indicate that the two treatment regimes have actually the same efficacy when taken as prescribed. Therefore, differently from the ITT principle, which measures treatment *effectiveness*, the AACE measures treatment *efficacy*². For this reason, our interest lies in AACE. We believe the value of this estimand, compared to the NI margin, should determine whether the experimental treatment is non-inferior to the SoC.

3.2.4 Relevance of Principal Causal Effects

Building on the tripartite framework conceptualized by Akacha et al. (2017), and in line with the *principal stratum strategies* to construct an estimand discussed in the ICH E9(R1) guideline (European Medicines Agency, 2020), Qu et al. (2020) proposed targeting the causal effect for those who can adhere to one treatment or both treatments. These authors stated that the choice of the stratum of main interest, and thus of the specific principal stratum estimand, can depend on whether the control is a placebo or an active control and on the primary scientific objective of the trial. Lou et al. (2019) proposed, specifically, to use the AACE as a co-primary causal estimand in bioequivalence (BE) studies in the presence of intercurrent events. They called it the SACE, in line with the work of Rubin on applications of PS to studies with truncation by death (Rubin, 2006). The work of Lou et al. could be applied to NI trials as well as to BE studies. We, too, propose to target the AACE in NI trials, while departing from Lou et al. in terms of the chosen estimation strategy, as we will explain in Section 3.2.6.

In addition to being an efficacy measure that accounts for non-adherence, we believe that in many circumstances the AACE is more meaningful than other estimands that can be constructed using the strategies listed in the ICH E9(R1) guideline. These include *hypothetical strategies*, which involve reconstructing the outcome of those

²Effectiveness is a measurement of how well a treatment works in real-world clinical practice (Revicki & Frank, 1999). Efficacy, on the other hand, is the ability of a treatment to produce a desired effect under perfect adherence and (partially) controlled conditions, such as in a clinical trial. The advantage of targeting efficacy, rather than effectiveness, is that results are internally valid. Internal validity refers to the degree to which the study design, conduct, and analysis ensure that the observed effects of the treatment are attributable to the treatment itself. Internal validity is crucial in clinical research, as it determines the reliability of study conclusions. Once internal validity is established, external validity, or generalizability, can be evaluated by determining whether the study results apply to similar patients under circumstances that more closely approach real-world practice. Yet, if the results of a trial are not internally valid, external validity is irrelevant.

who do not adhere to the randomized treatment as if they had done so. The use of the resulting estimand, in fact, is limited to scenarios in which it makes sense to perform such reconstruction. For example, imputing the outcome of those who discontinued the randomized treatment and took rescue medication as if they had adhered to the randomized treatment up to the endpoint and did not need the rescue medication is of limited interest from a clinical and regulatory perspective. In fact, it involves both envisioning a scenario in which non-adherents adhered and in which they did not need the rescue medication. This scenario is too far from reality to be the one of primary interest. Therefore, it is necessary to think carefully about the cases that are suitable for the use of the hypothetical strategy and those that are not. In the situation just described, assessing NI by considering only patients who would adhere to both treatments and would not need the rescue medication would give more concrete indications of treatment efficacy, at least for the Always Adherents sub-population.

Regarding Non-Adherents, we believe they should not play a role in the NI assessment: roughly speaking, it is not appropriate to compare the NACE to the NI margin because, since NACE is an associative effect, it may lead to misleading NI conclusions. However, information on Non-Adherents is important to answer the first two questions of the tripartite framework (Akacha et al., 2017).

3.2.5 Commonly invoked assumptions for identification: a critical discussion

In the setting we refer to, patients randomized to the SoC who do not adhere to it are necessarily Non-Adherents, whereas those who do adhere to it are Always Adherents. However, the principal stratum membership of those randomized to the experimental treatment is not known. The latent nature of the principal strata complicates inference. As we explained in Section 1.4, some assumptions can be made, where appropriate, to facilitate the identification and estimation of the PCEs. In this Section, we critically discuss some of these assumptions.

First of all, the ER (Assumption 9), which we have extensively discussed in Chapters 1 and 2. The ER could be applied either to both Non-Adherents and Always Adherents, as it would be done in an IV analysis, or only to subjects whose intermediate is not affected by the assignment, here Always Adherents, as it is commonly done in applications of PS to experiments characterized by non-compliance. In any

case, since causal effects for Always Adherents are dissociative, imposing the ER for this principal stratum would amount to assuming that the value of AACE is zero. As a result, the very scientific question we are trying to answer would be discarded. Therefore, the Assumption is not really applicable in this context.

The applicability of monotonicity assumptions in medical applications is also not obvious. In the setting we are referring to, Strong Monotonicity (Assumption 6 for $z = 1$ and $c = 1$) holds by design, but this is not always the case. More generally speaking, it is not always plausible to assume the absence of one or more principal strata. Such an assumption is particularly critical in clinical trials comparing two active treatments, since side effects leading to treatment discontinuation may arise in both treatment arms. It is usually more reasonable in placebo-controlled trials, in which the possibility that some patients would not adhere to the placebo because of side effects but would adhere to the experimental treatment can be ruled out.

We have identified three fairly common designs of NI trials in which some formulation of Monotonicity is generally valid. The first is NI trials conducted to assess whether a particular component of the SoC could be omitted without a clinically relevant worsening of the prognosis of the patients (e.g., Mulvenna et al., 2016); or, also, whether a particular diagnostic examination could be omitted without significantly affecting clinical probability assessment (e.g., Righini et al., 2008). For example, the QUARTZ trial (Mulvenna et al., 2016) was designed to assess whether whole brain radiotherapy (WBRT) could be omitted from the standard treatment of patients with brain metastases from non-small cell lung cancer without a significant detriment in quality-adjusted life-years. WBRT causes, in fact, very strong side effects such as radiation-induced fatigue, hair loss, nausea, and scalp discomfort, which were expected to balance out the potential increase in overall survival and justify the omission of WBRT from the SoC. In this trial, it would be more than reasonable to assume that people who adhere to the standard treatment (dexamethasone and WBRT) would also adhere to the experimental one (dexamethasone alone). The second type of NI trials in which Monotonicity is likely to hold is NI trials designed to evaluate whether an experimental treatment regime, identical to the SoC but of shorter duration, is non-inferior to the SoC (e.g., Bielicki et al., 2021; Turkova et al., 2018). In similar studies, one could safely assume that those who would adhere to the long treatment regime would also adhere to the short one. The third type is NI trials in which the NI of a treatment identical to the SoC but administered in

a lower dose is assessed (e.g., Bielicki et al., 2021; Marks et al., 2018). In such studies, it would be reasonable to assume that those who would adhere to the SoC (higher dose) would also adhere to the experimental treatment (lower dose). Anyhow, we want to emphasize that, except for these macro-categories of NI study design which generally allow to safely assume Monotonicity, the decision to impose it should be based primarily on available clinical knowledge of the two treatments to be compared.

In addition to Strong Monotonicity (Assumption 6 for $z = 1$ and $c = 1$), Strong PI (Assumption 13) could be assumed to achieve the point identification of the PCEs. However, in the setting we refer to, we directly observe the average outcomes for Always Adherents and Non-Adherents randomized to the SoC, so we could exploit this information and assume Weak PI (Assumption 14) instead of Strong PI. Weak PI is composed of Equations (1.11) and (1.12). Since non-adherence is only in the control arm, Equation (1.12) is not applicable. Equation (1.11) implies that, conditional on an auxiliary variable (here, the vector of observed pre-treatment covariates: $A_i \equiv \mathbf{X}_i$), the same outcome under the experimental treatment is expected for both an Always Adherent and a Non-Adherent. This assumption cannot be verified from the observed data. Moreover, since it postulates (conditional) independence between “cross-world” counterfactuals, i.e., potential outcomes that are realized in different states of the world, namely $D_i(0)$ and $Y_i(1)$, it is extremely difficult to relate to a real-world situation to understand its plausibility.

3.2.6 Estimation strategies

Lipkovich et al. (2022) have recently published a comprehensive tutorial on the use of PS in the analysis of clinical trials, so we do not dwell on the estimation aspect and refer to this work for a review. Here, we only report the estimation strategies we adopt in the simulations: the principal score method and the Bayesian model-based estimation strategy. The former requires Monotonicity (Assumption 6 or 7) and Strong PI (Assumption 13), or Monotonicity and Weak PI (Assumption 14), or Monotonicity and different combinations of ERs and PI assumptions. The latter, instead, does not require any of the assumptions discussed in Section 3.2.5; however, without them, inference can be very imprecise even in large samples (Imbens & Rubin, 1997). In this work, we refer to the use of the Bayesian model-based estimation

strategy under Strong Monotonicity (Assumption 6 for $z = 1$ and $c = 1$), which in the setting we refer to, holds by design.

Clearly, other estimation strategies can be adopted. And, depending on the assumptions on which they rely, they can lead to point estimates but also to bounds on the PCEs. Lou et al. (Lou et al., 2019), for example, relaxed Monotonicity and proposed a method to derive large sample bounds for the AACE in the context of BE studies. Specifically, after recognizing that the PP estimand is equal to the AACE plus a bias term, they exploit the PP estimator and varied the value of the parameters that compose the bias term in meaningful ranges. In this work, we focus on obtaining point estimates, since bounds are often too wide to be informative.

Principal score method The *principal score* is defined as the probability of belonging to a certain principal stratum given the covariates (Hill et al., 2002). Here, the probability that a patient i is an Always Adherent given the covariates, e_i , can be defined as $e_i \equiv p(G_i = 11 \mid \mathbf{X}_i)$. There are several ways to estimate principal scores, one is via modelling (Jo & Stuart, 2009; Schochet & Burghardt, 2007); in particular, information on patients randomized to the SoC can be used to fit a regression of D on the covariates. Then, to estimate the AACE, under Weak PI (Assumption 14) the principal score weighting method (Ding & Lu, 2017; Stuart & Jo, 2015) can be adopted. This method involves first the use of the fitted model to predict principal scores of patients randomized to the experimental treatment; then, their use to weigh the observed outcome of these patients. The resulting estimator for the AACE is

$$\hat{\vartheta}_{pscores}^{(11)} = \frac{\sum_{i:Z_i=1} Y_i \cdot \hat{e}_i}{\sum_{i:Z_i=1} \hat{e}_i} - \frac{\sum_{i:Z_i=0 \& D_i=1} Y_i}{\sum_{i=1}^n \mathbb{1}(Z_i = 0 \& D_i = 1)}. \quad (3.2)$$

Bayesian model-based estimation strategy Let $\boldsymbol{\psi}$ be the parameter vector, including the parameters of the distribution of the principal stratum membership given the covariates and of the conditional distributions of the potential outcome given the covariates and the principal stratum. Assume a prior distribution $p(\boldsymbol{\psi})$. Let $\pi_i^{(g)} \equiv p(G_i = g \mid \mathbf{X}_i, \boldsymbol{\psi})$ for $g \in \{11, 01\}$ and $f_i^{(z,g)} \equiv p(Y_i(z) = 1 \mid G_i = g, \mathbf{X}_i, \boldsymbol{\psi})$ for $z \in \{0, 1\}$ and $g \in \{11, 01\}$. Given the latent nature of the principal strata, the observed data likelihood, namely $\mathcal{L}_{obs}(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$, results in the

following finite mixture model likelihood:

$$\begin{aligned} \mathcal{L}_{obs}(\boldsymbol{\psi}|\mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) = & \prod_{i:Z_i=0, D_i=1} \left[\pi_i^{(11)} \cdot f_i^{(0,11)} \right] \times \prod_{i:Z_i=0, D_i=0} \left[\pi_i^{(01)} \cdot f_i^{(0,01)} \right] \\ & \times \prod_{i:Z_i=1} \left[\pi_i^{(01)} \cdot f_i^{(1,01)} + \pi_i^{(11)} \cdot f_i^{(1,11)} \right]. \end{aligned} \quad (3.3)$$

Consequently, the posterior distribution

$$p(\boldsymbol{\psi}|\mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) \propto p(\boldsymbol{\psi}) \times \mathcal{L}_{obs}(\boldsymbol{\psi}|\mathbf{Z}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$$

is analytically intractable. This issue can be addressed by implementing a Gibbs Sampling that first samples the principal strata memberships and then assesses the posterior distribution of the parameters, and thus the posterior of the causal estimands, given the augmented data. We have already described this Bayesian model-based estimation strategy in detail in Section 1.5. We denote the corresponding estimator, for the super-population average PCEs, by $\hat{\vartheta}_{Bayes_sp}^{(g)}$ ($g \in \{11, 01\}$); hereafter, we simplify the notation as follows: $\hat{\vartheta}_{Bayes}^{(g)} \equiv \hat{\vartheta}_{Bayes_sp}^{(g)}$.

3.3 Simulations

We have already introduced, from a theoretical point-of-view, the differences between some of the available strategies for defining an estimand – and thereby for analyzing NI trials – in the presence of non-adherence. To enrich the discussion of the same topic, in this Section we report relevant empirical results obtained through simulation studies.

3.3.1 Planning

Simulation studies have been planned using the ADEMP approach of T. P. Morris et al. (2019), thereby defining *Aims*, *Data-generating mechanisms*, *Estimands*, *Methods* and *Performance measures*.

Aims. The simulation studies performed aim to provide a numerical evaluation of the impact of choosing an estimand over another, comparing the results obtained with methods targeting different estimands in terms of performance measures and in particular NI conclusions.

Data-generating mechanisms. We simulate $n_{sim} = 500$ datasets³, each consisting of $n = 2000$ subjects, assigned to the experimental treatment or SoC with an allocation ratio of 1:1. We generate three n -dimensional vectors of pre-treatment covariates, namely $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, common to all datasets. x_1 is a binary variable drawn from a Bernoulli distribution with probability 0.5, representing gender. x_2 and x_3 are instead created to represent, respectively, age and a frailty score. Specifically, x_2 is a semi-continuous variable drawn from a truncated normal distribution with mean 60, standard deviation 20, and lower and upper truncation points 18 and 80, respectively. Also x_3 is a semi-continuous variable, drawn from a truncated normal distribution with mean 5, standard deviation 3, and lower and upper truncation points 0 and 10, respectively. All covariates are centered and scaled by their standard deviation so that they have zero mean and standard deviation equal to 1.

We use these covariates to generate the principal strata. In particular, we assume the following probit model for the principal stratum membership conditional on the covariates:

$$\begin{aligned}\pi_i^{(11)} &\equiv p(G_i = 11 | \mathbf{X}_i) = p(G_i^* > 0), \\ G_i^* &= \mathbf{X}_i' \boldsymbol{\alpha} + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, 1), \\ \pi_i^{(01)} &\equiv p(G_i = 01 | \mathbf{X}_i) = 1 - \pi_i^{(11)},\end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$ and $\mathbf{X}_i = (1, x_{1i}, x_{2i}, x_{3i})'$. In the following simulations, we set $\alpha_0 = 0.5$, $\alpha_1 = 0.7$, $\alpha_2 = -0.8$, $\alpha_3 = -1$. On average, the resulting proportions of Always Adherents and Non-Adherents are 60.70% and 39.30%, respectively. Furthermore, Non-Adherents are more often males, older and more frail compared to Always Adherents.

We generate the potential outcomes by specifying two probit models given the principal stratum membership and covariates:

$$\begin{aligned}f_i^{(z,g)} &\equiv p(Y_i(z) = 1 | G_i = g, \mathbf{X}_i) = p(Y_i^*(z) > 0), \\ Y_i^*(z) &= \mathbf{X}_i' \boldsymbol{\eta}^{(z,g)} + \epsilon_{zi}, \quad \epsilon_{zi} \sim \mathcal{N}(0, 1),\end{aligned}$$

³Datasets are simulated using as input seed “212”.

where $\boldsymbol{\eta}^{(z,g)} = (\eta_0^{(z,g)}, \eta_1^{(z,g)}, \eta_2^{(z,g)}, \eta_3^{(z,g)})'$, for $z \in \{0, 1\}$ and $g \in \{11, 01\}$. We set $\boldsymbol{\eta}^{(z,g)}$ such as

$$\begin{aligned} Y_i^*(1) &= 0.5 - 0.4x_{1i} + 0.9x_{2i} + x_{3i} & \forall_{i:G_i=11} \\ Y_i^*(1) &= \eta_0^{(1,01)} + 0.7x_{1i} + 0.8x_{2i} - 0.6x_{3i} & \forall_{i:G_i=01} \\ Y_i^*(0) &= -0.9 - 0.2x_{1i} + 0.35x_{2i} + x_{3i} & \forall_{i:G_i=11} \\ Y_i^*(0) &= 0.5 + 0.8x_{1i} - 0.6x_{2i} + 0.9x_{3i} & \forall_{i:G_i=01} \end{aligned}$$

We vary the value of $\eta_0^{(1,01)}$ to focus on scenarios of increasing heterogeneity in the treatment effect for Always Adherents and Non-Adherents: in particular, $\eta_0^{(1,01)} = 0.6$, $\eta_0^{(1,01)} = 0.4$ and $\eta_0^{(1,01)} = 0.2$ in the three data-generating mechanisms (DGMs) considered.

Estimands The estimand of interest is $\vartheta^{(11)}$. The DGMs considered correspond to the true values of $\vartheta^{(11)}$ and $\vartheta^{(01)}$ reported in Table 3.2. In what follows, we will denote these true values, respectively, by $\vartheta_*^{(11)}$ and $\vartheta_*^{(01)}$.

TABLE 3.2: Simulation scenarios. The values reported in the Table are estimated on a simulated dataset of 10'000'000 units (by averaging over the distribution of the covariates)

	DGM		
	1: $\eta_0^{(1,01)}=0.6$	2: $\eta_0^{(1,01)}=0.4$	3: $\eta_0^{(1,01)}=0.2$
$\mathbb{E}[Y_i(1) G_i = 11]$	0.447	0.447	0.447
$\mathbb{E}[Y_i(0) G_i = 11]$	0.147	0.147	0.147
$\vartheta_*^{(11)}$	0.300	0.300	0.300
$\mathbb{E}[Y_i(1) G_i = 01]$	0.588	0.537	0.484
$\mathbb{E}[Y_i(0) G_i = 01]$	0.598	0.598	0.598
$\vartheta_*^{(01)}$	-0.010	-0.062	-0.114

In this Section, we consider settings in which the experimental treatment is inferior to the control. The experimental treatment is considered inferior to the control when $\vartheta_*^{(11)}$ is greater than, or equal to, a pre-specified NI margin, M . Here, we set $M = \vartheta_*^{(11)}$.

Methods To estimate $\vartheta^{(11)}$, we use both the principal score method and a Bayesian model-based estimation strategy, as described in Section 3.2.6.

When using the principal score method, we expect the estimates of $\vartheta^{(11)}$ to be biased, since $\boldsymbol{\eta}^{(1,01)} \neq \boldsymbol{\eta}^{(1,11)}$ in all the DGMs considered, leading to a violation of Weak PI (Assumption 14). Nonetheless, we include this estimation method to evaluate its robustness. This is of interest because, as explained in Section 3.2.5, Weak PI cannot be verified from the observed data, and is extremely difficult to discern whether it is plausible or not.

When using a Bayesian model-based estimation strategy, we specify the same models from which the data are generated⁴. We assume that the parameters are a priori independent and choose conjugate prior distributions for them: $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$ and $\boldsymbol{\eta}^{(z,g)} \sim \mathcal{N}(\boldsymbol{\mu}_\eta^{(z,g)}, \boldsymbol{\Sigma}_\eta^{(z,g)})$. In addition, we set the hyperparameters as follows: $\boldsymbol{\mu}_\alpha = (0, 0, 0, 0)$, $\boldsymbol{\Sigma}_\alpha = 3 \cdot \mathbb{I}_{4 \times 4}$, $\boldsymbol{\mu}_\eta^{(z,g)} = (0, 0, 0, 0)$, $\boldsymbol{\Sigma}_\eta^{(z,g)} = 3 \cdot \mathbb{I}_{4 \times 4}$ (where $\mathbb{I}_{a \times a}$ denotes the identity matrix of size $a \times a$).

We simulate the posterior distributions of the causal estimands of interest using two independent chains from different starting values. We run each chain for 150'000 iterations, discarding the first 75'000 iterations and saving every 20th iterations. The Markov chains mix well. We combine the two chains and use the remaining 7500 iterations to draw inference. See Appendix B.1.5 for a detailed description of the MCMC algorithm and Appendix B.2 for convergence checks.

We also use other approaches to analyze the simulated data. In particular, we implement a Bayesian model-based ITT analysis, a Bayesian model-based PP analysis, and IPW. We report the estimands these methods target, and details on the estimation strategies adopted, in Appendix B.1. For the Bayesian analyses, i.e., the ITT and PP analyses, we also report the results of convergence checks performed in Appendix B.2.

Performance measures For each of the n_{sim} data sets, NI is tested as follows:

$$\begin{cases} H_0 : \vartheta_*^{(11)} \geq M \text{ (the experimental treatment is inferior to the SoC by } M \text{ or more)} \\ H_1 : \vartheta_*^{(11)} < M \text{ (the experimental treatment is inferior to the SoC by less than } M) \end{cases}$$

If the upper bound of the two-sided 95% confidence (or credible) interval for $\vartheta_*^{(11)}$ is less than M , the NI of the experimental treatment relative to the control is established. Type I error is estimated with the proportion of the n_{sim} repetitions where

⁴A future extension might investigate the robustness of the method to model misspecifications.

this happens, i.e., the proportion of the n_{sim} repetitions where NI is mistakenly concluded. Note that, although the different methods we use to analyze the simulated data target different estimand, we conduct this testing procedure as if they were all targeting $\vartheta^{(11)}$. This allows us to highlight how much the study conclusions may change depending on the approach adopted.

In addition to type I error, we derive the following statistics: average point estimate over the repeated samples; bias, estimated as the average difference between the point estimate and the true value of the estimand; coverage of confidence (or credible) interval, i.e., the proportion of iterations for which the interval includes the true value of the estimand; empirical standard error (SE), i.e., the standard deviation of the point estimates; average model-based SE, which targets the empirical SE ($\mathbb{E}[\text{model SE}^2] = \text{empirical SE}^2$); and relative % error in model SE, which is the proportional error in the model-based SE, estimated using the empirical SE as the gold standard.

Furthermore, we provide Monte Carlo SEs for all performance estimates considered. Monte Carlo SEs quantify uncertainty in the estimates due to using finite n_{sim} .

For the formulas of the performance estimates and their Monte Carlo SEs, see T. P. Morris et al. (2019).

3.3.2 Results

The results of the simulations are reported in Table 3.3. The Monte Carlo SEs across the different DGM and methods are all acceptable, thus the conclusions we can draw by looking at the performance estimates are reliable (there is no need to increase n_{sim}).

From Table 3.3 we can notice that the average point estimates that would be obtained by targeting an ITT, PP and IPW estimand would be substantially different from those that would be obtained by targeting $\vartheta^{(11)}$. Relatedly, very different NI conclusions would be drawn: in the scenarios considered, with the ITT method the NI of the experimental treatment would be accepted with probability one despite the experimental treatment being inferior to the SoC. Also with PP and IPW methods the probability of concluding NI of the experimental treatment is fairly high: it ranges from a minimum value of 0.294 to a maximum value of 0.970. Clearly, it is higher in the scenarios leading to average point estimates more distant from the true

TABLE 3.3: Estimates of the performance measures of interest (Monte Carlo SEs in parentheses).

Performance measure	DGM	ITT	PP	IPW	Principal scores	Bayesian model-based
Average point estimate	$\eta_0^{(1,01)} = 0.6$	0.177 (-)	0.246 (-)	0.237 (-)	0.281 (-)	0.300 (-)
	$\eta_0^{(1,01)} = 0.4$	0.157 (-)	0.230 (-)	0.217 (-)	0.272 (-)	0.299 (-)
	$\eta_0^{(1,01)} = 0.2$	0.137 (-)	0.213 (-)	0.196 (-)	0.262 (-)	0.298 (-)
Bias ¹	$\eta_0^{(1,01)} = 0.6$	-	-	-	-0.019 (0.001)	0.000 (0.001)
	$\eta_0^{(1,01)} = 0.4$	-	-	-	-0.028 (0.001)	-0.001 (0.001)
	$\eta_0^{(1,01)} = 0.2$	-	-	-	-0.038 (0.001)	-0.002 (0.001)
Coverage ¹	$\eta_0^{(1,01)} = 0.6$	-	-	-	0.882 (0.014)	0.950 (0.010)
	$\eta_0^{(1,01)} = 0.4$	-	-	-	0.792 (0.018)	0.946 (0.01)
	$\eta_0^{(1,01)} = 0.2$	-	-	-	0.602 (0.022)	0.936 (0.011)
α	$\eta_0^{(1,01)} = 0.6$	1.000 (0.000)	0.656 (0.021)	0.294 (0.020)	0.120 (0.015)	0.032 (0.008)
	$\eta_0^{(1,01)} = 0.4$	1.000 (0.000)	0.874 (0.015)	0.518 (0.022)	0.212 (0.018)	0.032 (0.008)
	$\eta_0^{(1,01)} = 0.2$	1.000 (0.000)	0.970 (0.008)	0.738 (0.02)	0.400 (0.022)	0.036 (0.008)
Empirical SE	$\eta_0^{(1,01)} = 0.6$	0.018 (0.001)	0.022 (0.001)	0.061 (0.002)	0.022 (0.001)	0.027 (0.001)
	$\eta_0^{(1,01)} = 0.4$	0.018 (0.001)	0.023 (0.001)	0.061 (0.002)	0.022 (0.001)	0.028 (0.001)
	$\eta_0^{(1,01)} = 0.2$	0.018 (0.001)	0.023 (0.001)	0.061 (0.002)	0.022 (0.001)	0.029 (0.001)
Model SE	$\eta_0^{(1,01)} = 0.6$	0.018 (0.000)	0.022 (0.000)	0.049 (0.002)	0.023 (0.000)	0.027 (0.000)
	$\eta_0^{(1,01)} = 0.4$	0.018 (0.000)	0.022 (0.000)	0.049 (0.002)	0.023 (0.000)	0.027 (0.000)
	$\eta_0^{(1,01)} = 0.2$	0.019 (0.000)	0.022 (0.000)	0.049 (0.002)	0.023 (0.000)	0.028 (0.000)
Relative % error in model SE	$\eta_0^{(1,01)} = 0.6$	0.908 (3.194)	-0.458 (3.152)	-19.504 (4.459)	5.264 (3.333)	-2.61 (3.092)
	$\eta_0^{(1,01)} = 0.4$	0.88 (3.193)	-0.403 (3.154)	-19.382 (4.467)	4.596 (3.312)	-3.118 (3.077)
	$\eta_0^{(1,01)} = 0.2$	2.467 (3.243)	-0.588 (3.148)	-19.679 (4.428)	4.525 (3.309)	-4.613 (3.033)

¹ To investigate bias and coverage of estimators which do not target $\vartheta^{(11)}$ is beyond the scope of this paper, so we do not report these performance estimates for them.

value of $\vartheta^{(11)}$; this is a fair comparison given that empirical SE of all the methods is virtually invariant among the different DGMs considered.

It is worth noting that IPW has a considerably higher model SE than the other methods. The consequence is that, in a situation in which there is another method providing a similar average point estimate (as it is the case, in these simulations, of the PP method), IPW leads to mistakenly rejecting H_0 much less often (see the estimated α for the two methods).

Let us also examine the results obtained with the estimators that target $\vartheta^{(11)}$. The principal score estimator is, as expected, biased. The amount of bias is low but increases with increasing heterogeneity, and is reflected in undercoverage (i.e., coverage below 0.90 in all the scenarios considered). The decline in coverage is very rapid because both empirical and model SE are small. Conversely, the Bayesian model-based estimator is almost unbiased and has estimated nominal coverage of confidence intervals.

3.4 Practical considerations

3.4.1 Description of the principal strata

One of the major advantages of the Bayesian model-based PS analysis is that it allows getting a characterization of patients belonging to the different principal strata in terms of the observed covariates. For example, in the simulations presented so far, with the Bayesian model-based PS analysis we estimate Non-Adherents to be 39.3% of the population, and that, on average, about 30% of them are male, with an age of 61.5 years, and a frailty score of 6.4 (on a 1-10 scale). The remainder of the population is estimated to be Always Adherents, who are mostly males (63%), younger, on average, than Non-Adherents (mean age of 51.1 years), and less frail (mean frailty score of 4.1). Clearly, the different information on the principal strata one can get depends on which covariates are included in the principal stratum membership model.

We believe that this type of description of the members of the different latent strata can be of fundamental importance to clinicians. Indeed, they would be able to make more informed decisions about which treatment to prescribe for a given patient, in light of (i) the likelihood that that patient has to adhere to the different treatment options, i.e., his or her imputed principal stratum membership, and (ii) his or her expected potential outcomes given the imputed principal stratum membership and covariates.

3.4.2 Power

Through simulation studies, we have shown that the Bayesian model-based PS approach provides an estimator of the AACE, i.e., $\hat{\vartheta}_{Bayes}^{(11)}$, that is not anti-conservative. Specifically, the latter has been shown to lead to incorrect conclusions of NI in less than 3% of the datasets in the least favorable case where $\vartheta_*^{(11)} = M$ (Table 3.3).

Another important aspect to consider when proposing to use a method is its power. In NI testing, power is the probability of correctly rejecting the null hypothesis of inferiority; i.e., of detecting when the experimental treatment is truly non-inferior to the SoC. To evaluate power, we change the value of the NI margin from $M = \vartheta_*^{(11)}$ to $M = \vartheta_*^{(11)} + m$, with $m \in \{0.01, 0.02, \dots, 0.1\}$. Therefore, we now refer to scenarios in which H_0 is false (i.e., $\vartheta_*^{(11)} < M$). Moreover, m provides the distance between the NI margin and the true value of $\vartheta^{(11)}$: as m increases, it should be

easier to conclude the NI of the experimental treatment relative to the SoC. With this modified margin, we repeat the simulation under the first DGM model (Table 3.2) presented in Section 3.3 with $n \in \{2000, 4000\}$.

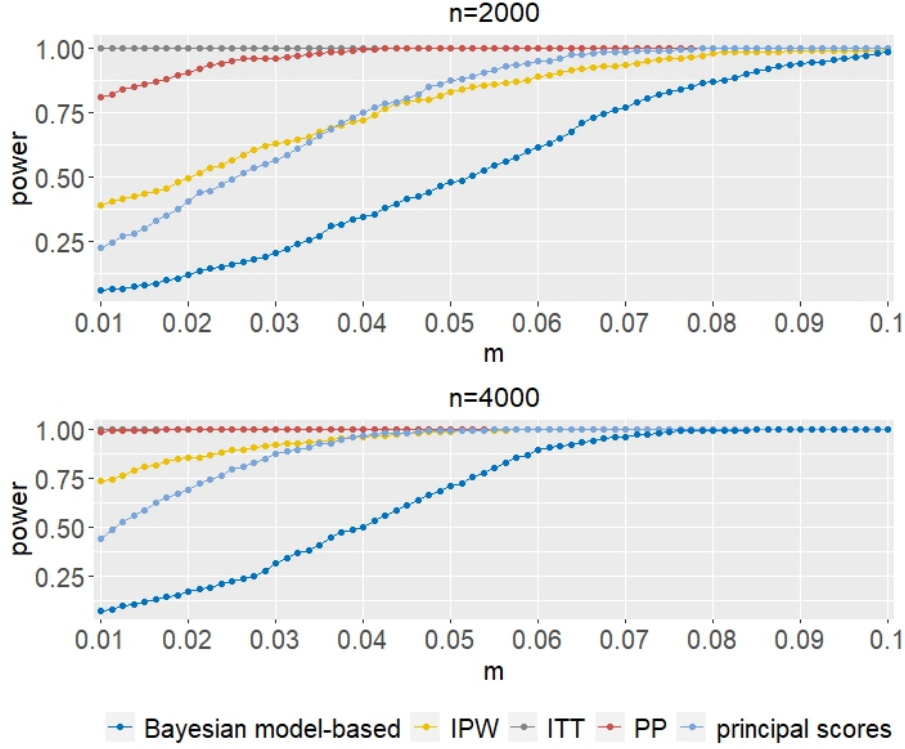


FIGURE 3.1: Estimated power of the different methods considered as m increases for $n \in \{2000, 4000\}$.

In Figure 3.1, we can visually appreciate that the estimated power of the different methods increases as m increases. Also, we can notice that the power resulting from the Bayesian model-based PS analysis is typically much lower than that resulting from the other analysis methods, regardless of the sample size. When $n = 2000$, it reaches an estimated value of 0.8 only when the difference between the NI margin and $\vartheta_*^{(11)}$, m , is more than 0.07. When $n = 4000$, it reaches the same result when $m \approx 0.055$. Therefore, the method would not be able to detect true NI of the experimental treatment in case the sample size is small and the true value of $\vartheta^{(11)}$ is very close to the NI margin. This is not a limitation specific to PS, but to all subgroup analyses. Indeed, the sample size needed to distinguish a treatment effect measured in two different subgroups (whether observed or latent) from the null is much larger than the sample size that would be needed if the hypothesis-testing

procedure concerned the overall treatment effect (Burke et al., 2015). Our advice is to adopt the method for analyzing big NI trials, so that true NI can be detected with a satisfying probability.

3.4.3 Weak predictors of adherence

In the simulations reported so far, $\widehat{\vartheta}_{Bayes}^{(11)}$ has shown good properties; among them, it has been shown to be almost unbiased and to have nominal coverage probability. Now we want to assess whether its performance is affected by the ability of the observed covariates to predict treatment adherence. Indeed, in the presence of weak predictors, there is greater uncertainty about the attribution of subjects to one principal stratum or another, and this may influence effect estimates.

To measure the strength of the predictors of treatment adherence, we computed the Area under the Receiver Operating Characteristic (AUROC) curve⁵ (Brown & Davis, 2006). This measure ranges between 0.5 and 1, with 0.5 indicating that patients are randomly assigned to the principal strata (which means that the covariates do not predict principal strata membership at all), and 1 indicating that all patients are correctly attributed to the principal strata they belong. In all the simulations reported in Section 3.3, the AUROC curve was 0.9022, so we were using very strong predictors of adherence. In this Section, we want to repeat those simulations using much weaker predictors of adherence. For this purpose, we use the same DGMs we have already used, but divide the coefficients in the principal stratum model by 3. The resulting AUROC curve is 0.7029. We believe that this value can be achieved in most clinical trials by carefully planning the baseline information to be collected. Note that this aspect is very important not only for a PS analysis but also to use the IPW method; in that case, in fact, good predictors of adherence are needed to construct the weights. More generally, it is important when using any method which takes into account non-adherence instead of excluding patients who do not adhere (e.g., PP) or ignoring non-adherence (e.g., ITT).

The simulation scenarios are reported in Table 3.4. Note that the values in the Table are different from the values in Table 3.2 because the modification of the coefficients in the principal stratum model leads to a partially different attribution of the patients to the principal strata. In these simulations, the proportions of

⁵The AUROC curve was computed on a single simulated dataset of 10'000'000 units.

TABLE 3.4: Simulation scenarios. The values reported in the Table are estimated on a simulated dataset of 10'000'000 units (by averaging over the distribution of the covariates)

	DGM		
	1: $\eta_0^{(1,01)}=0.6$	2: $\eta_0^{(1,01)}=0.4$	3: $\eta_0^{(1,01)}=0.2$
$\mathbb{E}[Y_i(1) G_i = 11]$	0.518	0.518	0.518
$\mathbb{E}[Y_i(0) G_i = 11]$	0.204	0.204	0.204
$\vartheta_*^{(11)}$	0.313	0.313	0.313
$\mathbb{E}[Y_i(1) G_i = 01]$	0.620	0.571	0.521
$\mathbb{E}[Y_i(0) G_i = 01]$	0.605	0.605	0.605
$\vartheta_*^{(01)}$	0.015	-0.034	-0.084

Always Adherents and Non-Adherents in the sample are, respectively, 55.91% and 44.09%, on average.

Again, we simulate the posterior distributions of the causal estimands of interest using two independent chains from different starting values. We run each chain for 150'000 iterations, discarding the first 75'000 iterations and saving every 20th iterations. We combine the two chains and use the remaining 7500 iterations to draw inference. For a detailed description of the MCMC algorithm, see Appendix B.1. We report that for a minor, but not negligible, percentage of datasets (10.9% of the simulated datasets, on average⁶), Markov chains of one or more parameters exhibit poor mixing. We believe this is due to the diminished ability of the covariates to predict treatment adherence, and consequently, principal strata memberships. We discuss this issue in detail in Appendix B.2, where we also report formal convergence checks.

For space constraints, we do not report here all the performance estimates presented in Table 3.3. Rather, we opt for a graphical representation of the main findings. Figure 3.2 shows lolly plots, which present estimates of bias, coverage, and empirical SE of the Bayesian model-based estimator of the AACE and the NACE, i.e., $\hat{\vartheta}_{Bayes}^{(11)}$ and $\hat{\vartheta}_{Bayes}^{(01)}$. By looking at these plots, we can notice that in all the scenarios considered the empirical SE of $\hat{\vartheta}_{Bayes}^{(11)}$ is higher than that estimated in the previous simulations (see Table 3.3), despite we have used the same n_{sim} . This is likely to be due to the greater uncertainty in the imputation of the principal strata memberships. However, the bias of both $\hat{\vartheta}_{Bayes}^{(11)}$ and $\hat{\vartheta}_{Bayes}^{(01)}$ is still very low. Also, their coverage is lower than

⁶Average calculated over the three DGMs considered.

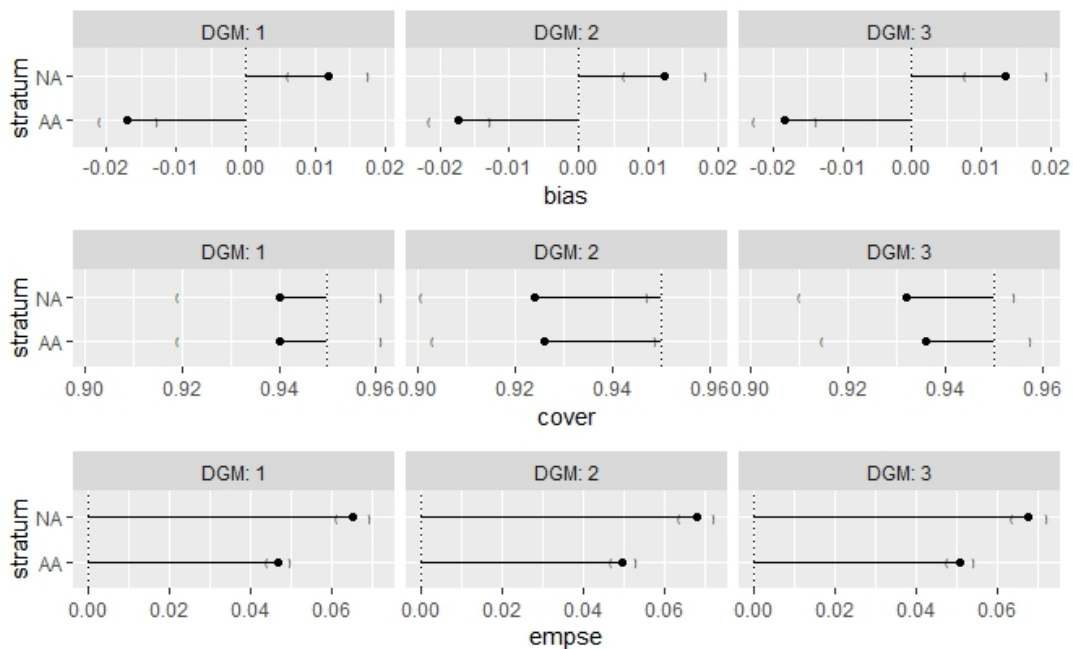


FIGURE 3.2: Lolly plots of bias (*bias*), coverage (*cover*) and empirical SE (*empse*) of $\hat{\vartheta}_{Bayes}^{(11)}$ and $\hat{\vartheta}_{Bayes}^{(01)}$. The dots indicate point estimates of these performance measures, while the parentheses indicate the corresponding confidence intervals based on Monte Carlo SEs.

0.95 but still quite satisfying, as it is between 0.924 and 0.940 in all the scenarios considered.

3.4.4 Sensitivity to model misspecifications

In the previous Section, we proved that the method can be useful even if strong predictors of adherence do not exist, albeit with limitations. However, the situation is different in the case where strong predictors of adherence exist but are not observed. Indeed, in this case, the model for principal stratum membership is necessarily misspecified. Although our estimation method does not require that all predictors of adherence are observed, the omission of one or more relevant covariates from the model decreases performance. It is important to acknowledge, however, that this is not a problem specific to Bayesian model-based PS analyses, but to all the estimation methods which involve explicit modeling of treatment adherence as a function of the covariates. Among them is the IPW, which involves the specification of a model for treatment adherence using patients in a given treatment arm, and its use to construct the weights. Yet, the alternative to using these methods is to keep

using the standard approaches, i.e., PP and/or ITT analyses. However, as already argued, the former can be strongly biased if the reasons for non-adherence are linked to the outcome, as is typically the case, whereas the latter does not target treatment efficacy at all.

3.5 Conclusions

In this work, we propose to use a principal stratum strategy to define an estimand that we believe answers a clinical question of fundamental importance, i.e., what is the treatment effect for patients who, if randomized, would be able to adhere to both the SoC and the experimental treatment whose NI is being evaluated. Specifically, we propose, on the one hand, to evaluate NI based on the causal effect for this latent subgroup of patients (we call them *Always Adherents*); and, on the other hand, to exploit the value of information on members of the other strata to better understand what are their peculiar characteristics and the reasons why they do not adhere to one or both the treatment regimes.

Through simulations, we show that a Bayesian model-based PS estimator could be a good choice as it is almost unbiased and has nominal coverage of credible intervals. Naturally, the assumptions on which the estimation method relies must be valid. We have identified three macro-categories of NI trial designs that are quite common in practice and in which the Monotonicity assumption that we impose is very likely to be valid. Remember, however, that other estimation strategies are available, and one could, for instance, compromise by using a method that allows relaxing the Monotonicity assumption and estimating bounds for the stratum-specific causal effects.

Furthermore, our work provides some further practical indications on when it is appropriate to conduct a Bayesian model-based PS analysis based on its estimated power and the strength of the available predictors of adherence. Both of these aspects are crucial in deciding which method to use, but the importance of the latter is often undervalued. Specifically, the simulation results show that a large sample size is required to detect true NI in the case where the causal effect to be estimated is very close to the margin, but relatively weak predictors of adherence (classified as such based on the AUROC curve) are in most cases sufficient to obtain an almost

unbiased estimate of stratum-specific causal effects. This demands the planning of larger NI trials to estimate causal effects specific to population subgroups.

Our considerations are based on the results of simulation studies, which refer to scenarios in which non-adherence occurs in only one treatment arm (specifically, every patient randomized to the experimental treatment adheres to it). However, non-adherence is more likely to occur in both treatment arms. Further simulation studies will be needed to assess the performances of the Bayesian model-based estimator we propose to adopt in this partially different context.

Chapter 4

Principal Stratification in the presence of network interference: an illustrative application to the Florentine Study

4.1 Research aim and objectives

Over the last few years, there has been an increasing interest in causal studies where units are part of networks (e.g., groups of friends). In this type of studies, social or physical interactions among units are a major concern, and thus, the no-interference assumption, under which the outcome of each unit cannot depend on the treatment assigned to others (Cox, 1958), is no longer plausible.

The literature on causal inference in the presence of networks has developed considerably in recent years (e.g. Forastiere, Airoidi, et al., 2021; Forastiere et al., 2018; Goldsmith-Pinkham & Imbens, 2013; Kang & Imbens, 2016; Ogburn et al., 2017; VanderWeele & An, 2013). The intuition characterizing this literature is that interference between units can be conveyed by the specific social links that tie these units together into a network, irrespective of treatment assignment. With this in mind, it is as if each unit was exposed to multiple treatments: its own treatment and the treatments of the other units in its network. The latter constitute an additional potential source of variation in the individual outcome. When this source is influential, i.e., when the treatments the neighbors receive significantly affect the individual outcome, we can say that *spillover effects* among neighbors are present (Del Prete et al., 2020).

Interference naturally arises in clustered encouragement designs (CEDs), where encouragement is randomized at the level of a cluster (e.g., class) of subjects (e.g., students) but compliance with the assigned encouragement is often at the individual level. CEDs are widely adopted, especially when randomization of the receipt of treatment is not feasible for ethical or practical reasons and the assignment of encouragement at the level of clusters appears preferable for administrative convenience, cost reduction, ethical issues and for improving treatment adherence. The literature abounds of studies designed as CEDs with individual non-compliance (e.g., Forastiere et al., 2016; Frangakis et al., 2002; S. S. Morris et al., 2004).

Clusters define a special network, where sets of units belonging to the same cluster are likely to share neighbors, but have no links with units belonging to different clusters. Therefore, in CEDs social interactions naturally occur among units belonging to the same group, and we can reasonably expect that among subjects who interact with one another, the treatment actually received by one subject may affect other subjects' outcomes, generating interference or spillover effects (Hong & Raudenbush, 2006; Hudgens & Halloran, 2008; Sobel, 2006). The presence of interference gives rise to intricate causal mechanisms, and disentangling them may provide precious information for researchers and policymakers to understand how an intervention works and how it can be improved.

A CED with individual non-compliance is the design underlying our illustrative study on teens' museum attendance. It is a field experiment based on an encouragement design clustered at the level of school classes, conducted in Florence (Italy) and designed to assess how appropriate incentives may prompt teens to visit museums in their free time. Throughout the text, we refer to this study as the Florentine study. A unique feature of this study is the availability of information about the friendship ties between the teens belonging to the same class, which makes the study a *CED with a hierarchical network structure*, with a high-level network defined by the clustered structure and a low-level network defined by the friendship structures within each cluster. The study was previously analyzed by Lattarulo et al. (2017) and Forastiere, Lattarulo, et al. (2021), but they neglect information on friendship ties, and thus, provide only a partial picture of the causal mechanisms by which the designed incentives affect teens' museums attendance.

Teens' museums attendance is a relevant topic because, while teens are likely to visit museums during school or family trips, museums are seldom their choice when it is

up to them to choose where to spend their free time (e.g. Hughes & Moscardo, 2019). In addition to granting free or low-price entrance to youths (Cellini & Cuccia, 2018), many museums are attempting to portray an image of educational and entertaining institutions and devising strategies to attract teens and young adults towards a tailored visit experience (Lattarulo et al., 2017; Manna & Palumbo, 2018). The hope is that such experience can raise further engagement with museums and cultural heritage in general (Kisida et al., 2014). In the Florentine experiment, classes of students are randomized to three incremental encouragements aimed at prompting students to visit Palazzo Vecchio, one of the most famous museums in the city. The treatment is the actual museum visit. The outcome of interest is the number of other museums visited in the follow-up period.

The stronger the encouragement, the higher the propensity of a student to visit Palazzo Vecchio is expected to be. However, students' compliance with the encouragement is imperfect, meaning that there are both students in classes assigned to the weakest encouragement who visit Palazzo Vecchio as well as students in classes assigned to the strongest encouragement who do not visit Palazzo Vecchio. We deal with the presence of non-compliance using the PS framework: we define principal strata based on the joint potential compliance statuses under the three encouragement conditions, and define stratum-specific causal effects. The role of PS in encouragement designs to draw inference on intention-to-treat effects within principal strata is uncontroversial (e.g., Angrist et al., 1996; Frangakis et al., 2002; Hirano et al., 2000; Imbens & Rubin, 1997).

How to deal with network interference in CEDs is still an open research issue. In principle, a number of causal pathways may take place within CEDs. The first path is the effect of the encouragement on a unit's outcome passing through that unit's treatment uptake. The second path is through spillover effects, i.e., changes in a unit's outcome originating from the treatment uptake by other units, which naturally arise among friends. The third path is the direct effect of the encouragement passing neither through the individual treatment uptake nor through spillovers, which ends up pooling further causal channels.

Two recent studies by Forastiere et al. (2016) and Forastiere, Lattarulo, et al. (2021) provide insightful contributions on the causal mechanisms in action in CEDs. Forastiere et al. (2016) propose to use the principal stratification framework to disentangle distinct causal mechanisms in CEDs, viewing the treatment variable as a

mediating variable and defining causal effects conditional on principal stratum membership based on hypothetical interventions on the treatment uptake. Forastiere, Lattarulo, et al. (2021) show how intention-to-treat effects within principal strata can be interpreted to get valuable information on different causal mechanisms for different types of units. Both Forastiere et al. (2016) and Forastiere, Lattarulo, et al. (2021) find evidence that encouragement and treatment effects in their CEDs are blended with spillovers, and highlight that disentangling these effects may be uneasy. We contribute to this literature by providing an innovative method for untying encouragement, treatment, and spillovers in CEDs with individual non-compliance under network interference, which capitalizes on the availability of within-cluster network data. Specifically, we take into account whether friends are treated, i.e., whether they actually visit Palazzo Vecchio. We summarize this information as the proportion of friends who actually take the treatment, and use this variable as a mediator in a principal stratum mediation analysis aimed at disentangling distinct causal mechanisms. We define principal stratum natural direct and indirect effects and principal stratum controlled direct effects, which are natural direct and indirect effects and controlled direct effects (Pearl, 2001; Robins & Greenland, 1992) for subpopulations of units defined by the principal stratification with respect to the individual receipt of treatment. We also introduce new causal estimands, which we refer to as principal controlled spillover effects, defined as comparisons within principal strata of hypothetical potential outcomes for the response variable under a fixed level of the encouragement and different values of the mediator. We interpret principal natural indirect effects and principal controlled spillover effects as spillover effects, which may be heterogeneous across different types of units defined by the compliance statuses. Principal natural and controlled direct effects, instead, can be viewed as pure effect of the encouragement or as a blend of the encouragement and the treatment/experience effects, depending on the principal stratum and the encouragement conditions being compared.

Inference on the causal estimands of interest is conducted under latent ignorability assumptions of the mediator using a Bayesian approach. We introduce two latent ignorability assumptions of the mediator: a latent ignorability assumption of the observed mediator, implying that the observed proportion of friends visiting Palazzo Vecchio are independent of potential outcomes for the number of museum visits conditional on principal stratum membership given the observed value of the encouragement and individual level and network level pre-treatment covariates; and

a latent ignorability assumption of the potential mediator, implying that potential outcomes for the proportion of friends visiting Palazzo Vecchio are independent of potential outcomes for the number of museum visits conditional on principal stratum membership given observed individual level and network level pretreatment covariates. Under latent ignorability assumptions, ignorability of the mediator only holds after conditioning on principal stratum membership, which is a latent variable, only partially observed for each unit. The latent nature of the principal strata complicates inference. Again we deal with inferential issues using a Bayesian approach, which does not require full identification (e.g. Gustafson, 2010). Specifically, we specify flexible parametric models for the principal stratum membership given individual-level covariates and for the potential outcomes conditional on principal stratum membership and individual- and network-level covariates.

Our work provides an innovative approach to the analysis of CEDs, which blends concepts and tools from principal stratification, mediation analysis, and causal inference with network data.

4.2 Experimental design and data

The clustered encouragement field experiment we re-analyze was run in Florence, Italy, at three different points in time during 2014. It involved $J = 15$ classes, indexed by $j = 1, \dots, 15$, from 3 different high schools in the city. Let n_j be the number of students in class j , $j = 1, \dots, 15$, and let $n = \sum_{j=1}^J n_j$ be the total number of students participating in the study: $n = 266$. Finally, let C_{ij} be an indicator for student i 's class membership: $C_{ij} = 1$ if student i is in class j and zero otherwise, $i = 1, \dots, n_j$ and $j = 1, \dots, J$. All high schools were of the same type and all students in the classes participating in the study were aged 17-18. Each of the 15 classes was randomly assigned to one of three incremental encouragements, in groups of five. All students, irrespective of the encouragement their class was assigned to, were offered the opportunity to visit individually the Florentine Museum of Palazzo Vecchio outside of school hours. To preclude the possibility of interference between classes, classes in the same school were located in different buildings or, when this was not possible, on different floors of the same building.

Under the weakest encouragement, which we name *flyer*, students received a flyer containing basic information about Palazzo Vecchio, and a short text written by

the experimenters briefly stating the importance of museum attendance. Under the intermediate encouragement, named *presentation*, in addition to the flyer and text, students received a short presentation conducted by an art expert from the museum: the aim of the presentation was to enhance students' curiosity about museum visits, in general, and Palazzo Vecchio, in particular, and to portray museum attendance as an intriguing and entertaining experience. Under the strongest encouragement, which we name *reward*, in addition to the flyer, text and presentation, students were promised a non-financial reward in the form of extra-credit points towards their final school grade if they visited Palazzo Vecchio within two months after assignment. Let $Z_j \in \mathcal{Z} \equiv \{1, 2, 3\}$ denote the encouragement to which class j is assigned: $Z_j=1$ if the class is assigned to the *flyer*, $Z_j=2$ if it is assigned to the *presentation* and $Z_j=3$ if it is assigned to the *reward*.

The field experiment was implemented through three visits to the classes. At time $t = 1$, all students were surveyed about their background characteristics and interviewed about their friendship network within the class. In particular, they were asked to mention all the classmates they consider to be their friends. In light of the location of the classes involved in the experiment, we believe that asking students to name friends from the same school, rather than the same class, would have produced the same result. After this baseline data collection, each class was randomly assigned to one of the three encouragement levels described above.

At time $t = 2$, two months after the assignment, information about whether students had visited or not Palazzo Vecchio was collected. Let $D_{ij} \in \mathcal{D} \equiv \{0, 1\}$ be a variable equal to 1 if student i in class j visits Palazzo Vecchio in the two months after the assignment, and zero otherwise. The variable D_{ij} is the treatment of interest.

A follow-up period of 6 months, starting from time $t = 2$ was considered. At the end of the study, that is, at time $t = 3$, information on the number of visits each student made to other museums in the follow-up period was collected. Let $Y_{ij} \in \mathcal{Y} \equiv \{0, 1, \dots, 18\}$ be a count variable indicating the number of museum visits made by student i in class j during the follow-up period: Y_{ij} is the outcome of primary interest. See Lattarulo et al. (2017) and Forastiere, Lattarulo, et al. (2021) for a more detailed description of the three forms of encouragement and the substantive results of the field experiment.

4.2.1 Descriptive statistics

Table 4.1 reports some descriptive statistics for the sample of the 266 students, grouped by class assignment. In small-scale randomized experiments, baseline char-

TABLE 4.1: The museum study: descriptive statistics of the individual and friends' level baseline covariates (proportions or means)

	<i>Flyer</i>	<i>Presentation</i>	<i>Reward</i>	<i>Overall</i>
<i>INDIVIDUAL BASELINE CHARACTERISTICS</i>				
Male (1/0)	0.19	0.29	0.53	0.34
Already visited Palazzo Vecchio (1/0)	0.66	0.75	0.72	0.71
No. of museums visited previous year	3.31	4.75	3.54	3.86
GPA (0-10) ^a	6.73	6.82	6.98	6.84
Interest in sciences (1-3) ^b	2.12	1.74	2.23	2.03
Parental education (1-5) ^c	3.26	3.54	3.56	3.45
No. of friends (individual degree)	4.94	4.00	3.07	4.00
<i>FRIENDS' BASELINE CHARACTERISTICS</i>				
Already visited Palazzo Vecchio (1/0)	0.62	0.74	0.69	0.68
No. of museums visited previous year	3.07	4.59	3.20	3.61
No. of friends of friends (friends' degree)	5.62	4.81	3.43	4.61
No. of classes	5	5	5	15
No. of students per class	17.8	17.4	18	17.7
No. of students	89	87	90	266

Note. Estimates are based on sampling-theory estimation of a ratio for cluster sampling (Cochran, 1963, pag. 30).

^a Grade Point Average (GPA) is continuous on a 1-10 point scale.

^b Levels of interest in sciences: 1 if the interest is only in humanities, 2 if it is in both humanities and sciences, 3 if it is only in sciences.

^c Levels of parental education: 1 if both parents completed only compulsory education, 2 if only one of them completed high school, 3 if both completed high school, 4 if only one of them completed university, 5 if both parents completed university.

acteristics are not always well-balanced across assignment groups. Here, imbalances arise in almost all the individual-level and network(friends)-level baseline characteristics. In particular, the number of museums visited the previous year and the proportion of males differ remarkably across students in classes exposed to alternative encouragement levels. Students assigned to the presentation visited more museums, on average, than those assigned to the other encouragements; not surprisingly, they are also more interested in the humanities. The proportion of males is, instead, much higher among students assigned to the reward. Imbalances also arise in the structure of friendship networks. Specifically, the average number of

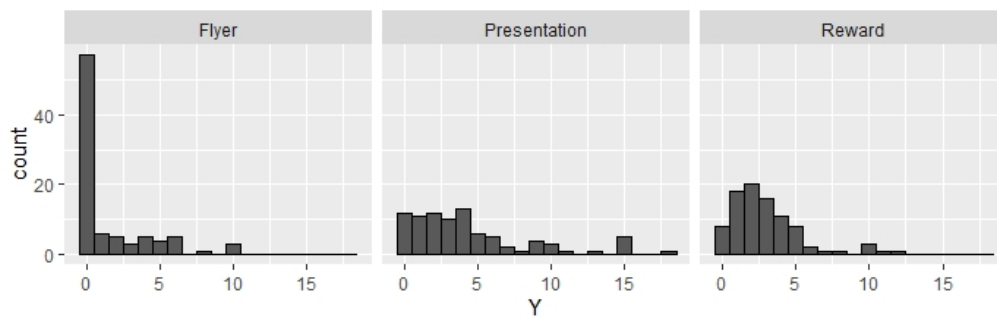
friends is about 5, 4 and 3 in the flyer, presentation and reward group, respectively. These differences result in a different number of friends of friends across encouragement groups, too: students assigned to the reward tend to have fewer friends – and fewer friends of friends – than those assigned to the presentation and, to a greater extent, than those assigned to the flyer.

Table 4.2 reports descriptive statistics regarding treatment uptake (Palazzo Vecchio visit) and final outcome (number of museum visits during the follow-up), again at both individual/student and network/student’s friends level. In addition, Figure 4.1 shows the bar charts of the final outcome by encouragement status. The

TABLE 4.2: The museum study: descriptive statistics of the treatment and the outcome variables

	<i>Flyer</i>	<i>Presentation</i>	<i>Reward</i>	<i>Overall</i>
<i>INDIVIDUAL TREATMENT AND OUTCOME</i>				
No. of students performing the proposed visit	3	10	40	53
Share of students performing the proposed visit	0.03	0.11	0.44	0.20
Count of museum visits during the follow-up	1.49	4.39	3.00	2.95
Students visiting at least one museum during the follow-up	0.36	0.86	0.91	0.71
<i>FRIENDS’ TREATMENT AND OUTCOME</i>				
Share of friends performing the proposed visit	0.03	0.10	0.41	0.18
Count of museum visits during the follow-up	1.57	4.02	2.86	2.80
Friends visiting at least one museum during the follow-up	0.56	0.98	0.92	0.82

FIGURE 4.1: Outcome bar charts



percentage of students who visit Palazzo Vecchio after receiving the flyer is only 3%. Among students who also attend the presentation made by the art expert, this percentage increases to 11%. A sharper increase is observed when the reward is also promised in the event of a visit to Palazzo Vecchio: in this case, in fact, 44% of students performed the proposed visit. However, it is worth noting that

the interest of the field experiment is not on investigating which incentive could better motivate students to perform a particular museum visit once, but on assessing which incentive is more effective to prompt teens' museums attendance in general. Simple comparisons of the outcomes by encouragement status suggest that the flyer is insufficient to motivate students to visit art museums in the follow-up period, just as it is insufficient to motivate students to visit Palazzo Vecchio in the first place. As we can see in Figure 4.1, the empirical distributions of the number of museum visits during the follow up by encouragement are left-skewed and have a mass at zero. The proportion of students who do not visit any museum is especially high under the flyer encouragement. The number of museum visits has a similar distribution under the presentation and reward encouragements, which seem to be more effective than the flyer encouragement. The average number of total museum visits is relatively high under both the presentation and the reward (4.39 and 3.00, respectively). Therefore, there is some evidence that these two types of encouragement may have at least partially achieved the purpose of motivating students to visit art museums.

4.3 Methodology

4.3.1 Network and neighborhood treatment

Let us denote students' friendship network as a pair $(\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes (subjects) and $\mathcal{E} = \{E_{ikj}, i = 1, \dots, n_j; k = 1, \dots, n_j, k \neq i; j = 1, \dots, J\}$ is the set of edges (friendship ties) that links the subjects; in particular, $E_{ikj} = E_{kij} = 1$ if subject i in class j declared classmate k as a friend and/or viceversa, and $E_{ikj} = E_{kij} = 0$ otherwise. It is worth noting that, by design, $E_{ikj} = E_{kij'} = 0$ for each pair of students i and k belonging to two different classes, j and j' . Thus, we are considering a binary and undirected network. "Binary" because we assign equal weight to all the friendships, guided by the willingness not to introduce speculation about whether or not there are sources more influential than others. "Undirected" because we do believe that an influence between subject i and subject k is likely even if only one of them declared that the other is a friend; in fact, that would at least mean that they interact often and are in close contact, more than with the rest of the classmates who have not been declared friends by any of them.

For each student i in class j , let $\mathcal{N}_{ij}$ denote the neighborhood of student i , comprising student i 's friends in class j . Let us also denote with $|\mathcal{N}_{ij}|$ the cardinality of $\mathcal{N}_{ij}$,

i.e., the number of ties (friends) that student has. Spillover effects may arise from the treatment friends receive. We indicate with $\mathbf{D}_{\mathcal{N}_{ij}}$ the neighborhood treatment vector, i.e., the $|\mathcal{N}_{ij}|$ –dimensional vector with elements equal to $D_{i'j}$, $i' \neq i \in \mathcal{N}_{ij}$. In a recent strand of literature on causal inference in the presence of networks, it is quite diffuse the practice of using a mapping function $m(\cdot)$ to summarize the treatment received by units in the individual network, from which spillover effects may arise (e.g., Del Prete et al., 2020; Forastiere, Airoidi, et al., 2021; Forastiere et al., 2018). Here, we denote with $S_{\mathcal{N}_{ij}}$ the variable resulting from applying this function to the neighborhood treatment vector, i.e., $S_{\mathcal{N}_{ij}} = m(\mathbf{D}_{\mathcal{N}_{ij}})$. For each student i in class j , $i = 1 \dots, n_j$, $j = 1, \dots, J$, we define $S_{\mathcal{N}_{ij}}$ as the proportion of student i 's friends who visit Palazzo Vecchio:

$$S_{\mathcal{N}_{ij}} = \frac{1}{|\mathcal{N}_{ij}|} \sum_{i' \neq i \in \mathcal{N}_{ij}} D_{i'j}$$

Any function $m(\mathbf{D}_{\mathcal{N}_{ij}})$ can be chosen, but it is worth mentioning that it must be assumed to be well-specified. Therefore, it is important to choose a function that is deemed to be relevant, either on an intuitive basis or on the basis of knowledge of the subject matter, for the specific application at hand. Here, we believe the proportion of friends visiting Palazzo Vecchio is influential because, if this proportion is high, it is implausible that the student, during his/her daily interactions with friends, would not receive any feedback on the Palazzo Vecchio experience. A future extension of this work might be to assess whether spillover passes from a different function of the neighborhood treatment vector.

For each student i in class j , $i = 1 \dots, n_j$, $j = 1, \dots, J$, we observe a vector of K covariates, $\mathbf{X}_{ij}$. This vector includes both individual and neighborhood-level covariates. The latter are of different types: some are a synthesis of individual-level covariates in the neighborhood, such as the average number of museum visits made by friends prior to the experiment; others represent the structure of the neighborhood, such as the number of ties a unit has.

4.3.2 Potential outcomes

Let $\mathbf{z}$ denote a J –dimensional vector of the encouragement assignments for the whole sample of classes with j th element equal to z_j : $\mathbf{z} \in \{1, 2, 3\}^J$. For each student i in class j , $i = 1 \dots, n_j$, $j = 1, \dots, J$, let $D_{ij}(\mathbf{z})$ denote the potential indicator for visiting Palazzo Vecchio given the encouragement vector $\mathbf{z}$; also let $Y_{ij}(\mathbf{z})$ denote

the potential outcome of main interest, that is, the number of museums visited in the follow-up period given the encouragement vector $\mathbf{z}$. Because $S_{\mathcal{N}_{ij}}$ is a function of $D_{i'j}$, $i' \neq i \in \mathcal{N}_{ij}$, it is a post-encouragement variable, and thus, we introduce potential outcomes for it, too:

$$S_{\mathcal{N}_{ij}}(\mathbf{z}) = \frac{1}{|\mathcal{N}_{ij}|} \sum_{i' \neq i \in \mathcal{N}_{ij}} D_{i'j}(\mathbf{z}).$$

We assume that a cluster-level Stable Unit Value Assumption (SUTVA, Rubin, 1980) holds for the potential outcomes $D_{ij}(\mathbf{z})$, $Y_{ij}(\mathbf{z})$, and $S_{\mathcal{N}_{ij}}(\mathbf{z})$:

Assumption 16 (*SUTVA for $D_{ij}(\mathbf{z})$, $Y_{ij}(\mathbf{z})$, and $S_{\mathcal{N}_{ij}}(\mathbf{z})$*)

No hidden variation of encouragement: For all $\mathbf{z}, \mathbf{z}'$: $\mathbf{z} = \mathbf{z}'$, $D_{ij}(\mathbf{z}) = D_{ij}(\mathbf{z}')$, $Y_{ij}(\mathbf{z}) = Y_{ij}(\mathbf{z}')$, $S_{\mathcal{N}_{ij}}(\mathbf{z}) = S_{\mathcal{N}_{ij}}(\mathbf{z}')$.

Cluster-level no-interference: For all $\mathbf{z}, \mathbf{z}'$: $z_j = z'_j$, $D_{ij}(\mathbf{z}) = D_{ij}(\mathbf{z}')$, $Y_{ij}(\mathbf{z}) = Y_{ij}(\mathbf{z}')$, $S_{\mathcal{N}_{ij}}(\mathbf{z}) = S_{\mathcal{N}_{ij}}(\mathbf{z}')$.

Assumption 16 allows to write $D_{ij}(\mathbf{z})$, $Y_{ij}(\mathbf{z})$ and $S_{\mathcal{N}_{ij}}(\mathbf{z})$ as $D_{ij}(z_j)$, $Y_{ij}(z_j)$ and $S_{\mathcal{N}_{ij}}(z_j) = |\mathcal{N}_{ij}|^{-1} \sum_{i' \neq i \in \mathcal{N}_{ij}} D_{i'j}(z_j)$. Assumption 16 implies that the outcomes of a given student does not vary with the encouragement assigned to students of other classes. Under SUTVA, $D_{ij} = D_{ij}(Z_j)$, $S_{ij} = S_{ij}(Z_j)$, and $Y_{ij} = Y_{ij}(Z_j)$.

We investigate whether spillover effects pass through the variable $S_{\mathcal{N}_{ij}}$. We formalize this idea by viewing it as a mediator. For $j = 1, \dots, J$, let $\mathbf{S}_j$ and $\mathbf{S}_j(z_j)$ denote n_j -dimensional vectors with i th element equal to $S_{\mathcal{N}_{ij}}$ and $S_{\mathcal{N}_{ij}}(z_j)$, respectively, $i = 1 \dots, n_j$. For each student i , we assume the existence of potential outcomes of the form $Y_{ij}(z_j, \mathbf{s}_j)$ and $Y_{ij}(z_j, \mathbf{S}_j(z'_j))$: $Y_{ij}(z_j, \mathbf{s}_j)$ would be the value of the outcome Y if, possibly contrary to fact, the encouragement were set to the level z_j and the mediator vector, $\mathbf{S}_j$, were set to a specific prefixed value, $\mathbf{s}_j$; and $Y_{ij}(z_j, \mathbf{S}_j(z'_j))$ would be the value of the outcome Y if, possibly contrary to fact, the encouragement were set to the level z_j and the mediator vector, $\mathbf{S}_j$, were set to the value it would have taken if the encouragement had been set to an alternative level, z'_j .

We believe to be plausible that a cluster version of the *Stable Unit Treatment on Neighborhood Value Assumption* (SUTNVA) introduced by Forastiere, Airolidi, et al. (2021) holds for potential outcomes of the form $Y_{ij}(z_j, \mathbf{s}_j)$ and $Y_{ij}(z_j, \mathbf{S}_j(z'_j))$. Formally,

Assumption 17 (*Cluster SUTNVA*)

No hidden variation of mediators:

For all $\mathbf{s}_j, \mathbf{s}'_j$: $\mathbf{s}_j = \mathbf{s}'_j$, $Y_{ij}(z_j, \mathbf{s}_j) = Y_{ij}(z_j, \mathbf{s}'_j)$.

For all $\mathbf{S}_j(z'_j), \mathbf{S}_j(z''_j)$: $\mathbf{S}_j(z'_j) = \mathbf{S}_j(z''_j)$, $Y_{ij}(z_j, \mathbf{S}_j(z'_j)) = Y_{ij}(z_j, \mathbf{S}_j(z''_j))$.

Cluster Neighborhood Interference:

For all $\mathbf{s}_j, \mathbf{s}'_j$: $\mathbf{s}_{\mathcal{N}_{ij}} = \mathbf{s}'_{\mathcal{N}_{ij}}$, $Y_{ij}(z_j, \mathbf{s}_j) = Y_{ij}(z_j, \mathbf{s}'_j)$.

For all $\mathbf{S}_j(z'_j), \mathbf{S}_j(z''_j)$: $\mathbf{S}_{\mathcal{N}_{ij}}(z'_j) = \mathbf{S}_{\mathcal{N}_{ij}}(z''_j)$, $Y_{ij}(z_j, \mathbf{S}_j(z'_j)) = Y_{ij}(z_j, \mathbf{S}_j(z''_j))$.

On the one hand, the *no hidden variation of mediators* component of cluster SUTNVA implies that potential outcomes of the form $Y_{ij}(z_j, \mathbf{s}_j)$ and $Y_{ij}(z_j, \mathbf{S}_j(z'_j))$ are identical as long as the value of the mediator vector is the same, i.e., as long as the proportion of friends visiting Palazzo Vecchio, for students in class j , is the same. The validity of this assumption is strongly related to the validity of the assumption that all friends of a student can influence his/her future museums attendance to the same extent. That is, if the influence of a friend were stronger than that of another, the assumption would not hold: there would be hidden variation in the proportion of friends visiting Palazzo Vecchio. On the other hand, the *Cluster Neighborhood Interference* component of cluster SUTNVA implies that potential outcomes of the form $Y_{ij}(z_j, \mathbf{s}_j)$ and $Y_{ij}(z_j, \mathbf{S}_j(z'_j))$ for student i in class j depend on proportion of student i 's friends who visit Palazzo Vecchio, but do not vary with the proportion of friends of other classmates who visit Palazzo Vecchio. Under cluster SUTNVA (Assumption 17), we can write $Y_{ij}(z_j, \mathbf{s}_j)$ and $Y_{ij}(z_j, \mathbf{S}_j(z'_j))$ as $Y_{ij}(z_j, \mathbf{s}_{\mathcal{N}_{ij}})$ and $Y_{ij}(z_j, \mathbf{S}_{\mathcal{N}_{ij}}(z'_j))$, respectively. Note that, under Assumptions 16 and 17, $Y_{ij} = Y_{ij}(Z_j) = Y_{ij}(Z_j, \mathbf{S}_{\mathcal{N}_{ij}}(Z_j))$.

It is worth noting that we have to know the network that links the units to exploit Assumptions 16 and 17, which are based on the idea that interference can take place between immediate neighbors. Moreover, we must assume that this network is fixed and that the links between nodes are correctly measured. In our setting this assumption means that any change in the vector of assignment $\mathbf{z}$ and in $\{\mathbf{S}_j(z_j)\}_{j=1}^J$ does not modify the network that links the subjects.

In the causal inference literature, potential outcomes of the form $Y_{ij}(z_j, \mathbf{s}_{\mathcal{N}_{ij}})$ and $Y_{ij}(z_j, \mathbf{S}_{\mathcal{N}_{ij}}(z'_j))$ are also called *a priori* counterfactual potential outcomes (e.g., Frangakis & Rubin, 2002), because some of them can never be observed for some units.

For instance, $Y_{ij}(z_j, S_{\mathcal{N}_{ij}}(z'_j))$ cannot be observed for any units with $S_{\mathcal{N}_{ij}}(z'_j) \neq S_{\mathcal{N}_{ij}}(z_j)$.

In the sequel, we denote by $D_{ij}(z)$, $S_{\mathcal{N}_{ij}}(z)$ and $Y_{ij}(z)$ potential outcomes for D_{ij} , $S_{\mathcal{N}_{ij}}$ and Y_{ij} for a student i in class j under class encouragement z . Similarly, we use the notation $Y_{ij}(z, s_{\mathcal{N}_{ij}})$ and $Y_{ij}(z, S_{\mathcal{N}_{ij}}(z'))$ for class encouragements z and z' .

4.3.3 Principal strata

In clustered encouragement designs with individual non-compliance, we are often interested in causal effects defined accounting for the compliance behavior. We deal with this issue using the PS framework. Accordingly, we classify students into principal strata defined by their compliance behavior. Because the encouragement is a three-valued variable, $Z_j \in \{1, 2, 3\}$, and the intermediate variable is a binary indicator for the Palazzo Vecchio visit, $D_{ij} \in \{0, 1\}$, the basic PS with respect to D_{ij} classifies units into eight latent subgroups according to the joint potential values of D_{ij} under each encouragement level: $D_{ij}(1)$, $D_{ij}(2)$ and $D_{ij}(3)$. More formally, let $G_{ij} = (D_{ij}(1), D_{ij}(2), D_{ij}(3))$, then the basic principal strata are given by the possible values of G_{ij} :

$$(d_1, d_2, d_3) := \{ij : D_{ij}(1) = d_1, D_{ij}(2) = d_2, D_{ij}(3) = d_3\} \text{ with } d_1, d_2, d_3 \in \{0, 1\}.$$

We relabel the possible values of G_{ij} as $d_1d_2d_3$, so $G_{ij} \in \{111, 011, 101, 110, 001, 010, 100, 000\}$, and name the principal strata as shown in Table 4.3.

TABLE 4.3: Principal strata and their labels

$D_{ij}(1)$	$D_{ij}(2)$	$D_{ij}(3)$	<i>Principal Stratum</i>	<i>Stratum Label</i>
1	1	1	$G = 111$	<i>Always Takers</i>
0	1	1	$G = 011$	<i>Presentation Compliers</i>
0	0	1	$G = 001$	<i>Reward Compliers</i>
0	0	0	$G = 000$	<i>Never Takers</i>
1	0	0	$G = 100$	<i>Defiers</i>
0	1	0	$G = 010$	<i>Defiers</i>
1	1	0	$G = 110$	<i>Defiers</i>
1	0	1	$G = 101$	<i>Defiers</i>

As we can see in Table 4.3, we refer to the union of principal strata 100, 010, 110, 101 as *Defiers*. They are students who would visit Palazzo Vecchio under the flyer and/or

presentation condition(s), but would not visit the museum under a stronger encouragement (presentation and/or reward). In our application study, similar behaviors are implausible: if the flyer and/or the presentation is/are enough to motivate a student to visit Palazzo Vecchio, we see no reason why receiving additional motivators should discourage the same visit. Therefore, we rule out the presence of Defiers by assuming Monotonicity. Formally,

Assumption 18 (*Monotonicity*). For all students i in each class j ,

$$D_{ij}(1) \leq D_{ij}(2) \leq D_{ij}(3).$$

Under this assumption, $G_{ij} \in \{111, 011, 001, 000\}$ for all ij , that is, each student ij can belong to one of the following four groups: *Always Takers* (AT), *Presentation Compliers* (PC), *Reward Compliers* (RC) and *Never Takers* (NT). *Always Takers* and *Never Takers* are students who, respectively, would perform, and would not perform the proposed visit to Palazzo Vecchio regardless of the encouragement. *Presentation Compliers* are students who would visit Palazzo Vecchio if they received the reward or at least attended the presentation. *Reward Compliers* are students who would visit Palazzo Vecchio only upon receiving the reward.

4.3.4 Overall Principal Causal Effects

The main advantage of the PS approach is that it allows for the proper definition of stratum-specific causal effects, named Principal Causal Effects (PCEs). Specifically, we focus on finite-sample average principal causal effects. Overall principal causal effects of an encouragement condition, say z , versus another, say z' , are defined as follows:

$$\begin{aligned} PCE(z \text{ vs } z')^{(g)} &= \frac{\sum_{ij:G_{ij}=g} Y_{ij}(z) - Y_{ij}(z')}{n_g} \\ &= \frac{\sum_{ij:G_{ij}=g} Y_{ij}(z, S_{N_{ij}}(z)) - Y_{ij}(z', S_{N_{ij}}(z'))}{n_g} \end{aligned} \quad (4.1)$$

where n_g is the total number of subjects in the sample belonging to the principal stratum $g \in \{000, 001, 011, 111\}$.

For Presentation and Reward Compliers, overall principal causal effects are *associative*, in the sense that they measure effects on the outcome that are associative with

effects on the intermediate variable, D_{ij} , combining three effects: the *pure encouragement effect*, the *experience effect*, and the *spillover effect* (Forastiere, Lattarulo, et al., 2021; Frangakis & Rubin, 2002). Pure encouragement effects measure changes in the individual primary outcome, i.e., the number of museum visits in the follow-up period, directly induced by the encouragement. Experience effects measure changes in the individual outcome that originate from individual treatment uptake, i.e., from visiting Palazzo Vecchio first-hand. Finally, spillover effects measure changes in the individual outcome ascribable to mechanisms of interference sparked by friends performing the encouraged visit to Palazzo Vecchio in the first place. Spillover effects can be interpreted as channeled effects: they measure the part of the encouragement effects mediated through the proportion of friends visiting Palazzo Vecchio.

For Never Takers and Always Takers, overall principal causal effects are dissociative (Frangakis & Rubin, 2002): because the visit to Palazzo Vecchio is – by definition – not affected by the encouragement for Never Takers and Always Takers, it follows that any change in the individual outcome between the encouragement conditions being compared may be “only” ascribed to pure encouragement and spillover effects (Forastiere, Lattarulo, et al., 2021).

In Table 4.4, for each principal stratum, we resume the different causal pathways involved in the *PCEs*. As we can see in Table 4.4, we interpret the overall principal

TABLE 4.4: Causal mechanisms underlying Principal Causal Effects

<i>Principal Stratum</i>	$PCE(3 \text{ vs } 2)^{(g)}$	$PCE(2 \text{ vs } 1)^{(g)}$	$PCE(3 \text{ vs } 1)^{(g)}$
Never Takers	Spill	Enc, Spill	Enc, Spill
Always Takers	Spill	Enc, Spill	Enc, Spill
Presentation Compliers	Spill	Enc, Exp, Spill	Enc, Exp, Spill
Reward Compliers	Exp, Spill	Enc, Spill	Enc, Exp, Spill

Note. “Enc”, “Exp” and “Spill” stand for, respectively, pure encouragement effect, experience effect and spillover effect.

causal effects of reward ($z = 3$) versus presentation ($z = 2$) for Never Takers, Always Takers and Presentation Compliers as spillover effects, implicitly assuming that there do not exist pure encouragement and experience effects for these types of subjects. Indeed, for Never Takers ($G_{ij} = 000$), Always Takers ($G_{ij} = 111$), and Presentation Compliers ($G_{ij} = 011$), $D_{ij}(2) = D_{ij}(3)$, and thus, experience effects do not exist by definition. Moreover, for these types of students, we can reasonably

assume that there exists no pure encouragement effect of reward ($z = 3$) versus presentation ($z = 2$). Under assignment to the reward, students get the reward only if they visit Palazzo Vecchio within two months after the assignment; they do not get any reward if they visit art museums in the follow-up period but do not visit Palazzo Vecchio in the first place. Therefore, in terms of pure encouragement effects, we can argue that the reward adds nothing with respect to presentation.

4.3.5 Disentangling Spillovers from overall Principal Causal Effects Through Mediation

This section describes the core of our novel methodological proposal: using principal stratum mediation analysis to disentangle spillovers from the overall PCEs.

In principal stratum mediation analysis, the focus is on disentangling the part of each principal causal effect that works through another variable, a mediator – *indirect principal causal effect*, from the causal effects of the assignment/treatment on the response net of the effect of the mediator – *direct principal causal effect*. In our application study, we conduct principal stratum mediation analysis using the proportion of friends who visited Palazzo Vecchio as a mediator. Therefore, indirect principal causal effects measure changes in the response that can be ascribed to a variation in the summary of friends' compliance behavior by principal strata, and thus, they are interpretable as spillover effects. In other words, our analysis is aimed at quantifying, for each principal stratum, the relative importance of direct, $Z_j \rightarrow Y_{ij}$, and indirect, $Z_j \rightarrow S_{\mathcal{N}_{ij}} \rightarrow Y_{ij}$, causal pathways in the overall intervention effect. From these definitions, it is clear that indirect principal causal effects measure, for each principal stratum, spillovers passing through the selected mediator. Therefore, conclusions about whether spillover effects are present and their extent crucially depend on the definition of $S_{\mathcal{N}_{ij}}$.

Formally, we introduce the concepts of principal natural direct and indirect effects and principal control direct effects:

$$NDE(z \text{ vs } z', S_{\mathcal{N}_{ij}}(z''))^{(g)} = \frac{\sum_{ij: G_{ij}=g} [Y_{ij}(z, S_{\mathcal{N}_{ij}}(z'')) - Y_{ij}(z', S_{\mathcal{N}_{ij}}(z'')))]}{n_g}, \quad (4.2)$$

$$NIE(z'', S_{\mathcal{N}_{ij}}(z) \text{ vs } S_{\mathcal{N}_{ij}}(z'))^{(g)} = \frac{\sum_{ij: G_{ij}=g} [Y_{ij}(z'', S_{\mathcal{N}_{ij}}(z)) - Y_{ij}(z'', S_{\mathcal{N}_{ij}}(z')))]}{n_g}, \quad (4.3)$$

$$CDE(z \text{ vs } z', s_{\mathcal{N}_{ij}}^*)^{(g)} = \frac{\sum_{ij: G_{ij}=g} [Y_{ij}(z, s_{\mathcal{N}_{ij}}^*) - Y_{ij}(z', s_{\mathcal{N}_{ij}}^*)]}{n_g}, \quad (4.4)$$

for $z, z', z'' \in \{1, 2, 3\}$ with $z \neq z'$ and z'' set to z or to z' , and $s_{\mathcal{N}_{ij}}^* \in [0, 1]$ is a value of the mediator reasonably chosen by the researcher and not necessarily observed in the sample. In the application, we calculate these estimands for $z > z'$.

Principal natural direct effects measure average local effects (that is, for students of type g) of z versus z' on the response, the number of visits to art museums, not mediated through $S_{\mathcal{N}_{ij}}$, the proportion of friends visiting Palazzo Vecchio; in contrast, they consider the effect of z versus z' on the response intervening by fixing the proportion of friends visiting Palazzo Vecchio to the value it would have taken if the encouragement had been set to z'' . Principal natural direct effects embrace all causal channels other than spillovers, i.e., the pure encouragement effect and, where appropriate, also experience effects.

Principal natural indirect effects fix the encouragement to the value z'' and then measure average local effects (that is, for students of type g) on the response, the number of visits to art museums, mediated through the proportion of friends visiting Palazzo Vecchio, $S_{\mathcal{N}_{ij}}$; that is, setting $S_{\mathcal{N}_{ij}}$ to what it would have been if the encouragement were z , $S_{\mathcal{N}_{ij}}(z)$, in contrast to what it would have been if the encouragement were z' , $S_{\mathcal{N}_{ij}}(z')$. Principal natural indirect effects provide information on spillover effects.

Principal controlled direct effects measure local effects on the response not mediated through the proportion of friends visiting Palazzo Vecchio, as principal natural direct effects, but they compare potential outcomes for students belonging to the same principal stratum under two encouragements, z and z' , setting the mediator, $S_{\mathcal{N}_{ij}}$, to some pre-fixed value $s_{\mathcal{N}_{ij}}^*$.

It is worth noting that “Natural Direct Effects” (NDE), “Natural Indirect Effects” (NIE), and “Controlled Direct Effects” (CDE), defined for the whole population, rather than for latent sub-populations such as the principal strata, are the causal estimands usually considered in mediation analysis (e.g., Pearl, 2001; Robins & Greenland, 1992). Moreover, as in the classical mediation setting, overall average causal effects can be decomposed into the sum of natural direct and indirect effects, each average overall principal causal effect can be decomposed into the sum of a principal natural direct effect and a principal natural indirect effect:

$$\begin{aligned} PCE(z \text{ vs } z')^{(g)} &= NDE(z \text{ vs } z', S_{\mathcal{N}_{ij}}(z))^{(g)} + NIE(z', S_{\mathcal{N}_{ij}}(z') \text{ vs } S_{\mathcal{N}_{ij}}(z))^{(g)} \\ &= NDE(z' \text{ vs } z, S_{\mathcal{N}_{ij}}(z'))^{(g)} + NIE(z, S_{\mathcal{N}_{ij}}(z) \text{ vs } S_{\mathcal{N}_{ij}}(z'))^{(g)}. \end{aligned} \quad (4.5)$$

We also introduce new causal estimands, which we call principal controlled spillover effects, defined as follows:

$$CSE(z, s_{\mathcal{N}_{ij}}^{***} \text{ vs } s_{\mathcal{N}_{ij}}^{**})^{(g)} = \frac{\sum_{ij: G_{ij}=g} [Y_{ij}(z, s_{\mathcal{N}_{ij}}^{***}) - Y_{ij}(z, s_{\mathcal{N}_{ij}}^{**})]}{n_g} \quad (4.6)$$

where $z \in \{1, 2, 3\}$ and $s_{\mathcal{N}_{ij}}^{**}$ and $s_{\mathcal{N}_{ij}}^{***} \in [0, 1]$ are values of the proportion of friends visiting Palazzo Vecchio. Principal controlled spillover effects measure local effects on the response for students belonging to the same principal stratum and exposed to the encouragement level z , due to a change in the proportion of friends visiting Palazzo Vecchio from a pre-fixed value $s_{\mathcal{N}_{ij}}^{***}$ to another pre-fixed value $s_{\mathcal{N}_{ij}}^{**}$. Therefore, principal controlled spillover effects provide information on spillover effects.

Principal controlled direct and spillover effects are useful for simulating a controlled experimental setting in which it would be possible to arbitrarily set the proportion of friends visiting Palazzo Vecchio to a value of interest (e.g., zero friends visiting Palazzo Vecchio), or to vary it between two – possibly unobserved – values of interest (e.g., from all friends visiting Palazzo Vecchio to none). Although in a real intervention it is not possible to force a certain proportion of friends to visit the museum, these principal controlled effects are nonetheless interesting for answering substantive questions about causal mechanisms that would be impossible to answer by estimating principal natural effects only.

Interpretation of principal NDEs, NIEs, CDEs and CSEs

As discussed in Section 4.3.4, experience effects cannot exist for Always Takers and Never Takers. Therefore, for these two principal strata, principal natural and controlled direct effects provide information on pure encouragement effects, and principal natural indirect effects and principal controlled spillover effects provide information on spillover effects.

For Presentation Compliers and Reward Compliers, in general, there exist experience effects, and thus, overall principal causal effects for these principal strata combine encouragement, experience and spillover effects. Therefore, for Presentation Compliers and Reward Compliers, principal natural indirect effects and principal controlled spillover effects are interpretable as spillover effects, and principal natural and controlled direct effects entangle pure encouragement and experience effects. Table 4.5 summarizes the information on the causal pathways provided by principal natural and controlled direct effects. Note that there is no encouragement and ex-

TABLE 4.5: Principal natural direct effects and principal controlled direct effects: an interpretation

<i>Principal Stratum</i> g	$NDE(3 \text{ vs } 2, S_{N_{ij}}(z''))^{(g)}$ $CDE(3 \text{ vs } 2, s_{N_{ij}}^*)^{(g)}$	$NDE(2 \text{ vs } 1, S_{N_{ij}}(z''))^{(g)}$ $CDE(2 \text{ vs } 1, s_{N_{ij}}^*)^{(g)}$	$NDE(3 \text{ vs } 1, S_{N_{ij}}(z''))^{(g)}$ $CDE(3 \text{ vs } 1, s_{N_{ij}}^*)^{(g)}$
Never Takers	0	Enc	Enc
Always Takers	0	Enc	Enc
Presentation Compliers	0	Enc, Exp	Enc, Exp
Reward Compliers	Exp	Enc	Enc, Exp

Note. “Enc” and “Exp” stand for pure encouragement effect experience effect, respectively.

perience effect for Never Takers, Always Takers and Presentation Compliers when encouragement levels 2 and 3 are compared, and thus, $NDE(3 \text{ vs } 2, S_{N_{ij}}(z''))^{(g)}$ and $CDE(3 \text{ vs } 2, s_{N_{ij}}^*)^{(g)}$ for $g \in \{000, 111, 011\}$ are zero.

4.3.6 Identification issues

Our study is a CED, and thus, the encouragement is randomly assigned by design, implying that is independent of all the potential outcomes and covariates. Formally,

Assumption 19 (*Randomization of the cluster-level encouragement*)

$Z_j \perp\!\!\!\perp \{\{D_{ij}(z)\}_z, \{S_{N_{ij}}(z)\}_z, \{Y_{ij}(z, s_{N_{ij}})\}_{z, s_{N_{ij}}}, \mathbf{X}_{ij}\}$ for $z \in \{1, 2, 3\}$ and $s_{N_{ij}} \in [0, 1]$.

Under Monotonicity (Assumption 18), there exist four principal strata and most observed groups defined by the encouragement actually received and the observed value of D_{ij} host mixtures of principal strata as shown in Table 4.6.

TABLE 4.6: Observed groups of students and possible principal strata: NT=000, RC=001, PC=011, AT=111

Z_j	D_{ij}	No.	$\% Z_j = z$	$Mean(Y_{ij})$	$p(Y_{ij} > 0)$	<i>Possible Principal Strata</i>
1	0	86	96.6%	1.35	0.34	PC, RC, NT
1	1	3	3.4%	5.67	1.00	AT
2	0	77	86.5%	4.58	0.87	RC, NT
2	1	10	11.2%	2.90	0.80	AT, PC
3	0	50	55.6%	3.18	0.88	NT
3	1	40	44.4%	2.78	0.95	AT, PC, RC

Randomization guarantees that principal stratum membership has the same distribution in all encouragement arms. Therefore, under Randomization of the cluster-level encouragement and Monotonicity, we can non-parametrically point-identify principal stratum proportions. Let $\pi^{(d_1 d_2 d_3)}$ denote the proportion of units belonging to principal stratum $d_1 d_2 d_3 = \{ij : D_{ij}(1) = d_1, D_{ij}(2) = d_2, D_{ij}(3) = d_3\}$, that is, $\pi^{(d_1 d_2 d_3)} := p(G_{ij} = d_1 d_2 d_3)$. We have $\pi^{(100)} = \pi^{(010)} = \pi^{(110)} = \pi^{(101)} = 0$; $\pi^{(111)} = p(D_{ij} = 1 | Z_j = 1)$; $\pi^{(000)} = p(D_{ij} = 0 | Z_j = 3)$; $\pi^{(001)} = p(D_{ij} = 0 | Z_j = 2) - \pi^{(000)} = p(D_{ij} = 1 | Z_j = 3) - p(D_{ij} = 1 | Z_j = 2)$; and $\pi^{(011)} = p(D_{ij} = 1 | Z_j = 2) - \pi^{(111)} = p(D_{ij} = 1 | Z_j = 3) - \pi^{(111)} - \pi^{(001)}$. The method-of-moments estimates of the principal stratum proportions and also Table 4.6) are as follows (standard errors in parenthesis): $\hat{\pi}^{(111)} = 0.034$ (0.019); $\hat{\pi}^{(000)} = 0.556$ (0.052); $\hat{\pi}^{(001)} = 0.329$ (0.062); $\hat{\pi}^{(011)} = 0.081$ (0.039); suggesting that there are more than 50% Never Takers, and less than 3.5% Always Takers. About 1/3 of students are expected to be Reward Compliers and 8.1% of students are expected to be Presentation Compliers.

Although we can non-parametrically point identify principal stratum proportions, we cannot do the same with overall principal causal effects, i.e., $PCE(z \text{ vs } z')^{(g)}$. In the Florentine study, encouragement was not only aimed at prompting students to visit Palazzo Vecchio, but also to visit museums in general; therefore, we do not believe that ERs would be plausible. Our choice of not imposing additional assumptions,

among which the ER for non-Compliers¹, results in not point identifiable principal causal effects.

Ignorability Assumptions

Inference on principal natural direct and indirect effects (Equations 4.2 and 4.3) and principal controlled direct and spillover effects (Equations 4.4 and 4.6) is even more challenging, because these estimands involve potential outcomes of the form $Y_{ij}(z, s_{\mathcal{N}_{ij}})$ and $Y_{ij}(z, S_{\mathcal{N}_{ij}}(z'))$, which are never observed in this specific experiment for some students. Some structural assumptions that allow us to extrapolate information on this type of potential outcomes from at least potentially observable data are required. Hence, we invoke “latent” ignorability assumptions of the mediator, that is, ignorability assumptions of the mediator conditional on principal stratum membership.

Assumption 20 (*Latent ignorability of the mediator (I)*).

For each unit i in cluster j , for all $z \in \{1, 2, 3\}$ and $s_{\mathcal{N}_{ij}} \in [0, 1]$,

$$Y_{ij}(z, s_{\mathcal{N}_{ij}}) \perp\!\!\!\perp S_{\mathcal{N}_{ij}} \mid Z_j, G_{ij}, \mathbf{X}_{ij};$$

Assumption 21 (*Latent ignorability of the mediator (II)*).

For each unit i in cluster j , for all $z, z' \in \{1, 2, 3\}$ and $s_{\mathcal{N}_{ij}} \in [0, 1]$,

$$Y_{ij}(z, s_{\mathcal{N}_{ij}}) \perp\!\!\!\perp S_{\mathcal{N}_{ij}}(z') \mid G_{ij}, \mathbf{X}_{ij}.$$

In our application study, Assumption 20 means that potential outcomes of the form $Y_{ij}(z, s_{\mathcal{N}_{ij}})$ for future visits to other museums of a student i in class j are independent of the decision to visit Palazzo Vecchio of subject i ’s friends in class j conditional on student i ’s encouragement status, covariates and compliance status. Assumption 21 implies that $Y_{ij}(z, s_{\mathcal{N}_{ij}})$ and the decision to visit Palazzo Vecchio of subject i ’s friends in class j under encouragement z' , $S_{\mathcal{N}_{ij}}(z')$, are independent conditional on student

¹The ER for AT and NT could be formalized as $p(Y_{ij}(z, S_{\mathcal{N}_{ij}}(z)) \mid G_{ij} = g, \mathbf{X}_{ij}) = p(Y_{ij}(z', S_{\mathcal{N}_{ij}}(z')) \mid G_{ij} = g, \mathbf{X}_{ij})$ for $g \in \{000, 111\}$, and implies that $PCE(z \text{ vs } z')^{(g)} = 0$. In particular, for these two principal strata, the natural direct and indirect effects involved in Equation (4.5) would cancel each other out, on average.

i 's covariates and compliance status. Assumptions 20 and 21 are not testable. They would be violated if there existed unobserved variables correlated with both the mediator and outcome despite the conditioning on the assigned treatment, compliance status and observed covariates in case of Assumption 20, or compliance status and observed covariates in case of Assumption 21. The plausibility of these assumptions crucially depends on whether the amount of information provided by the observed covariates is sufficient to establish them. In other words, the presence of unobserved variables that could cause them not to hold is much less likely if $\mathbf{X}_{ij}$ is rich. In the Florentine study, we have observed several pre-treatment variables relative to both the individual and friends' background characteristics, but also to the friendship networks structure (see Table 4.1). We include most of them in $\mathbf{X}_{ij}$, as we will see in Section 4.4.1.

If the compliance status were observed for each unit, Assumption 20 and Assumption 21 would be the standard ignorability assumptions on the mediator usually invoked in the causal mediation literature (see, e.g., Ten Have & Joffe, 2012; VanderWeele & Vansteelandt, 2009, for a review). Therefore, we could identify principal controlled direct and spillover effects under Assumptions 19 and 20 and principal natural direct and indirect effects under Assumptions 19, 20 and 21. Because the compliance status is only partially observed, Assumptions 19, 20 and 21 do not solve identification issues. We deal with inferential issues arising from the latent nature of principal stratum membership using a model-based Bayesian approach (see Section 1.4.7 for a discussion on identification from a Bayesian perspective).

4.4 Analyses

4.4.1 Parametric Models

In our setting, randomization is at cluster-level, and thus, a within-class correlation among individuals may arise from reciprocal influence or from other unobserved common factors. In particular, students belonging to the same class are likely to show a resemblance in terms of both compliance behavior and final outcomes. We account for this systematic unexplained variation within classes by introducing a hierarchical structure, by specifying varying-intercept models for the compliance status and for the outcome.

Principal Strata Model

We specify a multinomial logit random effect model for principal stratum membership conditional on covariates:

$$\log \left(\frac{p(G_{ij} = g | \mathbf{X}_{ij})}{p(G_{ij} = 111 | \mathbf{X}_{ij})} \right) = \gamma^{(g)} + \mathbf{X}_{ij}' \boldsymbol{\delta}^{(g)} + a_j \quad \text{with } a_j \sim N(0, \sigma_a), \quad (4.7)$$

where $g \in \{000, 001, 011\}$ (the stratum 111 is chosen as the reference category) and a_j are the random intercepts. The vector $\mathbf{X}_{ij}$ includes the following individual background characteristics introduced in Table 4.1: *Male, Already visited Palazzo Vecchio, No. of museums visited previous year, GPA, Interest in sciences and Parental education.*

Potential Outcome Models

We specify Zero-Inflated Poisson (ZIP) models for the potential outcomes conditional on principal stratum membership and covariates. A ZIP model assumes that subjects are a mixture of two types. For subjects of the first type, the outcome is always zero. For subjects of the second type, the outcome follows a usual Poisson distribution, which can produce a zero outcome or some other. Let $\phi_{ij}^{(z,g)}$ denote the conditional probability given the covariates that an individual with compliance behavior g under assignment z is of the first type, and let $\mu_{ij}^{(z,g)}$ denote the Poisson mean, which we model using a hierarchical generalized linear model with a log link:

$$\begin{aligned} \log \left(\mu_{ij}^{(z,g)} \right) &= \alpha^{(z,g)} + \beta_S^{(z,g)} S_{\mathcal{N}_{ij}}(z) + \beta_{S\mathcal{N}}^{(z,g)} |\mathcal{N}_{ij}| S_{\mathcal{N}_{ij}}(z) \\ &\quad + \beta_{\mathcal{N}}^{(z,g)} |\mathcal{N}_{ij}| + \mathbf{X}_{ij}' \boldsymbol{\beta}_X^{(z,g)} + b_j^{(z)} \end{aligned} \quad (4.8)$$

with random intercepts $b_j^{(z)} \sim N(0, \sigma_b^{(z)})$. Then,

$$\begin{aligned} p(Y_{ij}(z, S_{\mathcal{N}_{ij}}(z)) = y \mid G_{ij} = g, \mathbf{X}_{ij}) &= \\ &\begin{cases} \phi_{ij}^{(z,g)} + (1 - \phi_{ij}^{(z,g)}) \cdot \text{Pois}(y; \mu_{ij}^{(z,g)}) & \text{if } y = 0 \\ (1 - \phi_{ij}^{(z,g)}) \cdot \text{Pois}(y; \mu_{ij}^{(z,g)}) & \text{if } y > 0 \end{cases} \end{aligned}$$

where $\text{Pois}(y; \mu_{ij}^{(z,g)})$ denotes the mass probability function of the Poisson distribution.

In the ZIP models for the potential outcomes, in addition to the proportion of friends

visiting Palazzo Vecchio under the encouragement assignment z , $S_{\mathcal{N}_{ij}}(z)$, we include the total number of friends, $|\mathcal{N}_{ij}|$, as main effect, and its interaction with $S_{\mathcal{N}_{ij}}(z)$. In fact, we want to account for the possibility that the impact of the proportion of friends visiting Palazzo Vecchio may vary with the total number of friends the student has. For instance, it might make a difference if 5 out of 10 friends or 1 out of 2 visited Palazzo Vecchio, although the value of $S_{\mathcal{N}_{ij}}(z)$ would be unchanged. Also, we condition on the same individual background characteristics we include in the principal strata model except for *Already visited Palazzo Vecchio*, which we expect to affect the decision to visit Palazzo Vecchio but not the decision to visit other museums in the follow-up period. Finally, we condition on a network variable indicating the number of friends' past museums visits, since we can reasonably believe that having museum-goers friends may influence individual museums attendance.

As we discussed in Section 4.3.4, we can reasonably assume that there are no encouragement effects or experience effects for Never Takers, Always Takers and Presentation Compliers in the comparison of potential outcomes under reward ($z = 3$) versus presentation ($z = 2$) (see Table 4.5). This assumption implies that for $g \in \{000, 111, 011\}$, $\phi_{ij}^{(2,g)} = \phi_{ij}^{(3,g)}$ and $\mu_{ij}^{(2,g)} = \mu_{ij}^{(3,g)}$, which, in turn, implies that $\alpha^{(2,g)} = \alpha^{(3,g)}$, $\beta_S^{(2,g)} = \beta_S^{(3,g)}$, $\beta_{SN}^{(2,g)} = \beta_{SN}^{(3,g)}$, $\beta_N^{(2,g)} = \beta_N^{(3,g)}$ and $\beta_X^{(2,g)} = \beta_X^{(3,g)}$.

For the sake of parsimony, we assume that the probabilities of extra zeros with respect to the Poisson distribution, $\phi_{ij}^{(z,g)}$, depend on the encouragement and the principal stratum membership, but not on covariates: $\phi_{ij}^{(z,g)} = \phi^{(z,g)}$. Moreover, we impose the following prior equalities of the slope coefficients and the standard deviations of the random intercepts in the Poisson regressions: $\beta_S^{(z,g)} \equiv \beta_S^{(z)}$, $\beta_{SN}^{(z,g)} \equiv \beta_{SN}^{(z)}$ and $\beta_N^{(z,g)} \equiv \beta_N^{(z)}$ for $g \in \{000, 001, 011, 111\}$ and $z \in \{1, 2\}$; $\beta_X^{(z,g)} \equiv \beta_X$ for all $z \in \{1, 2, 3\}$ and $g \in \{000, 001, 011, 111\}$; and $\sigma_b^{(z)} \equiv \sigma_b$, for all $z \in \{1, 2, 3\}$.

Prior Specifications

We assume that parameters are a priori independent, and specify relatively weakly informative prior distributions for them. Specifically, we postulate independent normal priors with prior mean equal to 0 and standard deviations equal to 2.5 for the coefficients of the principal strata model, $\gamma^{(g)}$ and $\delta^{(g)}$, and independent normal priors with prior mean equal to 0 and standard deviations equal to 1 for the coefficients of the potential outcome models, $\alpha^{(z,g)}$, $\beta_S^{(z,g)}$, $\beta_{SN}^{(z,g)}$, $\beta_N^{(z,g)}$ and β_X . Finally, we specify half-normal prior distributions with a standard deviation equal to

0.5 for the standard deviations σ_a and σ_b of the random intercepts a_j and b_j , entering the principal strata model and the potential outcome models (Gelman, 2006). The choice of the values of the parameters of the prior distributions, in general, and of the standard deviations of the Normal priors was partially driven by stability issues arising from running the algorithm we use to derive the posterior distribution.

4.4.2 Bayesian Inference

Under the assumption that the probabilities of extra zeros with respect to the Poisson distribution, $\phi_{ij}^{(z,g)}$, depend on the encouragement and the principal stratum membership, but not on covariates: $\phi_{ij}^{(z,g)} = \phi^{(z,g)}$, and the prior equalities of the slope coefficients and the standard deviations of the random intercepts in the Poisson regressions described in Section 4.4.1, the full parameter vector is $\boldsymbol{\psi} = \{(\gamma^{(g)}, \boldsymbol{\delta}^{(g)}; g \in \{000, 001, 011\}), \sigma_a, \phi^{(1,000)}, \phi^{(2,000)} = \phi^{(3,000)}, \phi^{(1,001)}, \phi^{(2,001)}, \phi^{(3,001)}, \phi^{(1,011)}, \phi^{(2,011)} = \phi^{(3,011)}, \phi^{(1,111)}, \phi^{(2,111)} = \phi^{(3,111)}, \alpha^{(1,000)}, \alpha^{(2,000)} = \alpha^{(3,000)}, \alpha^{(1,001)}, \alpha^{(2,001)}, \alpha^{(3,001)}, \alpha^{(1,011)}, \alpha^{(2,011)} = \alpha^{(3,011)}, \alpha^{(1,111)}, \alpha^{(2,111)} = \alpha^{(3,111)}, \beta_S^{(1)}, \beta_S^{(2,000)} = \beta_S^{(3,000)} = \beta_S^{(2,001)} = \beta_S^{(2,011)} = \beta_S^{(3,011)} = \beta_S^{(2,111)} = \beta_S^{(3,111)} = \beta_S^{(2)}, \beta_S^{(3,001)}, \beta_{SN}^{(1)}, \beta_{SN}^{(2,000)} = \beta_{SN}^{(3,000)} = \beta_{SN}^{(2,001)} = \beta_{SN}^{(2,011)} = \beta_{SN}^{(3,011)} = \beta_{SN}^{(2,111)} = \beta_{SN}^{(3,111)} = \beta_{SN}^{(2)}, \beta_N^{(1)}, \beta_N^{(2,000)} = \beta_N^{(3,000)} = \beta_N^{(2,001)} = \beta_N^{(2,011)} = \beta_N^{(3,011)} = \beta_N^{(2,111)} = \beta_N^{(3,111)} = \beta_N^{(2)}, \beta_N^{(3,001)}, \boldsymbol{\beta}_X, \sigma_b\}$.

Define $\pi_{ij}^{(g)} \equiv p(G_{ij} = g \mid \mathbf{X}_{ij}, \boldsymbol{\psi})$ and $f_{ij}(y)^{(z,g)} \equiv p(Y_{ij}(z, S_{N_{ij}}(z)) = y \mid G_{ij} = g, \mathbf{X}_{ij}, \boldsymbol{\psi})$. The observed data likelihood results in the following finite mixture model

likelihood:

$$\begin{aligned}
\mathcal{L}_{obs}(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{C}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) = & \prod_{j:Z_j=1} \left[\prod_{i:C_{ij}=1, D_{ij}=0} \left[\pi_{ij}^{(011)} \cdot f_{ij}(Y_{ij})^{(1,011)} + \pi_{ij}^{(001)} \cdot f_{ij}(Y_{ij})^{(1,001)} + \pi_{ij}^{(000)} \cdot f_{ij}(Y_{ij})^{(1,000)} \right] \times \right. \\
& \left. \prod_{i:C_{ij}=1, D_{ij}=1} \left[\pi_{ij}^{(111)} \cdot f_{ij}(Y_{ij})^{(1,111)} \right] \right] \times \\
& \prod_{j:Z_j=2} \left[\prod_{i:C_{ij}=1, D_{ij}=0} \left[\pi_{ij}^{(001)} \cdot f_{ij}(Y_{ij})^{(2,001)} + \pi_{ij}^{(000)} \cdot f_{ij}(Y_{ij})^{(2,000)} \right] \times \right. \\
& \left. \prod_{i:C_{ij}=1, D_{ij}=1} \left[\pi_{ij}^{(111)} \cdot f_{ij}(Y_{ij})^{(2,111)} + \pi_{ij}^{(011)} \cdot f_{ij}(Y_{ij})^{(2,011)} \right] \right] \times \\
& \prod_{j:Z_j=3} \left[\prod_{i:C_{ij}=1, D_{ij}=0} \left[\pi_{ij}^{(000)} \cdot f_{ij}(Y_{ij})^{(3,000)} \right] \times \right. \\
& \left. \prod_{i:C_{ij}=1, D_{ij}=1} \left[\pi_{ij}^{(111)} \cdot f_{ij}(Y_{ij})^{(3,111)} + \pi_{ij}^{(011)} \cdot f_{ij}(Y_{ij})^{(3,011)} + \pi_{ij}^{(001)} \cdot f_{ij}(Y_{ij})^{(3,001)} \right] \right]
\end{aligned}$$

where $\mathbf{Z}$ is a J -dimensional vector with j th element equal to Z_j , $\mathbf{C}, \mathbf{D}$, and $\mathbf{Y}$ are n -dimensional vectors with ij th elements equal to C_{ij} , D_{ij} and Y_{ij} , respectively, and $\mathbf{X}$ is an $n \times K$ matrix with ij th row equal to $\mathbf{X}_{ij}$, $i = 1, \dots, n_j$, $j = 1, \dots, J$.

Let $p(\boldsymbol{\psi})$ denote the joint prior distribution of $\boldsymbol{\psi}$. The posterior distribution,

$$p(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{C}, \mathbf{D}, \mathbf{Y}, \mathbf{X}) \propto p(\boldsymbol{\psi}) \times \mathcal{L}_{obs}(\boldsymbol{\psi} \mid \mathbf{Z}, \mathbf{C}, \mathbf{D}, \mathbf{Y}, \mathbf{X})$$

is analytically intractable due to the complex mixture structure of the likelihood function. We deal with computational issues approximating the posterior distribution via Stan (Stan Development Team, 2014), and by using a data augmentation approach. The latter involves, at each iteration, the imputation of the missing potential outcomes, first for D_{ij} , then for $S_{N_{ij}}$ and Y_{ij} , for each unit, and thus the estimation of the effects as a function of the observed and missing information.

Specifically, the algorithm for the estimation of PCEs consists of the following steps. First, for each student, the missing potential outcomes for D_{ij} are imputed from its posterior predictive distribution. Second, principal stratum memberships are determined based on the joint (observed and imputed) potential outcomes for D_{ij} . Third,

the missing potential outcomes for $S_{\mathcal{N}_{ij}}$ are determined based on the imputed principal stratum memberships of friends: for example, if the subject has half friends imputed as Never Takers and half friends imputed as Reward Compliers, for that subject this imputation step results in $S_{\mathcal{N}_{ij}}(1) = 0$, $S_{\mathcal{N}_{ij}}(2) = 0$ and $S_{\mathcal{N}_{ij}}(3) = 0.5$. Fourth, the imputed potential outcomes for $S_{\mathcal{N}_{ij}}$ are used to impute the missing potential outcomes of the form $Y_{ij}(z, S_{\mathcal{N}_{ij}}(z))$ from its posterior predictive distribution. Last, PCEs are computed as function of the observed and imputed $Y_{ij}(z, S_{\mathcal{N}_{ij}}(z))$.

To compute CDEs and CSEs, under Assumption 20, we use the following relationship conditionally on the imputed principal stratum membership:

$$\mathbb{E}[Y_{ij}(z, s_{\mathcal{N}_{ij}}) \mid G_{ij} = g, \mathbf{X}_{ij}] = \mathbb{E}[Y_{ij} \mid Z_j = z, S_{\mathcal{N}_{ij}} = s_{\mathcal{N}_{ij}}, G_{ij} = g, \mathbf{X}_{ij}]. \quad (4.9)$$

Equation (4.9) links the conditional mean of the potential outcomes, $Y_{ij}(z, s_{\mathcal{N}_{ij}})$, given principal stratum membership and covariates and the conditional mean of the observed outcomes, Y_{ij} , given principal stratum membership, encouragement status, mediator and covariates. It allows computing $\mathbb{E}[Y_{ij}(z, s_{\mathcal{N}_{ij}}) \mid G_{ij} = g, \mathbf{X}_{ij}]$ by plugging in the model for $\mu_{ij,g,z}$ (Equation 4.8), i.e., the Poisson mean, the value $s_{\mathcal{N}_{ij}}$. Then, $Y_{ij}(z, s_{\mathcal{N}_{ij}})$ can be drawn from its posterior predictive distribution and used to compute controlled effects.

Similarly, to compute NDEs and NIEs, under Assumptions 20 and 21, we use the mediation formula (Pearl, 2001) conditionally on imputed potential outcomes for $S_{\mathcal{N}_{ij}}$:

$$\begin{aligned} \mathbb{E}[Y_{ij}(z, S_{\mathcal{N}_{ij}}(z')) \mid G_{ij} = g, \mathbf{X}_{ij}] = \\ \int \mathbb{E}[Y_{ij} \mid Z_j = z, S_{\mathcal{N}_{ij}} = s_{\mathcal{N}_{ij}}, G_{ij} = g, \mathbf{X}_{ij}] f_{S_{ij}}(s_{\mathcal{N}_{ij}} \mid Z_j = z', G_{ij} = g, \mathbf{X}_{ij}) \end{aligned} \quad (4.10)$$

In other words, we compute $\mathbb{E}[Y_{ij}(z, S_{\mathcal{N}_{ij}}(z')) \mid G_{ij} = g, \mathbf{X}_{ij}]$ by plugging in the model for $\mu_{ij,g,z}$ (Equation 4.8) the imputed value for $S_{\mathcal{N}_{ij}}(z')$. Then, $Y_{ij}(z, S_{\mathcal{N}_{ij}}(z'))$ can be drawn from its posterior predictive distribution and used to compute natural effects.

We simulate the posterior distributions of NDEs, NIEs, CDEs and CSEs by implementing the algorithm just outlined and executing it in three independent chains from different starting values. We run each chain for 3000 iterations, discarding the first half and saving the second half. The Markov chains mix well (see Appendix

C.1 for convergence checks). We combine the three chains and use the remaining 4500 iterations to draw inference.

4.5 Results

4.5.1 Principal Strata

Table 4.7 shows some summary statistics for the posterior distribution of the principal strata membership. It is worth noting that the posterior means for the prin-

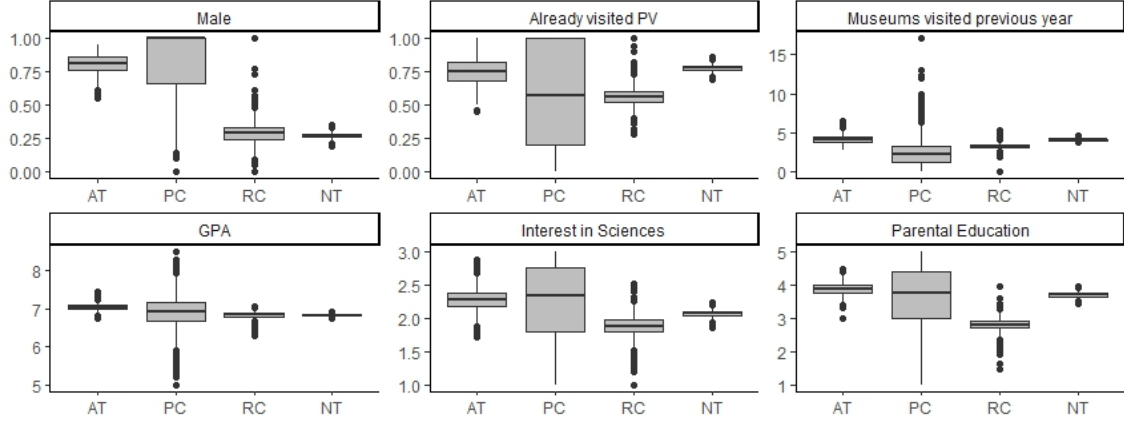
TABLE 4.7: Summary statistics of the posterior probabilities of principal stratum membership

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>5%</i>	<i>95%</i>	<i>97.5%</i>
Always Takers	0.097	0.028	0.053	0.056	0.143	0.158
Presentation Compliers	0.017	0.020	0.000	0.000	0.057	0.068
Reward Compliers	0.291	0.054	0.177	0.199	0.376	0.387
Never Takers	0.596	0.039	0.523	0.534	0.662	0.680

incipal stratum proportions are slightly different from the method-of-moment estimates, especially for the proportion of Always Takers and Presentation Compliers. These difference are mainly attributable to the fact that the method-of-moment estimates do not account for individual and network level student’s pre-experimental characteristics, which may affect the estimates given the observed imbalance across encouragement groups.

An appealing advantage of the model-based Bayesian approach we use is that it easily allows us to investigate the distribution of baseline characteristics within each principal stratum, which may provide precious information on principal stratum members. Figure 4.2 shows the boxplots of the distribution of the posterior mean of the covariates for each principal stratum. All strata comprise students with similar GPAs, but they are somewhat different in terms of the other characteristics. We do not elaborate further on the principal strata characterization here. Since it is consistent with found by Forastiere, Lattarulo, et al. (2021), we refer to that work for a full description.

FIGURE 4.2: Distribution of the posterior mean of the covariates by principal strata



4.5.2 Overall Principal Causal Effects and Principal Natural Direct and Indirect Effects

Of the four strata described above, we now focus on the two largest ones, i.e., Never Takers and Reward Compliers. For these principal strata, Table 4.8 reports summary statistics of the posterior distributions of the finite-population overall principal causal effects and principal natural direct and indirect effects on count of museum visits during the follow-up period.

Let us first focus on the contrast between Presentation and Flyer. We find evidence that principal causal effects of Presentation versus Flyer for Never Takers and Reward Compliers are positive. The posterior mean of the PCE for Never Takers is 1.59 (SD=0.89) and the 90% credible interval (CI) only covers positive values; the posterior probability that this effect is positive is very high (0.97). The posterior mean of the PCE for Reward Compliers suggests an increase of 5.32 museum visits due to presentation versus flyer (SD=1.40 and 90% CI=(3.27, 7.62)) and this principal causal effect takes on positive values with probability one. Both results are consistent with those in Forastiere, Lattarulo, et al. (2021). Based on these results, Presentation appears as an effective encouragement towards attending museums in the future for Reward Compliers and, to a smaller extent, for Never Takers.

As argued in Section 4.3.4, overall PCEs may originate from both a direct and an indirect causal pathway. Disentangling these pathways is our main contribution, moving a step forward with respect to the analyses in Forastiere, Lattarulo, et al.

TABLE 4.8: Summary statistics of the posterior distributions of overall principal causal effects and principal natural direct and indirect effects for Never Takers and Reward Compliers

Causal Estimand	Never Takers ($G_{ij} = 000$)					Reward Compliers ($G_{ij} = 001$)				
	Mean	SD	5%	95%	$p(\cdot > 0)$	Mean	SD	5%	95%	$p(\cdot > 0)$
$PCE(2 \text{ vs } 1)^{(g)}$	1.59	0.89	0.19	3.11	0.97	5.32	1.40	3.27	7.62	1.00
$NDE(2 \text{ vs } 1, S_{N_{ij}}(2))^{(g)}$	1.58	0.93	0.07	3.12	0.96	5.31	1.41	3.23	7.61	1.00
$NDE(2 \text{ vs } 1, S_{N_{ij}}(1))^{(g)}$	1.57	0.91	0.17	3.13	0.96	5.13	1.53	2.89	7.69	1.00
$NIE(2, S_{N_{ij}}(2) \text{ vs } S_{N_{ij}}(1))^{(g)}$	0.02	0.21	-0.32	0.36	0.52	0.19	0.50	-0.62	0.99	0.66
$NIE(1, S_{N_{ij}}(2) \text{ vs } S_{N_{ij}}(1))^{(g)}$	0.00	0.22	-0.33	0.38	0.48	0.01	0.17	-0.23	0.29	0.46
$PCE(3 \text{ vs } 2)^{(g)}$	0.13	0.32	-0.36	0.66	0.64	-2.33	1.66	-5.03	0.26	0.07
$NDE(3 \text{ vs } 2, S_{N_{ij}}(3))^{(g)}$	0.00	0.00	0.00	0.00	0.00	-2.36	1.74	-5.21	0.32	0.07
$NDE(3 \text{ vs } 2, S_{N_{ij}}(2))^{(g)}$	0.00	0.00	0.00	0.00	0.00	-2.02	2.20	-5.31	1.57	0.15
$NIE(3, S_{N_{ij}}(3) \text{ vs } S_{N_{ij}}(2))^{(g)}$	0.13	0.32	-0.36	0.66	0.64	-0.31	1.04	-2.12	0.89	0.42
$NIE(2, S_{N_{ij}}(3) \text{ vs } S_{N_{ij}}(2))^{(g)}$	0.13	0.32	-0.36	0.66	0.64	0.03	0.72	-1.07	1.24	0.49
$PCE(3 \text{ vs } 1)^{(g)}$	1.72	0.92	0.34	3.29	0.98	2.99	1.27	1.13	5.23	0.99
$NDE(3 \text{ vs } 1, S_{N_{ij}}(3))^{(g)}$	1.23	1.58	-1.10	3.12	0.87	2.80	1.96	0.70	5.11	0.97
$NDE(3 \text{ vs } 1, S_{N_{ij}}(1))^{(g)}$	1.57	0.91	0.17	3.13	0.96	3.31	1.93	0.82	6.72	0.99
$NIE(3, S_{N_{ij}}(3) \text{ vs } S_{N_{ij}}(1))^{(g)}$	0.15	0.33	-0.39	0.72	0.67	-0.31	1.06	-2.16	0.92	0.42
$NIE(1, S_{N_{ij}}(3) \text{ vs } S_{N_{ij}}(1))^{(g)}$	0.48	1.20	-0.46	2.04	0.68	0.20	1.44	-0.24	0.91	0.57

(2021) and enriching the literature on causal inference with interference. When comparing the Presentation and the Flyer, the direct pathway, whose strength is evaluated by principal NDEs, channels only encouragement effects for both Never Takers and Reward Compliers. Instead, the indirect pathway, whose strength is evaluated by principal NIEs, may only channel spillovers. Principal natural direct effects of Presentation versus Flyer can be evaluated setting the proportion of friends visiting Palazzo Vecchio to the value that it would have taken either under the Flyer encouragement, $S_{N_{ij}}(1)$, or under the Presentation encouragement, $S_{N_{ij}}(2)$. Similarly, principal natural indirect effects measure spillover effects on future museum attendance arising from a change in the proportion of friends visiting Palazzo Vecchio from $S_{N_{ij}}(1)$, the value that $S_{N_{ij}}$ would have taken under the Flyer encouragement ($z = 1$), to $S_{N_{ij}}(2)$, the value that $S_{N_{ij}}$ would have taken under the Presentation encouragement ($z = 2$), by fixing the individual encouragement to either the Flyer level ($z = 1$) or the Presentation level ($z = 2$). It is worthwhile to highlight that $S_{N_{ij}}(1)$ is equal to the proportion of Always Taker friends, because only Always Takers visit Palazzo Vecchio under the Flyer encouragement; analogously, $S_{N_{ij}}(2)$ is equal to the proportion of friends that are either Always Takers or Presentation Compliers, because only Always Takers and Presentation Compliers visit Palazzo Vecchio under the Presentation encouragement. Therefore, principal natural indirect effects quantify the spillover on future museum attendance that would come

from friends ready to visit Palazzo Vecchio under the Presentation but not under the Flyer, that is, from being exposed, in addition to Always Takers, to friends that react as Presentation Compliers. Since the proportions of Always Takers and Presentation Compliers are small, $S_{\mathcal{N}_{ij}}(1)$ and $S_{\mathcal{N}_{ij}}(2)$ describe a friendship environment where a student is hardly exposed to friends that might send spillovers.

The posterior means of $NDE(2 \text{ vs } 1, S_{\mathcal{N}_{ij}}(2))^{(g)}$ and $NDE(2 \text{ vs } 1, S_{\mathcal{N}_{ij}}(1))^{(g)}$ are, respectively, 1.58 and 1.57 for Never Takers and 5.31 and 5.13 for Reward Compliers, and the corresponding posterior probabilities that these effects are positive are higher than 0.95. Therefore, for Never Takers and Reward Compliers there exist pure encouragement effects of Presentation versus Flyer. The posterior means of $NIE(1, S_{\mathcal{N}_{ij}}(2) \text{ vs } S_{\mathcal{N}_{ij}}(1))^{(g)}$ and $NIE(2, S_{\mathcal{N}_{ij}}(2) \text{ vs } S_{\mathcal{N}_{ij}}(1))^{(g)}$ are very small for both Never Takers and Reward Compliers, and the corresponding 90% CIs cover zero, therefore there is no evidence that there exist spillover effects of Flyer versus Presentation for these type of subjects. Indeed, as we can see in Table 4.8, the size of the estimated principal natural direct effects for Never Takers and Reward Compliers is similar to the size of the overall PCEs. Therefore, causal effects for Never Takers and Reward Compliers are almost entirely due to the encouragement.

Let us now focus on the contrast between Reward ($z = 3$) and Presentation ($z = 2$). For Never Takers, principal natural direct effects of Presentation versus Reward are zero by assumption (see the discussion in Sections 4.3.4 and 4.3.5), therefore principal causal effects of $z = 3$ versus $z = 2$ for this type of subjects are interpretable as spillover effects. Consistently with the insights of Forastiere, Lattarulo, et al. (2021), we find no evidence that there exist statistically significant spillover effects of Presentation versus Reward for Never Takers: the posterior means of $NIE(2, S_{\mathcal{N}_{ij}}(3) \text{ vs } S_{\mathcal{N}_{ij}}(2))^{(000)}$ and $NIE(3, S_{\mathcal{N}_{ij}}(3) \text{ vs } S_{\mathcal{N}_{ij}}(2))^{(000)}$ are relatively close to zero and the corresponding 90% CIs are spread around zero. For Reward Compliers, the posterior probability that overall PCEs of Presentation versus Reward are negative is 0.93, in line with the results previously presented in Forastiere, Lattarulo, et al. (2021). We interpret these overall PCEs for Reward Compliers as a mixture of experience and spillover effects, which we propose to disentangle using principal natural direct and indirect effects. Specifically, we attribute principal natural direct effects of Presentation versus Reward for Reward Compliers solely to the experience (the visit at Palazzo Vecchio), because they visit Palazzo Vecchio only under Reward, and principal natural indirect effects to spillovers. Summary

statistics of the posterior distributions of these effects shown in Table 4.8 suggest that experience effects account for almost all of the negative overall PCEs. Thus, Reward Compliers do not seem to have enjoyed their experience at Palazzo Vecchio very much, which may have even discouraged further museum visits in the follow-up period. If the effect of the Presentation itself is to enhance Reward Compliers' museum attendance, such effect is partially offset by the experience at Palazzo Vecchio. Spillovers are, instead, negligible. This conclusion on spillovers is particularly relevant because, contrary to the Presentation/Flyer contrast, the Reward is able to increase the proportion of friends visiting Palazzo Vecchio by attracting Reward Compliers, who are about 1/3 of students and may act as senders of spillovers.

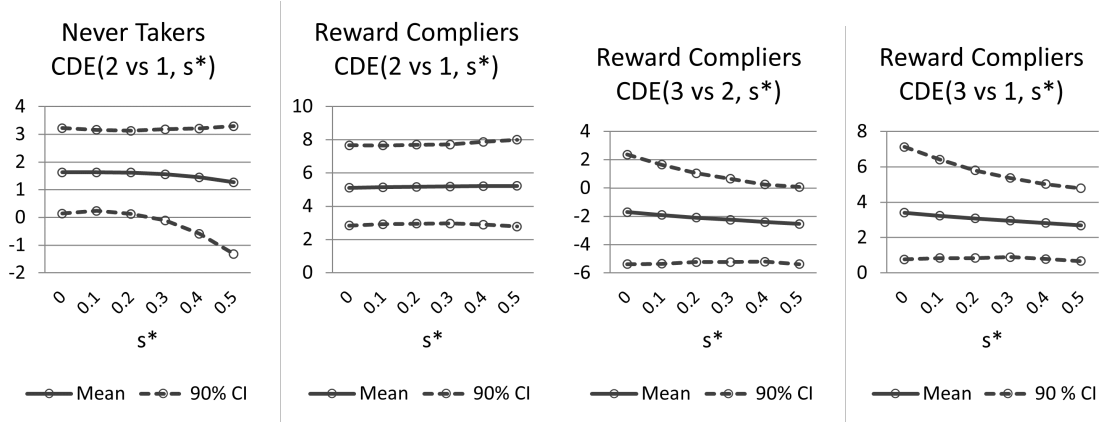
Finally, let us focus on the contrast between Reward and Flyer. Overall PCEs of Reward versus Flyer amounts to the sum of the overall PCEs of Presentation versus Flyer and of Reward versus Presentation. We find evidence that for both Never Takers and Reward Compliers, these principal causal effects are positive and mainly due to direct effects.

4.5.3 Principal Controlled Direct and Spillover Effects

Principal controlled direct effects measure causal effects of the encouragement in situations where the proportion of friends performing the visit to Palazzo Vecchio is arbitrarily set by the researcher to a fixed value s^* for all units. In principle, any value $s^* \in [0, 1]$ could be chosen. To avoid excessive extrapolation, we focus on a range of proportions that are not unreasonably distant from those actually supported by the data. Figure 4.3 reports the CDEs evaluated at six different levels of s^* .

For Never Takers, it is worth showing only the CDE of the Presentation relative to the Flyer, $CDE(2 \text{ vs } 1, s^*)^{(000)}$. As shown in Table 4.5, the CDE of the Reward relative to the Presentation for Never Takers, $CDE(3 \text{ vs } 2, s^*)^{(000)}$, is assumed to be zero because the Reward applies only to students who actually visit Palazzo Vecchio, and disappears in the follow-up period. As a consequence, $CDE(3 \text{ vs } 1, s^*)^{(000)} = CDE(2 \text{ vs } 1, s^*)^{(000)}$. The posterior means of the CDEs of the Presentation relative to the Flyer for Never Takers show some evidence that there exists a positive direct effect of Presentation versus Flyer for low proportions of friends visiting Palazzo Vecchio (for $s^* \in \{0, 0.1, 0.2\}$), and that this effect is almost invariant with respect

FIGURE 4.3: Estimated CDEs



to s^* . We cannot state the same for higher proportions of friends visiting Palazzo Vecchio, however, because the corresponding 90% CIs become quite large.

For Reward Compliers, it makes sense to consider CDEs relative to any encouragement contrast (see previous discussion and Table 4.5). $CDE(2 \text{ vs } 1, s^*)^{(001)}$ can be interpreted as a pure encouragement effect, $CDE(3 \text{ vs } 2, s^*)^{(001)}$ as the effect of the visit experience at Palazzo Vecchio, and $CDE(3 \text{ vs } 1, s^*)^{(001)}$ as the sum of the previous two quantities.

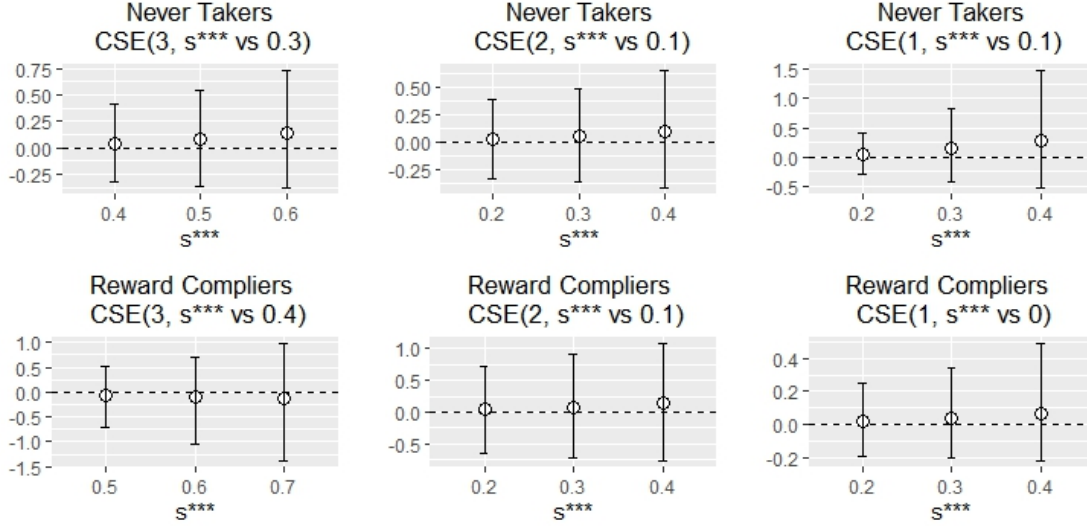
The CDEs of the Presentation relative to the Flyer for Reward Compliers are to raise the number of subsequent visits by more than 5 points, regardless of the fixed proportion of friends visiting Palazzo Vecchio. However, when s^* increases, the estimated CDEs slightly increase, suggesting that a higher proportion of friends visiting Palazzo Vecchio could prompt the positive direct effect of the Presentation (relative to the Flyer) for Reward Compliers.

The posterior means of the CDEs of Reward versus Presentation for Reward Compliers are, instead, negative but statistically negligible, even if the upper bound of the 90% CIs approach zero when s^* increases. Therefore, there is no evidence that the Palazzo Vecchio experience may prompt future museum visits for Reward Compliers.

Finally, Figure 4.4 shows the means and the 90% CIs of the posterior distributions of CSEs. In particular, the CSEs are evaluated by incrementally varying the observed average proportion of friends visiting Palazzo Vecchio under a given encouragement. This allows to evaluate how spillovers would change if the proportion of friends

visiting Palazzo Vecchio were moderately higher than the latter, but the encouragement was kept at a fixed level. The spillover generated by any of these changes on

FIGURE 4.4: Estimated CSEs for Never Takers and Reward Compliers



Note. The blue dashed line indicates zero.

future museum attendance of both Never Takers and Reward Compliers are small and statistically negligible: the posterior means are close to zero and the 90% CIs cover zero and are approximately symmetric around zero. Therefore, no matter how much an individual is exposed to friends visiting Palazzo Vecchio, spillovers are hardly relevant in the decision to attend museums in the follow-up period.

4.6 Posterior predictive checks

We evaluate the goodness of fit of the models by performing posterior predictive checks (e.g., Rubin, 1984). For this purpose, we chose to refer again to the discrepancy measures defined in Section 1.6, and to estimate the corresponding Bayesian p -values. Yet, the procedure required to generate replicates of the Florentine Study data is complicated by the presence of network interference. We provide details below.

Let ψ^k ($k = 1, \dots, K$) denote the draw of the parameters vector from its posterior distribution at iteration k . We simulate $K = 3000$ replicated data sets as follows. First, we generate the compliance status for every subject in the sample from the

hypothesized principal strata model given ψ^k , i.e., $G_{ij}^{rep,k}$. Once we have this information, we know – by definition – who would visit Palazzo Vecchio under the three encouragements, i.e., $D_{ij}^{rep,k}(z)$ for $z = 1, 2, 3$. Given that we also know the network that links the subjects (that we consider fixed while conducting posterior predictive checks), we can determine the proportion of friends who would visit Palazzo Vecchio under the different encouragement levels, i.e., $S_{\mathcal{N}_{ij}}^{rep,k}(z)$ for $z = 1, 2, 3$. Thus, it is possible to simulate the outcome $Y_{ij}^{rep,k}$ from its posterior predictive distribution, keeping the encouragement Z_j fixed at its observed value and using the proportion of friends who would visit Palazzo Vecchio under that encouragement in the replicated data set, i.e., $S_{\mathcal{N}_{ij}}^{rep,k}$. We repeat the whole procedure K times and thus we obtain K replicated data sets, $(G_{ij}^{rep,k}, S_{\mathcal{N}_{ij}}^{rep,k}, Y_{ij}^{rep,k}, Z_j)$. Once we have these replicated data sets, we can use them to compute all the discrepancy measures of interest.

Let us indicate with $\mathcal{D}_{z,g}^{study} = \{ij : Z_j = z \text{ and } G_{ij}^{study} = g\}$ the group of students with compliance status $G_{ij}^{study} = g$ assigned to $Z_j = z$ in the study data, where $study = obs$ denotes the observed data and $study = rep$ a replication of them obtained from their posterior predictive distribution as described above. Let $|\mathcal{D}_{z,g}^{study}|$ indicate the cardinality of $\mathcal{D}_{z,g}^{study}$, i.e., the number of students in this group. Finally, let $\bar{Y}_{z,g}^{study}$ and $s_{z,g}^{2,study}$ denote the mean and the variance of the outcome variable Y^{study} for units in the $\mathcal{D}_{z,g}^{study}$ group. We choose the following discrepancy measures:

$$SI_{z,z',g}^{study} = |\bar{Y}_{z,g}^{study} - \bar{Y}_{z',g}^{study}|, \quad NO_{z,z',g}^{study} = \sqrt{\frac{s_{z,g}^{2,study}}{|\mathcal{D}_{z,g}^{study}|} + \frac{s_{z',g}^{2,study}}{|\mathcal{D}_{z',g}^{study}|}}, \quad \text{and}$$

$$SN_{z,z',g}^{study} = \frac{SI_{z,z',g}^{study}}{NO_{z,z',g}^{study}} \quad \text{for } z, z' \in \{1, 2, 3\}, \text{ with } z > z', z' \neq z, \text{ and for all } g.$$

The estimated PPPVs for Never Takers and Reward Compliers, reported in Table 4.9, are all faraway from zero and one, suggesting good model fit.

4.7 Conclusions

In this work we show how information on networks can be used to investigate spillover effects within cluster encouragement designs with individual non-compliance. We propose an original methodological interplay between principal stratification and

TABLE 4.9: Estimated PPPVs for Never Takers and Reward Compliers

Discrepancy Measure	Never Takers ($G_{ij}^{study} = 000$)	Reward Compliers ($G_{ij}^{study} = 001$)
$SI_{3,2,g}^{study}$	0.652	0.410
$SI_{2,1,g}^{study}$	0.451	0.415
$SI_{3,1,g}^{study}$	0.460	0.161
$NO_{3,2,g}^{study}$	0.366	0.394
$NO_{2,1,g}^{study}$	0.516	0.312
$NO_{3,1,g}^{study}$	0.307	0.593
$SN_{3,2,g}^{study}$	0.657	0.424
$SN_{2,1,g}^{study}$	0.447	0.501
$SN_{3,1,g}^{study}$	0.529	0.158

mediation analysis, where a synthesis of the network information is used as a mediator which may generate spillover effects. We formally define principal natural direct and indirect effects and principal controlled direct and spillover effects, and explain how these quantities can be interpreted to get information on different causal pathways, including spillover. We revisit results from a small field experiment, based on a clustered encouragement design, conducted in Florence (Italy) to study how appropriate incentives may lead students to visit museums in their free time. Possible causal pathways can either originate from such incentives directly affecting the student's outcome, or follow an indirect trail where the student's outcome is finally affected by the spillovers received from other students involved in the experiment. We use a Bayesian model-based approach to inference under latent sequential ignorability assumptions of the mediator. Previous causal studies of these experimental data (Forastiere, Lattarulo, et al., 2021; Lattarulo et al., 2017) have not exploited the available information on friendship networks within each cluster of students, which we use here to disentangle and estimate spillover effects, and have thus settled for causal effects where spillovers are usually blended with other sources of change in the outcome.

Since our application is based on a small field experiment, the conclusions we reach would perhaps require corroboration in similar studies conducted on a larger scale. However, the results of our analysis suggest that spillovers from friends did not lead

to significant changes in the students' museum attendance, that the Presentation encouragement was effective in motivating students to devote some free time to museum attendance, and that the visit experience that students were offered at Palazzo Vecchio was not effective, at least for the subpopulation of Reward Compliers.

In conclusion, note that our considerations are based on the assumption that spillovers pass through the proportion of friends visiting Palazzo Vecchio. Future works might investigate whether a different synthesis of the information about the visit to Palazzo Vecchio of friends could be a vehicle for such effects. Also, in principle, different types of networks might be at work, although the important role played by friends in influencing teens' behavior is hardly questionable (e.g., Berndt & Murphy, [2002](#)).

Chapter 5

Final remarks

The evaluation of the causal impact of a treatment or an intervention is particularly challenging in the presence of complications arising after treatment has been assigned or received. One of such complications is non-compliance, which frequently arises in experimental or quasi-experimental designs where the treatment cannot be enforced for ethical or practical reasons. In social studies, non-compliance means that one or more units refuse to take the assigned treatment. In clinical trials, it usually specifically indicates how many doses one or more patients missed while still on treatment. Rather than non-compliance, researchers in the medical field usually have to account for non-adherence, which means that one or more patients deviate from the protocol in several different ways, which do not necessarily involve the number of doses of the drug taken.

The contributions of this Thesis are twofold. First, we provide a description of the Principal Stratification framework, together with an overview of how it has been declined to address non-compliance but also several different post-treatment complications. Then, we dwell on aspects related to the identification of the causal effects of interest, providing a systematic review of the assumptions that could be made to either reach or ease it. Also, we focus on the Bayesian model-based estimation of the causal effects, but also provide the basis to understand how the strategy could be extended to Bayesian non-parametrics. Finally, we explain how the goodness-of-fit of the model can be assessed, when a parametric approach is used, by performing posterior predictive checks.

Second, we enrich the Principal Stratification literature with three original works including novel declinations and uses of the approach. In the first work, we show

how Principal Stratification allows to relax one of the Instrumental Variables assumptions, specifically, the Exclusion Restriction. The focus is on the consequences of violations of this assumption, which we illustrate through the impact evaluation of a Community Health Financing intervention implemented in rural Uganda and affected by non-compliance. In the second work, we provide a novel declination of Principal Stratification in the context of Non-inferiority trials, which are designed to evaluate whether an experimental treatment is non-inferior to the standard-of-care. The proposal is relevant because, in this specific type of trials, the presence of treatment non-adherence leads other commonly used analysis methods, such as the Intention-To-Treat principle, to be anti-conservative, i.e., to make it more likely to conclude the non-inferiority. We propose to evaluate the non-inferiority based on the causal effects for the sub-population of those who would be able to adhere to both the treatment being compared. Also, we show through simulations that the Bayesian model-based estimator we propose to adopt has desirable properties; among them, it is not anti-conservative. In the third and last work, we exploit Principal Stratification to address non-compliance in Cluster Encouragement Designs. In these studies, the cluster-level encouragement makes it impossible to assume the absence of interference between the units. The concomitant presence of non-compliance and interference makes inferring causal effects in these designs particularly demanding. We specifically refer to a Cluster Encouragement Design implemented in Florence as an illustrative example. In this study, classes of high school students were randomized to different incentives to visit a specific museum in their free time; non-compliance is present since the decision to visit the museum was on a voluntary basis. The final aim was to prompt the general museums attendance of these students. The possibility of peer influence motivated us to develop a novel approach based on the idea that interference could spread through networks. The main contribution of this approach is that it allows to disentangle the part of the total causal effects which is due to interactions of units in the same network from the parts which pass through different causal channels, for different population sub-groups. Such quantification of spillover effects is relevant for many reasons. One of the most important ones is that allows understanding the indirect effects of a treatment on units that are not directly exposed to it, but are part of the network of a treated unit. This is fundamental in the design phase of any intervention on potentially interacting units.

Appendix A

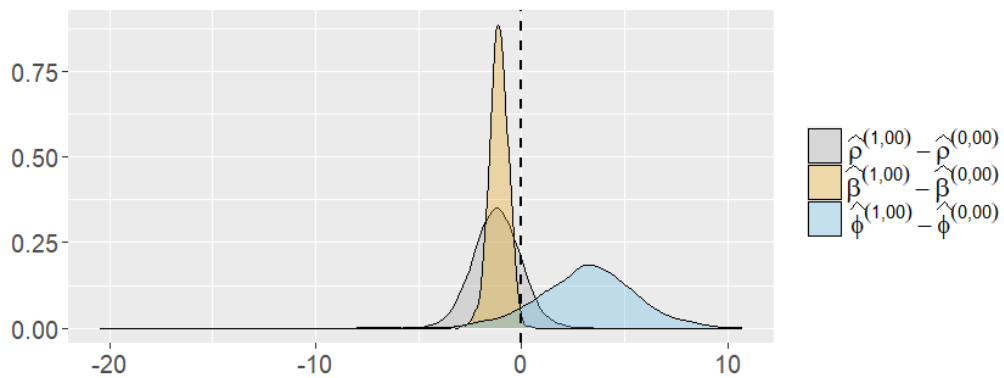
Appendix to the Oyam Health Expenditures Study

A.1 Assessing the validity of the Exclusion Restriction for Never Takers

To conduct a sensitivity analysis for the ER for Never Takers, it is possible to contrast different features of the posterior distributions of the estimated potential outcomes for Never Takers, including the parameters of these distributions. Specifically, by assessing the posterior distributions of $\hat{\rho}^{(1,00)} - \hat{\rho}^{(0,00)}$, $\hat{\beta}^{(1,00)} - \hat{\beta}^{(0,00)}$ and $\hat{\phi}^{(1,00)} - \hat{\phi}^{(0,00)}$, it is possible to understand whether the assumption would hold if it were imposed. In fact, these distributions could be centered around zero, thus lending credibility to the ER, or not.

We show the result of this sensitivity analysis in Figure A.1. The posterior distribu-

FIGURE A.1: Posterior distributions of $\hat{\rho}^{(1,00)} - \hat{\rho}^{(0,00)}$, $\hat{\beta}^{(1,00)} - \hat{\beta}^{(0,00)}$ and $\hat{\phi}^{(1,00)} - \hat{\phi}^{(0,00)}$. The dashed line indicates zero



tions in the figure are not centered around zero and, while the 95% credible intervals for $\rho^{(1,00)} - \rho^{(0,00)}$ and $\phi^{(1,00)} - \phi^{(0,00)}$ still cover zero, the one for $\beta^{(1,00)} - \beta^{(0,00)}$ does not. Therefore, this examination shows that the ER for Never Takers is not valid in this study.

A.2 Convergence checks

We use the $\hat{R}$ convergence diagnostic to assess the convergence of the MCMC algorithm, which compares the between- and within-chain estimates for model parameters and other univariate quantities of interest. If chains have not mixed well (i.e., the between- and within-chain estimates do not agree), $\hat{R}$ is larger than 1. The $\hat{R}$ we refer to is not the traditional $\hat{R}$ of Gelman and Rubin (1992), but the maximum of rank normalized split- $\hat{R}$ and rank normalized folded-split- $\hat{R}$, as proposed by Vehtari et al. (2021). We also use the 0.01 threshold suggested by these authors, which is much tighter than the one recommended by Gelman and Rubin. We assess the value of this $\hat{R}$ diagnostic for all the model parameters when assuming and when relaxing the ER for NT. The statistic always takes on values lower than 1.01, showing no evidence against convergence.

For space constraints, we report only the trace plots of the model parameters when ER for Never Takers is not assumed. The trace plots of the principal stratum model parameters are shown in Figure A.2. Those of the potential outcomes model parameters are instead shown in Figures A.3 and A.4. From the visualization of these trace plots, we can confirm that convergence has been reached. Furthermore, the posterior distributions of the model parameters are generally well-shaped.

FIGURE A.2: Trace plots of the principal stratum model parameters

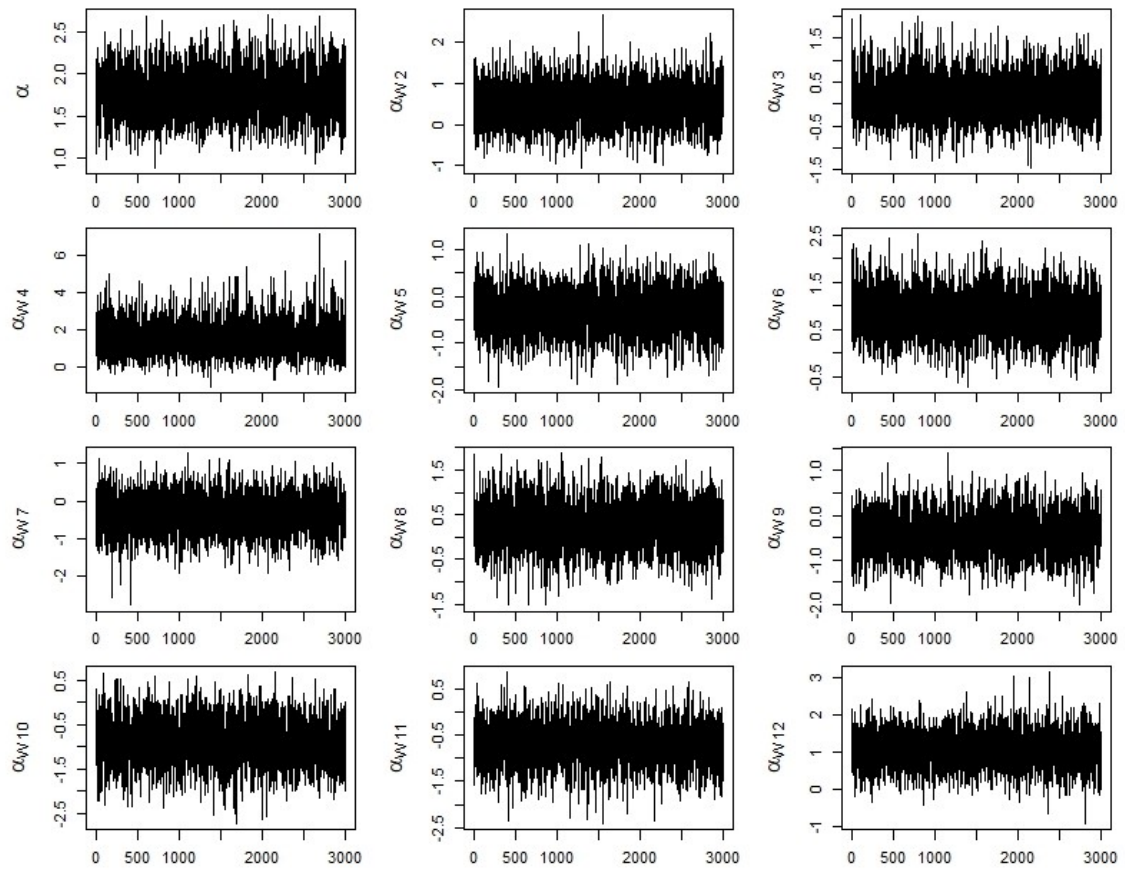


FIGURE A.3: Trace plots of the potential outcomes model parameters (part I)

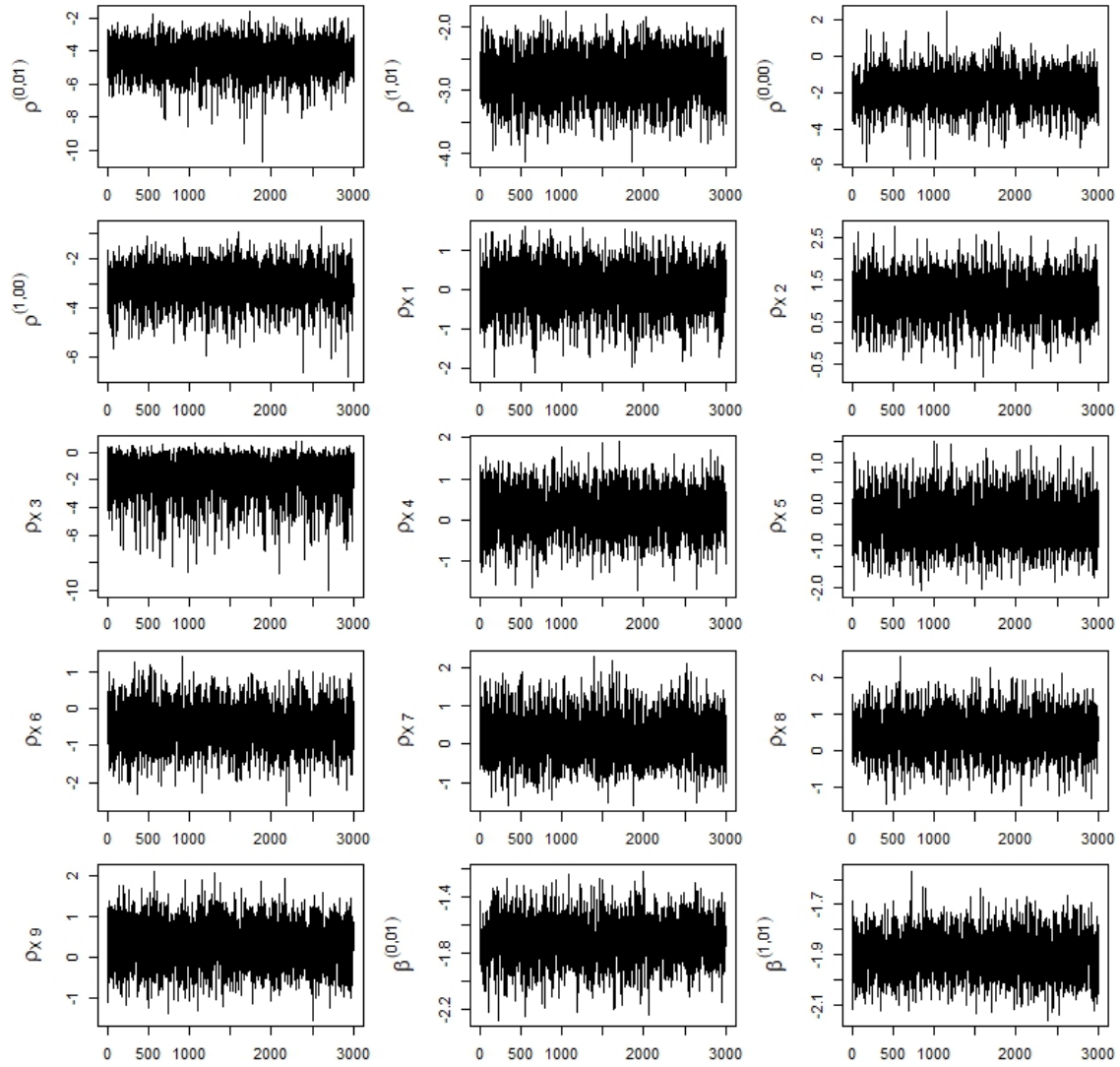
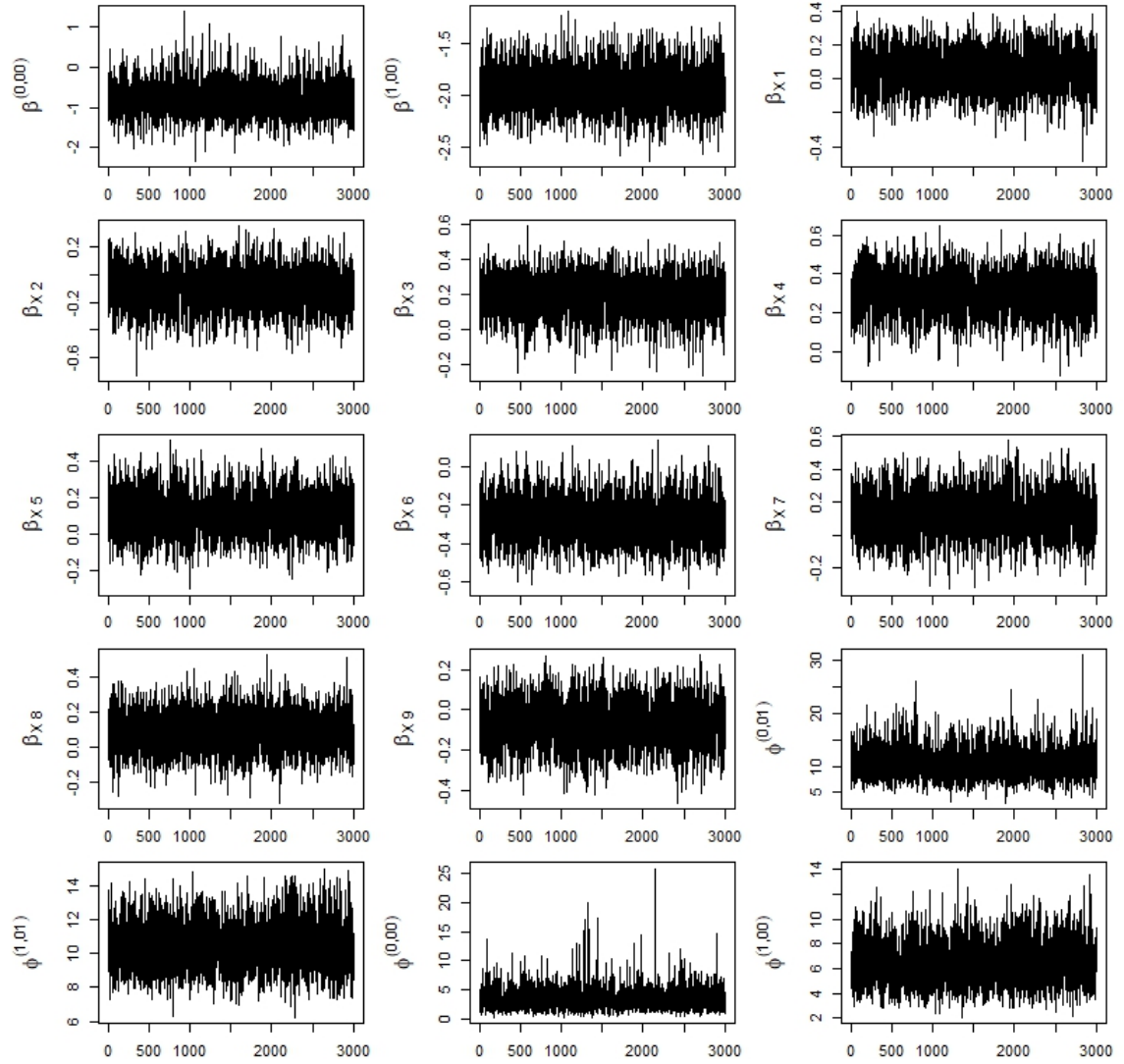
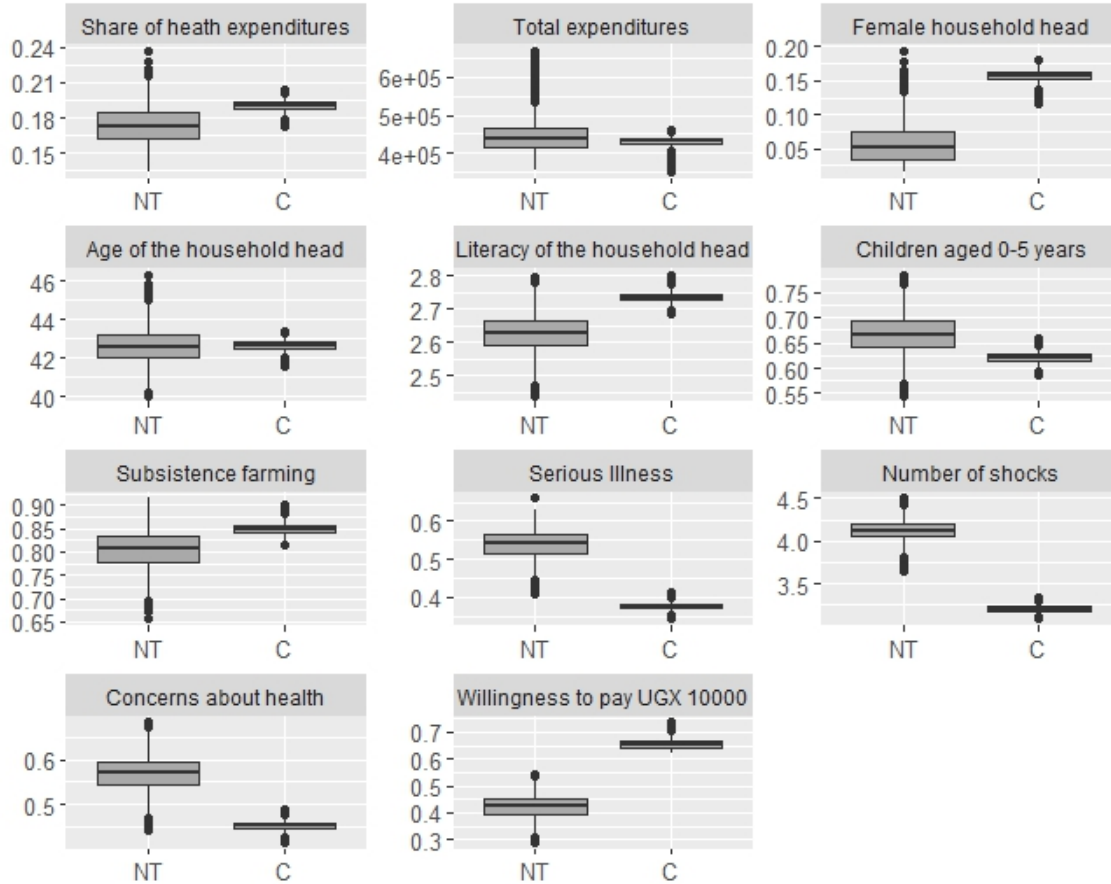


FIGURE A.4: Trace plots of the potential outcomes model parameters (part II)



A.3 Additional figures

FIGURE A.5: Results of the analysis on the share of health costs over the total monthly expenditures when the ER (Assumption 9) for NT is not stated. Boxplots summarizing the distributions of the covariates in the principal strata model by principal stratum; “NT” stands for Never Takers and “C” for Compliers



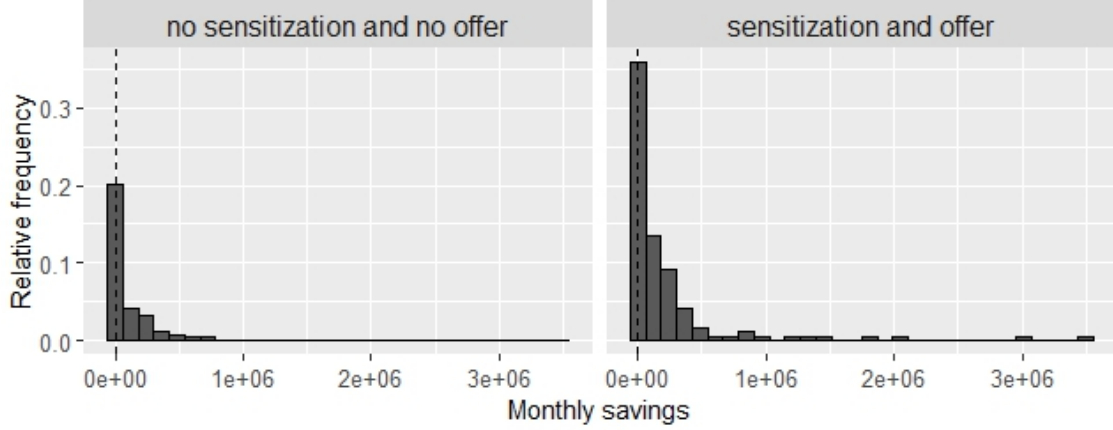
A.4 Supplementary analysis

We now want to assess whether the impact of the intervention in decreasing the monthly health costs of most households is also reflected in an increase in their monthly savings.

A.4.1 The outcome

The histograms of the empirical distribution of monthly savings by assignment group are shown in Figure A.6. From this plot, it is evident that the distribution of this

FIGURE A.6: Histogram of the empirical distribution of monthly savings by treatment assignment



outcome for households that did not receive sensitization is more shrunk towards zero than that of households that receive sensitization and were offered CHF. Moreover, we can notice the presence of several outliers in the latter distribution, i.e., several households with an unusually large amount of monthly savings among those which received sensitization. The complication due to the presence of these outliers can be addressed in several ways. A simple option could be to remove them and conduct the analysis on a subset of households only. However, we want to have here the same sample we had in our primary analysis so we can interpret the results of both analyses as a whole. Thus, we decided to categorize monthly savings according to the tertiles of its empirical distribution. We denote with $\mathbf{Y}^*$ the N -dimensional vector containing the resulting categorical outcome for the whole sample: $Y_i^* = t$ if the monthly savings of i -th household are in the t tertile ($t \in \{1, 2, 3\}$). The sample size at hand does not allow a finer categorization. Relying on SUTVA (Assumption 1), we again index the individual potential outcomes by the individual treatment assignment only; so, we indicate with $Y_i^*(z)$ the monthly savings tertile of i -th household if it is assigned to treatment condition z (with $z \in \{0, 1\}$). The rest of the notation remains the same as in the previous analyses.

A.4.2 Methodology

We define the same principal strata and we again estimate PCEs, i.e., stratum-specific effects of the intervention. We refer to the PCEs on savings as *Savings*

Effect (SE), and we define them as follows:

$$SE(t)^{(g)} = \frac{\sum_{i:G_i=g} [\mathbb{1}(Y_i^*(1) = t) - \mathbb{1}(Y_i^*(0) = t)]}{n_g} \quad t = 1, 2, 3$$

Therefore, we have a PCE on each tertile t of the empirical distribution of savings for each principal stratum g . We do not assume, here, the ER for Never Takers given that the previous analysis revealed the existence of some causal effects even for this subgroup.

We again use Bayesian inference methods. We specify the same model for the principal strata membership given the covariates, but a different model for the potential outcomes conditional on the covariates and the principal strata. Specifically, given the categorical nature of $\mathbf{Y}^*$, we model it using a multinomial logistic model of the form:

$$\log \left(\frac{p(Y_i^*(z) = t \mid G_i = g, \mathbf{S}_i)}{p(Y_i^*(z) = 3 \mid G_i = g, \mathbf{S}_i)} \right) = \gamma_t^{(z,g)} + \mathbf{S}_i' \boldsymbol{\gamma}_{S,t}$$

with $z \in \{0, 1\}$ and $t = 1, 2$ (the last category of Y^* is taken as the reference). $\boldsymbol{\gamma}_{S,t} = (\gamma_{1,t}, \dots, \gamma_{k,t})'$ are vectors of unknown regression parameters and $\mathbf{S}_i = (S_{i,1}, \dots, S_{i,k})'$ are observations on k covariates. The vector $\mathbf{S}_i$ includes: the monthly savings prior to the intervention; the total monthly expenditures and the amount of monthly savings; whether the household is single, widow or divorced; whether in the household there are members aged 5 or less and, more generally, the size of the household; whether the main source of income is subsistence farming; and, finally, the level of satisfaction with the health status of household members.

We specify weakly informative prior distributions for the parameters. Specifically, for the parameters of the principal strata model, we specify the same prior distributions that we specified in the analysis on health costs; for the parameters of the outcome model, we assume Normal prior distributions with mean 0 and standard deviation equal to 2.

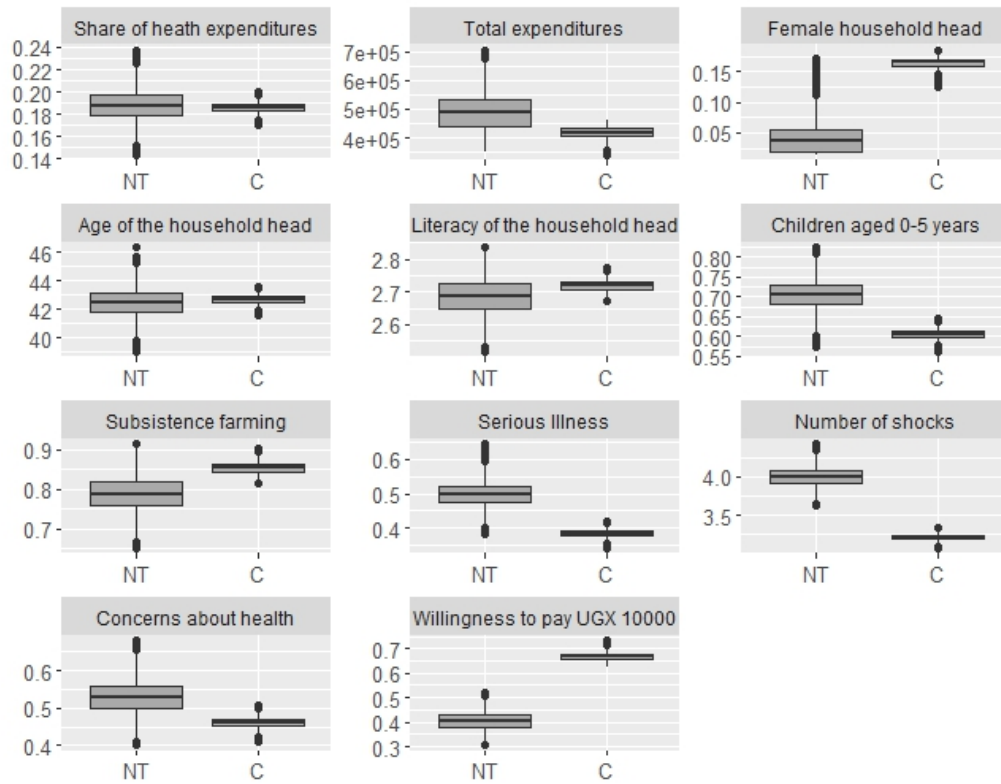
A.4.3 Results

Regarding the principal strata, the estimated proportion of Compliers is more or less the same we found in the previous analyses (Table 10). Likewise, also the observed characteristics of these principal strata are very similar to those we estimated before; we report the resulting boxplots in Figure A.7.

TABLE 10: Estimated posterior probabilities of principal strata membership

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>
Proportion of Compliers	0.765	0.021	0.720	0.807
Proportion of Never Takers	0.235	0.021	0.193	0.280

FIGURE A.7: Results of the analysis on monthly savings. Boxplots summarizing the distributions of the covariates in the principal strata model by principal stratum; “NT” stands for Never Takers and “C” for Compliers



Regarding the estimated PCEs, reported in Table 11, we found some interesting results for Compliers. On the one hand, we estimate a higher incidence of savings in the third tertile following the intervention: 16.2% more Compliers have savings in the third tertile, on average (95% credible interval -3.8% – 35.1%). On the other hand, we cannot conclude that the intervention clearly decreases the incidence of savings in the first tertile for this principal stratum, although the posterior distribution of the estimated effect is mostly negative. Finally, for Never Takers, the 95% credible intervals for the estimated PCEs are wider – because these households are fewer in

TABLE 11: Summary statistics of the posterior distribution of the estimated PCEs

	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>	<i>p</i> ($\cdot < 0$)
$SE(1)^{(01)}$	-0.150	0.126	-0.416	0.066	0.881
$SE(2)^{(01)}$	-0.013	0.112	-0.210	0.237	0.567
$SE(3)^{(01)}$	0.162	0.104	-0.038	0.351	0.054
$SE(1)^{(00)}$	-0.156	0.248	-0.569	0.347	0.725
$SE(2)^{(00)}$	0.152	0.209	-0.415	0.418	0.179
$SE(3)^{(00)}$	0.004	0.194	-0.418	0.288	0.413

number than Compliers – and include zero, so we cannot conclude that there is an effect of the sensitization received on saving.

Appendix B

Appendix to the Simulation Study

B.1 Details of the calculations

Let us introduce some additional notation that is convenient to present the details of the calculations (Table 1).

TABLE 1: Description of additional notation

h	The method adopted; $h \in \{ITT, PP, IPW, pcores, Bayes\}$
ϑ_h or $\vartheta_h^{(g)}$	The estimand method h targets. The superscript (g) is added if it is a stratum-specific estimand ($g \in \{01, 11\}$), otherwise it is avoided.
$r = 1, \dots, n_{sim}$	Indexes the repetitions of the simulation.
$\hat{\vartheta}_h$ or $\hat{\vartheta}_h^{(g)}$	The estimator of ϑ_h . The superscript (g) is added if ϑ_h is a stratum-specific estimand.
$\hat{\vartheta}_h^r$ or $\hat{\vartheta}_h^{(g)r}$	The estimate of ϑ_h from the r th repetition. The superscript (g) is added if ϑ_h is a stratum-specific estimand.

B.1.1 ITT

In terms of potential outcomes, the super-population ITT estimand can be defined as $\vartheta_{ITT} := \mathbb{E}[Y_i(1) - Y_i(0)]$, i.e., the expected difference in the potential outcomes for a patient ignoring information on adherence. In the simulations, its true value can be computed as

$$\begin{aligned} \vartheta_{ITT} &= \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \\ &\quad \left[\mathbb{E}[Y_i(1) \mid G_i = 11] \cdot p(G_i = 11) + \mathbb{E}[Y_i(1) \mid G_i = 01] \cdot p(G_i = 01) \right] - \\ &\quad \left[\mathbb{E}[Y_i(0) \mid G_i = 11] \cdot p(G_i = 11) + \mathbb{E}[Y_i(0) \mid G_i = 01] \cdot p(G_i = 01) \right]. \end{aligned}$$

An ITT analysis can or cannot be covariates-adjusted. Here, we choose to adjust for the covariates and we use Bayesian model-based inference for obtaining posterior draws of the parameters. In particular, we assume the following generalized linear model with probit link for the potential outcomes:

$$p(Y_i(z) = 1 \mid \mathbf{X}_i) = p(Y_i^*(z) > 0) \quad \text{for } z \in \{0, 1\},$$

$$Y_i^*(z) = \mathbf{X}_i' \boldsymbol{\beta}^{(z)} + u_i, \quad u_i \sim \mathcal{N}(0, 1).$$

Therefore, $p(Y_i(z) = 1 \mid \mathbf{X}_i) = \Phi(\mathbf{X}_i' \boldsymbol{\beta}^{(z)})$. We assume that the parameters are a priori independent and we set conjugate prior distributions for them: $\boldsymbol{\beta}^{(z)} \sim \mathcal{N}(\boldsymbol{\mu}_\beta^{(z)}, \boldsymbol{\Sigma}_\beta^{(z)})$. In addition, we set the hyperparameters as follows in order to have weakly informative prior distributions: $\boldsymbol{\mu}_\beta^{(z)} = (0, 0, 0, 0)$, $\boldsymbol{\Sigma}_\beta^{(z)} = 5 \cdot \mathbb{I}_{4 \times 4}$.

The posterior distributions of the parameters are approximated by using an MCMC method, specifically, a Gibbs Sampling. Let $\boldsymbol{\Omega}^{(k)}$, where $\boldsymbol{\Omega}^{(k)} = \{\boldsymbol{\beta}^{(0)(k)}, \boldsymbol{\beta}^{(1)(k)}\}$, denote the state of the chain at iteration k . The state of the chain at iteration $k+1$ follows from applying the following steps.

1. The full conditional distribution of $Y_i^*(z)$ is given by the following Truncated Normal ($\mathcal{TN}$) distributions:

$$Y_i^*(z) \mid Y_i, \mathbf{X}_i; \boldsymbol{\beta}^{(z)} \sim \begin{cases} \mathcal{TN}(\mathbf{X}_i' \boldsymbol{\beta}^{(z)}, 1, \text{lower} = 0, \text{upper} = +\infty) & \text{if } Y_i = 1 \\ \mathcal{TN}(\mathbf{X}_i' \boldsymbol{\beta}^{(z)}, 1, \text{lower} = -\infty, \text{upper} = 0) & \text{if } Y_i = 0 \end{cases}$$

Therefore, we sample the latent variable $Y_i^*(z)$ from $\mathcal{N}(\mathbf{X}_i' \boldsymbol{\beta}^{(z)}, 1)$ truncated to $(-\infty, 0)$ if $Y_i = 0$ and to $(0, +\infty)$ if $Y_i = 1$.

2. Sample $\boldsymbol{\beta}^{(z)}$ from the multivariate Normal distribution $\mathcal{N}(\boldsymbol{\mu}_\beta^{(z)*}, \boldsymbol{\Sigma}_\beta^{(z)*})$, where

$$\boldsymbol{\mu}_\beta^{(z)*} = \boldsymbol{\Sigma}_\beta^{(z)*} \left(\boldsymbol{\Sigma}_\beta^{(z)-1} \boldsymbol{\mu}_\beta^{(z)} + \mathbf{X}_z' \mathbf{Y}_z^* \right) \quad \boldsymbol{\Sigma}_\beta^{(z)*} = \left(\boldsymbol{\Sigma}_\beta^{(z)-1} + \mathbf{X}_z' \mathbf{X}_z \right)^{-1}$$

where $\mathbf{X}_z$ denotes the sub-matrix of $\mathbf{X}$ and $\mathbf{Y}_z^*$ the sub-vector of $\mathbf{Y}^*$ for units with $Z_i = z$.

3. The value of the ITT estimator, $\widehat{\vartheta}_{ITT}$, is calculated by averaging over the empirical distribution of the covariates:

$$\frac{\sum_{i=1}^n \Phi(\mathbf{X}_i' \boldsymbol{\beta}^{(1)})}{n} - \frac{\sum_{i=1}^n \Phi(\mathbf{X}_i' \boldsymbol{\beta}^{(0)})}{n}.$$

Step 3 is only made for iterations following the burn-in period.

In the simulations, we run a single chain for 10000 iterations, with a burn-in period of 5000. Therefore, for each dataset r , we store 5000 $\widehat{\vartheta}_{ITT}^r$; we combine the latter by taking their average, thus obtaining a single point estimate for each dataset.

B.1.2 PP

In terms of potential outcomes, the super-population PP estimand can be defined as $\vartheta_{PP} \equiv \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$, i.e., the expected difference in the potential outcomes for a patient who adhered to the randomized treatment. Define $\pi_i^{(g|1)} \equiv p(G_i = g | D_i = 1)$ for $g \in \{11, 01\}$. In the simulations, the true value of the super-population PP estimand can be computed as

$$\begin{aligned} \vartheta_{PP} &= \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1] = \\ &\quad \mathbb{E}[Y_i(1) | D_i = 1] - \mathbb{E}[Y_i(0) | D_i = 1] = \\ &\quad \left\{ \mathbb{E}[Y_i(1) | D_i = 1, G_i = 11] \cdot \pi_i^{(11|1)} + \mathbb{E}[Y_i(1) | D_i = 1, G_i = 01] \cdot \pi_i^{(01|1)} \right\} \\ &\quad - \mathbb{E}[Y_i(0) | Z_i = 0, D_i = 1] = \\ &\quad \left\{ \mathbb{E}[Y_i(1) | D_i = 1, G_i = 11] \cdot \pi_i^{(11|1)} + \mathbb{E}[Y_i(1) | D_i = 1, G_i = 01] \cdot \pi_i^{(01|1)} \right\} \\ &\quad - \mathbb{E}[Y_i(0) | Z_i = 0, D_i = 1, G_i = 11] = \\ &\quad \left\{ \mathbb{E}[Y_i(1) | D_i = 1, G_i = 11] \cdot \pi_i^{(11|1)} + \mathbb{E}[Y_i(1) | D_i = 1, G_i = 01] \cdot \pi_i^{(01|1)} \right\} \\ &\quad - \mathbb{E}[Y_i(0) | D_i = 1, G_i = 11]. \end{aligned}$$

To perform the PP analysis, we use again Bayesian model-based inference. In particular, we specify the same models that we assumed for the ITT analysis (Section B.1.1), and use an analogous Gibbs sampling for obtaining posterior draws of the parameters. The only difference is that we exclude from the analysis patients with $D_i = 0$, i.e., patients who do not adhere to the assigned treatment. Therefore, the

PP estimator is the following:

$$\hat{\vartheta}_{PP} = \frac{\sum_{i:D_i=1} \Phi(\mathbf{X}_i' \boldsymbol{\beta}^{(1)})}{\sum_{i=1}^n \mathbb{1}(D_i = 1)} - \frac{\sum_{i:D_i=1} \Phi(\mathbf{X}_i' \boldsymbol{\beta}^{(0)})}{\sum_{i=1}^n \mathbb{1}(D_i = 1)}.$$

B.1.3 IPW

Let us now introduce a different potential outcome. In particular, let us define $Y_i^{hyp}(D_i(z) = 1)$ as the potential outcome of a randomly selected subject if he/she is assigned to treatment z and is able to adhere to it entirely. The IPW estimand is the following:

$$\vartheta_{IPW} = \mathbb{E}[Y_i^{hyp}(D_i(1) = 1) - Y_i^{hyp}(D_i(0) = 1)],$$

i.e., the mean difference in potential outcomes between treatment groups if all participants were to adhere. Let us denote with n_z the total number of patients randomized to the treatment condition z . The IPW estimator is

$$\hat{\vartheta}_{IPW} = \frac{\sum_{i:Z_i=1} Y_i}{n_1} - \frac{\sum_{i:Z_i=0 \text{ \& } D_i=1} (Y_i \cdot w_i)}{\sum_{i:Z_i=0 \text{ \& } D_i=1} w_i}, \quad (\text{B.1})$$

$$w_i = \frac{(\sum_{i:Z_i=0} D_i) / n_0}{p(D_i = 1 \mid Z_i = 0, \mathbf{X}_i)}. \quad (\text{B.2})$$

Note that the first term in Equation (B.1) is not weighted because every subject adheres to the experimental treatment. The second term, instead, is weighted by w_i because there is non-adherence in the control arm. The numerator of w_i in Equation (B.2) is the proportion of adherents among those randomized to the SoC, while the denominator is the estimated probability of adherence for a patient i assigned to the SoC given his or her covariates $\mathbf{X}_i$. This probability is estimated through a probit regression of D on the covariates among patients assigned to the SoC.

In the simulations, after obtaining $\hat{\vartheta}_{IPW}^r$ for each dataset r , we compute its SE and 95% confidence interval based on $B = 1000$ bootstrap repetitions.

B.1.4 Principal scores

We have already described the principal scores method in Section 3.2.6. Here we provide further details on how it has been used in the simulations.

Principal scores are estimated through a probit regression of D on the covariates among patients assigned to the SoC. This regression model is identical to the model used to estimate the weights in the IPW analysis. In the simulations, after obtaining $\widehat{\vartheta}_{pscores}^{(11)r}$ for each dataset r , we compute its SE and 95% confidence interval based on $B = 1000$ bootstrap repetitions.

B.1.5 Bayesian model-based PS analysis

We provide here further details on the two-stage Gibbs Sampling which has been used to approximate the posterior distribution of the parameters in the Bayesian model-based PS analysis.

Let $(\mathbf{G}^{(k)}, \boldsymbol{\psi}^{(k)})$ denote the state of the chain at iteration k , where $\mathbf{G}^{(k)}$ is the n -dimensional vector of imputed principal strata memberships at iteration k and $\boldsymbol{\psi}^{(k)}$ is the value of the parameters at iteration k , i.e., $\boldsymbol{\psi}^{(k)} = \{\boldsymbol{\alpha}^{(k)}, \boldsymbol{\eta}^{(0,11)(k)}, \boldsymbol{\eta}^{(1,11)(k)}, \boldsymbol{\eta}^{(0,01)(k)}, \boldsymbol{\eta}^{(1,01)(k)}\}$. The state of the chain at iteration $k + 1$ follows from applying the following steps:

1. Sample $\mathbf{G}^{(k+1)}$ according to $p(\mathbf{G} \mid \boldsymbol{\psi}^{(k)}, \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{D})$:

$$p(G_i = 11 \mid \boldsymbol{\psi}^{(k)}, Y_i, Z_i = 0, D_i = 1, \mathbf{X}_i) = 1.$$

$$p(G_i = 01 \mid \boldsymbol{\psi}^{(k)}, Y_i, Z_i = 0, D_i = 0, \mathbf{X}_i) = 1.$$

For observations with $Z_i = 1$:

$$p(G_i = 11 \mid \boldsymbol{\psi}^{(k)}, Y_i, Z_i = 1, D_i = 1, \mathbf{X}_i) \propto \frac{\pi_i^{(11)} \times f_i^{(1,11)}}{(\pi_i^{(01)} \times f_i^{(1,01)}) + (\pi_i^{(11)} \times f_i^{(1,11)})}.$$

2. The full conditional distributions of G_i^* is given by the following Truncated Normal ($\mathcal{TN}$) distributions:

$$G_i^* \mid G_i, \mathbf{X}_i; \boldsymbol{\alpha} \sim \begin{cases} \mathcal{TN}(\mathbf{X}_i' \boldsymbol{\alpha}, 1, \text{lower} = -\infty, \text{upper} = 0) & \text{if } G_i \neq 11 \\ \mathcal{TN}(\mathbf{X}_i' \boldsymbol{\alpha}, 1, \text{lower} = 0, \text{upper} = +\infty) & \text{if } G_i = 11 \end{cases}$$

Thus, we sample the latent variable G_i^* from $\mathcal{N}(\mathbf{X}_i' \boldsymbol{\alpha}, 1)$ truncated to $(-\infty, 0)$ if $G_i \neq 11$, and to $(0, +\infty)$ if $G_i = 11$.

3. Sample $\boldsymbol{\alpha}$ from the multivariate Normal distribution $\mathcal{N}(\boldsymbol{\mu}_\alpha^*, \boldsymbol{\Sigma}_\alpha^*)$, where

$$\boldsymbol{\mu}_\alpha^* = \boldsymbol{\Sigma}_\alpha^* (\boldsymbol{\Sigma}_\alpha^{-1} \boldsymbol{\mu}_\alpha + \mathbf{X}' \mathbf{G}^*) \quad \boldsymbol{\Sigma}_\alpha^* = (\boldsymbol{\Sigma}_\alpha^{-1} + \mathbf{X}' \mathbf{X})^{-1}$$

where $\mathbf{G}^*$ is the n -dimensional vector with i th element equal to G_i^* .

4. The full conditional distribution of $Y_i^*(z)$ is:

$$Y_i^*(z) \mid Y_i, \mathbf{X}_i, G_i = g; \boldsymbol{\eta}^{(z,g)} \sim \begin{cases} \mathcal{TN}(\mathbf{X}_i' \boldsymbol{\eta}^{(z,g)}, 1, \text{lower} = 0, \text{upper} = +\infty) & \text{if } Y_i = 1 \\ \mathcal{TN}(\mathbf{X}_i' \boldsymbol{\eta}^{(z,g)}, 1, \text{lower} = -\infty, \text{upper} = 0) & \text{if } Y_i = 0 \end{cases}$$

Thus, we sample the latent variable $Y_i^*(z)$ from $\mathcal{N}(\mathbf{X}_i' \boldsymbol{\eta}^{(z,g)}, 1)$ truncated to $(-\infty, 0)$ if $Y_i = 0$ and to $(0, +\infty)$ if $Y_i = 1$.

5. Sample $\boldsymbol{\eta}^{(z,g)}$ from the multivariate Normal distribution $\mathcal{N}(\boldsymbol{\mu}_\eta^{(z,g)*}, \boldsymbol{\Sigma}_\eta^{(z,g)*})$, where

$$\boldsymbol{\mu}_\eta^{(z,g)*} = \boldsymbol{\Sigma}_\eta^{(z,g)*} \left(\boldsymbol{\Sigma}_\eta^{(z,g)-1} \boldsymbol{\mu}_\eta^{(z,g)} + \mathbf{X}_{gz}' \mathbf{Y}_{gz}^* \right) \quad \boldsymbol{\Sigma}_\eta^{(z,g)*} = \left(\boldsymbol{\Sigma}_\eta^{(z,g)-1} + \mathbf{X}_{gz}' \mathbf{X}_{gz} \right)^{-1}$$

where $\mathbf{X}_{gz}$ denotes the sub-matrix of $\mathbf{X}$ and $\mathbf{Y}_{gz}^*$ the sub-vector of $\mathbf{Y}^*$ for units with $G_i^{(k+1)} = g$ and $Z_i = z$.

6. Calculate the value of the Bayesian model-based estimator of the AACE, $\widehat{\vartheta}_{Bayes}^{(11)}$, and of the NACE, $\widehat{\vartheta}_{Bayes}^{(01)}$, as

$$\frac{\sum_{i=1}^n \left[\pi_i^{(g)} \times \Phi(\mathbf{X}_i' \boldsymbol{\eta}^{(1,g)}) \right]}{\sum_{i=1}^n \pi_i^{(g)}} - \frac{\sum_{i=1}^n \left[\pi_i^{(g)} \times \Phi(\mathbf{X}_i' \boldsymbol{\eta}^{(0,g)}) \right]}{\sum_{i=1}^n \pi_i^{(g)}}$$

with $g = 11$ and $g = 01$, respectively. Step 6 is only made for iterations following the burn-in period. In the simulations, we set the total number of iterations at 150'000, discarding the first 75'000 as burn-in and saving every 20th. We run two independent chains from different starting values. Therefore, for each dataset r , we store 7500 $\widehat{\vartheta}_{Bayes}^{(g)r}$; we combine the latter by taking their average, thus obtaining a single point estimate for each dataset.

B.2 Convergence checks

We use the $\hat{R}$ convergence diagnostic to assess the convergence of the MCMC algorithm. Again, we refer to the maximum of rank normalized split- $\hat{R}$ and rank normalized folded-split- $\hat{R}$, as proposed by Vehtari et al. (2021). We assess the value of this diagnostic for the model parameters in the ITT, PP, and Bayesian model-based PS analyses. We perform this check both for the analyses of the data simulated by using strong predictors of adherence (see Section 3.3) and for those simulated by using much weaker predictors (see Section 3.4.3). It is worth repeating that in the former case, the AUROC is equal to 0.9022, while in the latter it is equal to 0.7029.

Regardless of the value of the AUROC, the $\hat{R}$ diagnostic takes on values lower than 1.01 for all the model parameters in the ITT and PP analyses, for all n_{sim} repetitions of the dataset and for all DGMs considered.

The results of the Bayesian model-based PS analysis, on the other hand, require a more in-depth discussion. In the presence of strong predictors of adherence, the $\hat{R}$ diagnostic takes on values lower than 1.01 for all model parameters, for all the n_{sim} repetitions of the dataset and for all DGMs considered. This, together with the visualization of the trace plots, indicates that convergence has been reached. Furthermore, the posterior distributions of the model parameters are generally well-shaped. For the sake of brevity, we do not report the respective plots here.

In the presence of weak predictors of adherence, the $\hat{R}$ diagnostic indicates that convergence has been reached in the vast majority of cases. To ease the explanation, let us now refer specifically to the datasets simulated from the DGM where $\eta_0^{(1,01)} = 0.6$ (see Table 3.4). The results show that for 11.4% of these datasets (57 out of 500) the $\hat{R}$ diagnostic takes on values greater or equal to 1.01 for at least one parameter. This indicates that, in these cases, the chains have not mixed well¹. To allow a more in-depth understanding of these results, in Figures B.1, B.2 and B.3 we show through trace plots the problematic chains for the first 8 datasets with convergence issues².

¹When the datasets are simulated from the DGMs with $\eta_0^{(1,01)} = 0.4$ and $\eta_0^{(1,01)} = 0.2$, this percentage amounts, respectively, to 11.0% and 10.4%. The issue is analogous to that discussed for datasets simulated from the DGM where $\eta_0^{(1,01)} = 0.6$.

²We do not show results for all datasets with convergence problems for ease of reading. The results relative to the first 8 problematic datasets are representative of the results relative to the rest of the datasets with convergence issues.

FIGURE B.1: Problematic trace plots (i.e., trace plots of the parameters for which the $\hat{R}$ diagnostic takes on values greater or equal to 1.01), part I. The results are from the analysis of the datasets simulated by using weak predictors of adherence, and specifically from the DGM where $\eta_0^{(1,01)} = 0.6$ (see Table 3.4).

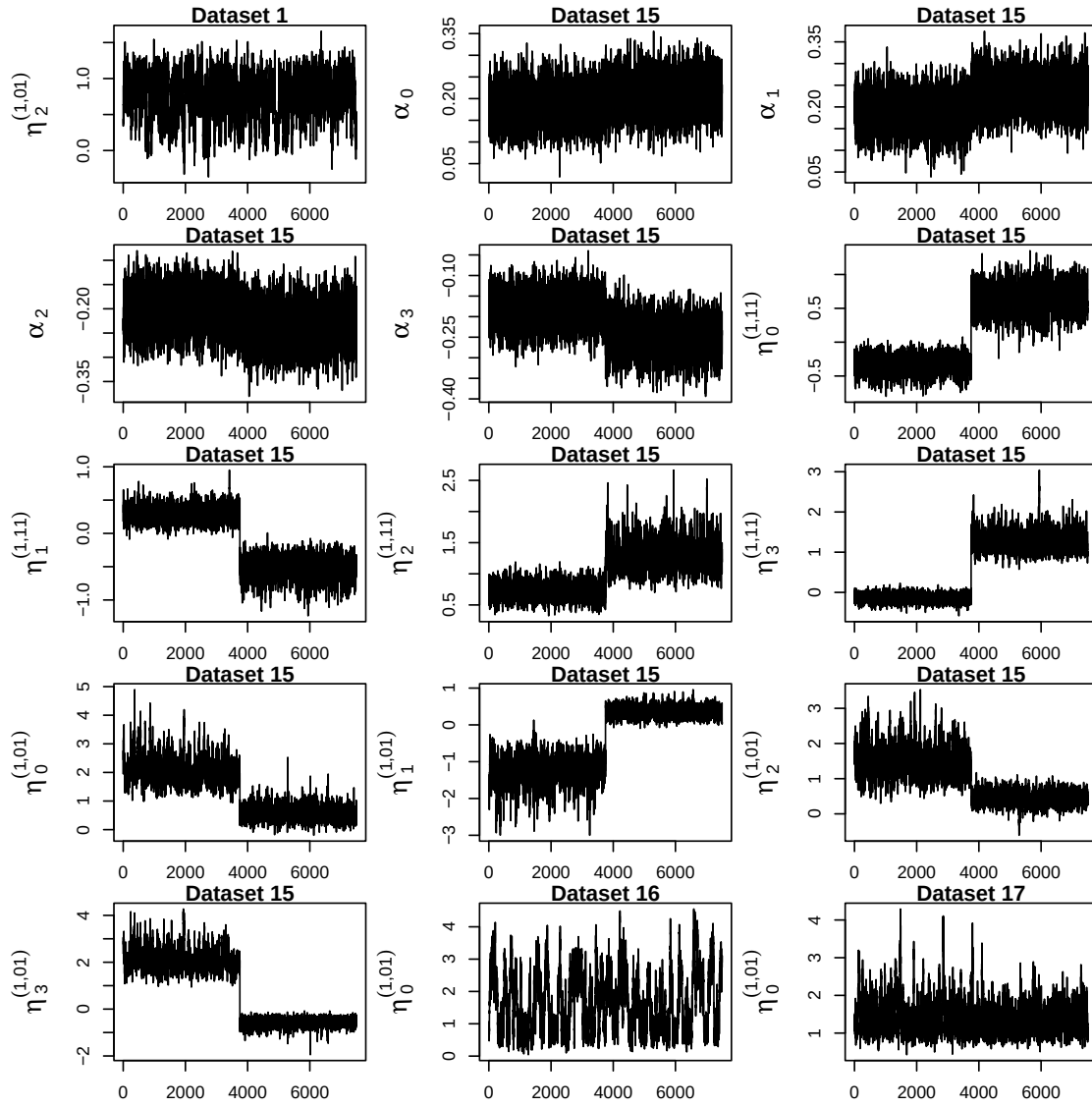


FIGURE B.2: Problematic trace plots (i.e., trace plots of the parameters for which the $\hat{R}$ diagnostic takes on values greater or equal to 1.01), part II. The results are from the analysis of the datasets simulated by using weak predictors of adherence, and specifically from the DGM where $\eta_0^{(1,01)} = 0.6$ (see Table 3.4).

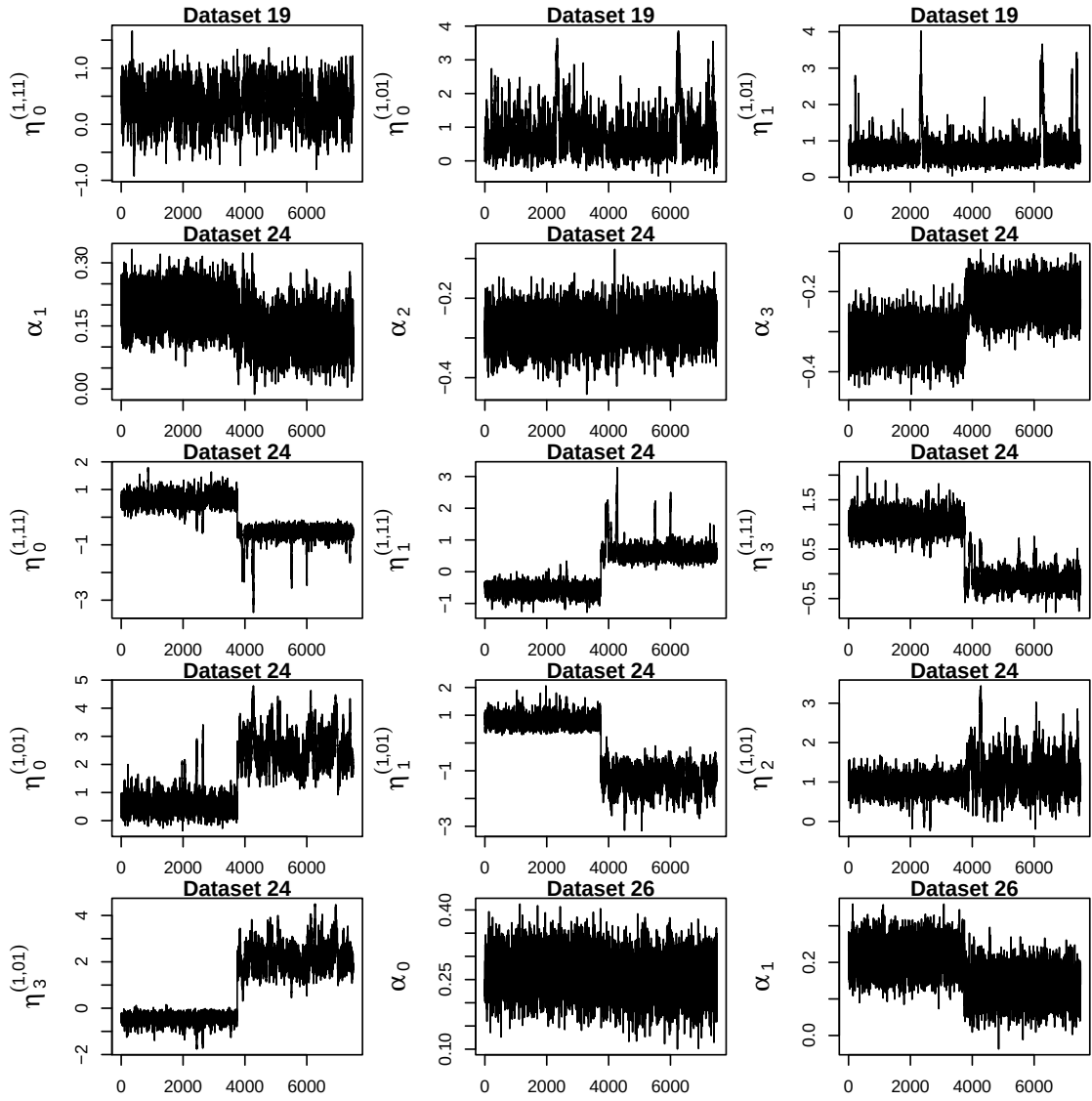
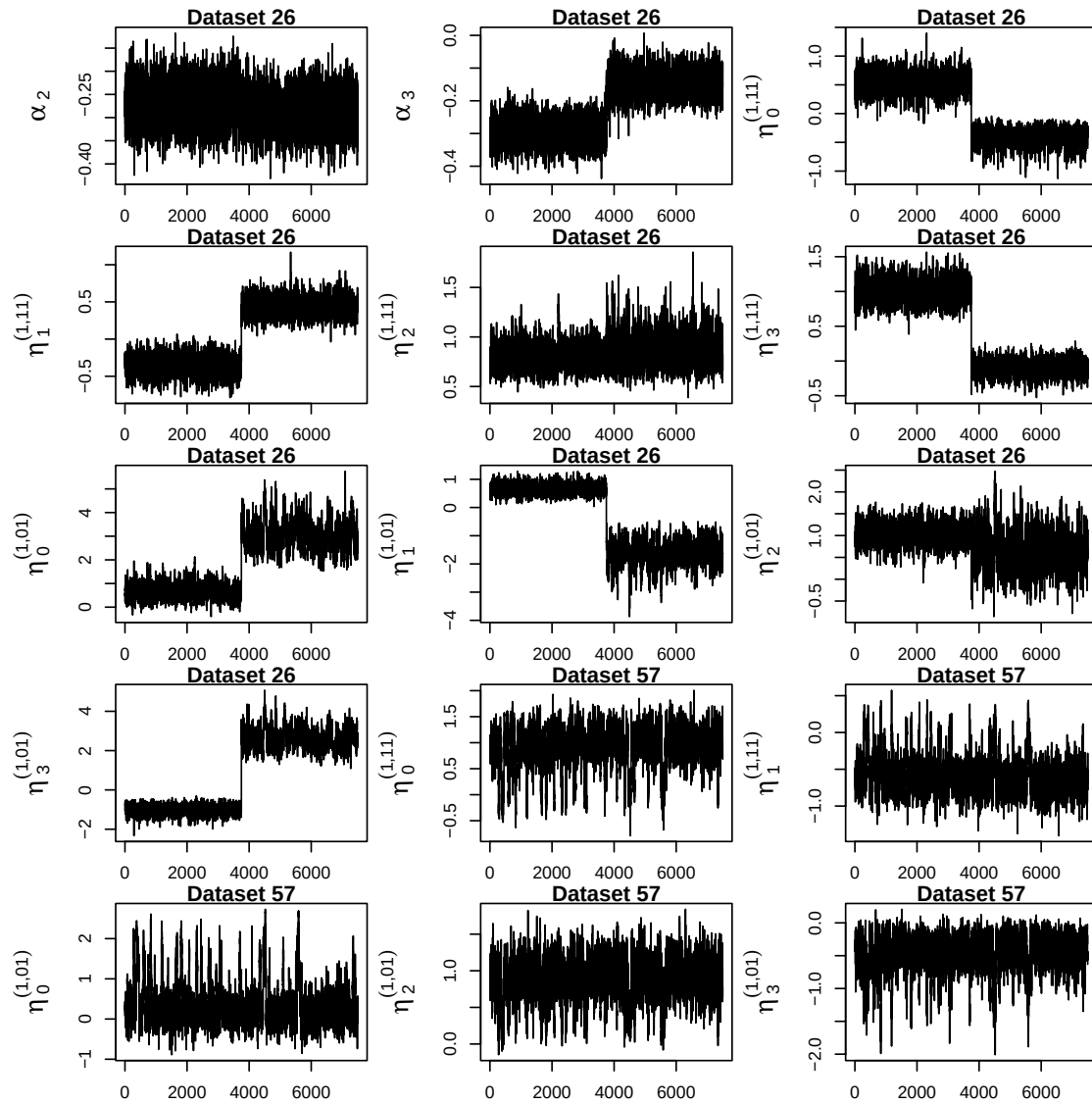


FIGURE B.3: Problematic trace plots (i.e., trace plots of the parameters for which the $\hat{R}$ diagnostic takes on values greater or equal to 1.01), part III. The results are from the analysis of the datasets simulated by using weak predictors of adherence, and specifically from the DGM where $\eta_0^{(1,01)} = 0.6$ (see Table 3.4).



As we can see from these Figures, poor convergence is always observed for one or more parameters in $\boldsymbol{\eta}^{(1,g)}$ for $g \in \{11, 01\}$, and less frequently for one or more parameters in the principal stratum model. It should be noticed that the parameters in $\boldsymbol{\eta}^{(1,g)}$ are involved in the mixture (i.e., principal stratum membership is unknown for those in the experimental arm). This suggests that the issue may arise from the greater uncertainty in the patients' attribution to the principal strata, which in turn is due to the weak predictors of adherence used. Thus, we do not expect that the issue can be solved by increasing the total number of iterations (which is now 150'000).

Confirming this, from the visualization of the chains it is evident that some of them are even stuck at different modes, therefore clearly indicating multimodality of the posterior distributions. This is confirmed by the estimated Bulk Effective Sample Size (Bulk-ESS), which is a measure of sampling efficiency in the bulk of the distribution (related, for example, to the efficiency of mean and median estimates). Bulk-ESS should be at least 100 (approximately) per chain in order to indicate that estimates of respective posterior quantiles are reliable (Vehtari et al., 2021). Here, since we run two independent chains from different starting values, the Bulk-ESS should be at least 200. However, in the case of multimodal distributions, the estimated Bulk-ESS is usually close to the number of distinct modes that are found. In Table 2, we show the estimated Bulk-ESS referred to the problematic chains (i.e., those shown in Figures B.1, B.2 and B.3). As is evident from the Table, some values are lower than the threshold. Above all, the estimated Bulk-ESS referred to the chains of the parameters for datasets 15, 24, and 26 is very low, indicating multimodality of the posterior distributions. Also, the visualization of the posterior distributions confirms that they are generally not well-shaped.

In conclusion, the problem occurs in a minority of datasets and concerns a few parameters. However, the risk of this problem occurring when analyzing weakly informative data on treatment adherence must be acknowledged. This is not a point against using the methodology, since poor convergence can easily be detected if multiple chains are run (possibly in parallel not to increase the required computational time) from different starting values.

TABLE 2: Estimated Bulk-ESS referred to the problematic chains (i.e., those shown in Figures B.1, B.2 and B.3)

	Dataset							
	1	15	16	17	19	24	26	57
α_0	-	21	-	-	-	-	44	-
α_1	-	7	-	-	-	5	4	-
α_2	-	12	-	-	-	79	39	-
α_3	-	4	-	-	-	3	3	-
$\eta_0^{(1,11)}$	-	3	-	-	472	3	3	331
$\eta_1^{(1,11)}$	-	3	-	-	-	3	3	361
$\eta_2^{(1,11)}$	-	3	-	-	-	-	51	-
$\eta_3^{(1,11)}$	-	3	-	-	-	3	3	-
$\eta_0^{(1,01)}$	-	3	121	481	241	3	3	325
$\eta_1^{(1,01)}$	-	3	-	-	339	3	3	-
$\eta_2^{(1,01)}$	363	3	-	-	-	16	10	379
$\eta_3^{(1,01)}$	-	3	-	-	-	3	3	379

Appendix C

Appendix to the Florentine Study

C.1 Convergence checks

We use the $\hat{R}$ convergence diagnostic to assess the convergence of the MCMC algorithm. Again, we refer to the maximum of rank normalized split- $\hat{R}$ and rank normalized folded-split- $\hat{R}$, as proposed by Vehtari et al. (2021). We assess the value of this $\hat{R}$ diagnostic for all the model parameters. The statistic always takes on values lower than 1.01, showing no evidence against convergence.

The trace plots of the principal stratum model parameters are shown in Figure C.1. Those of the potential outcomes model parameters are instead shown in Figures C.2 and C.3. In these Figures, we show only the trace plots of the models' parameters for Never Takers and Reward Compliers ($g \in \{000, 001\}$), because we focus on the results for these principal strata (which are the two largest ones). From the visualization of these trace plots, we can confirm that convergence has been reached.

FIGURE C.1: Trace plots of the principal stratum model parameters

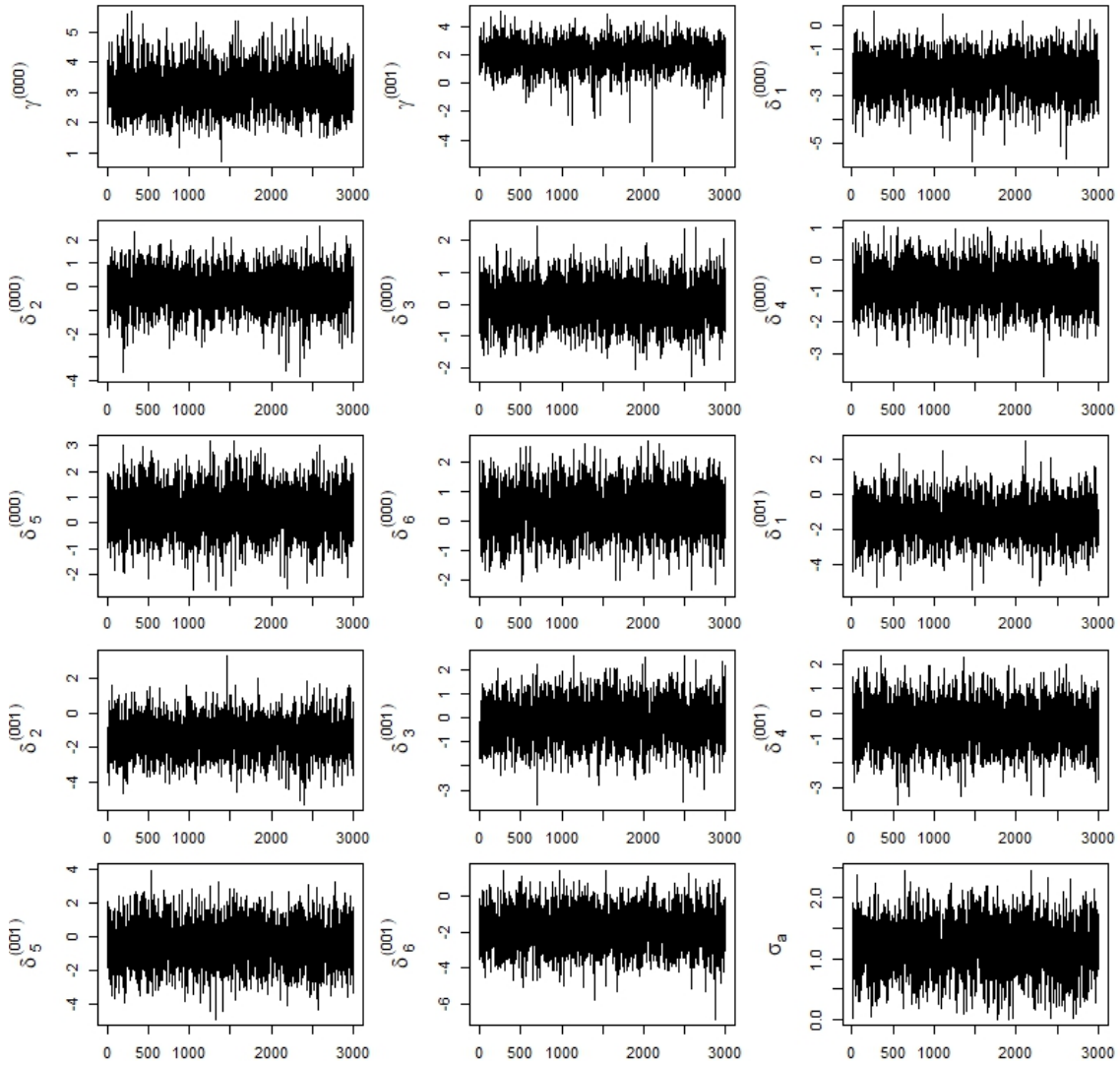


FIGURE C.2: Trace plots of the potential outcomes model parameters (part I)

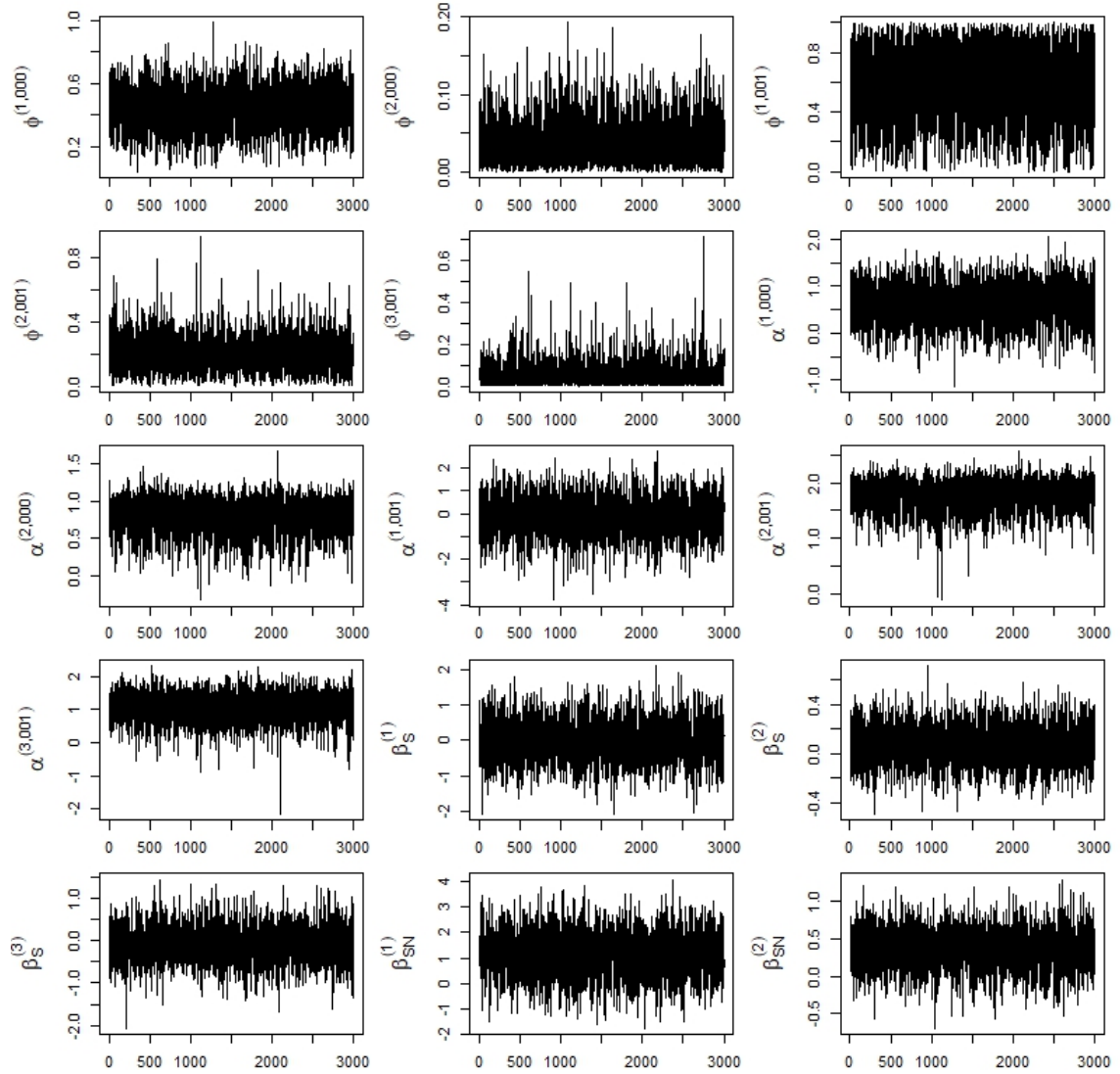
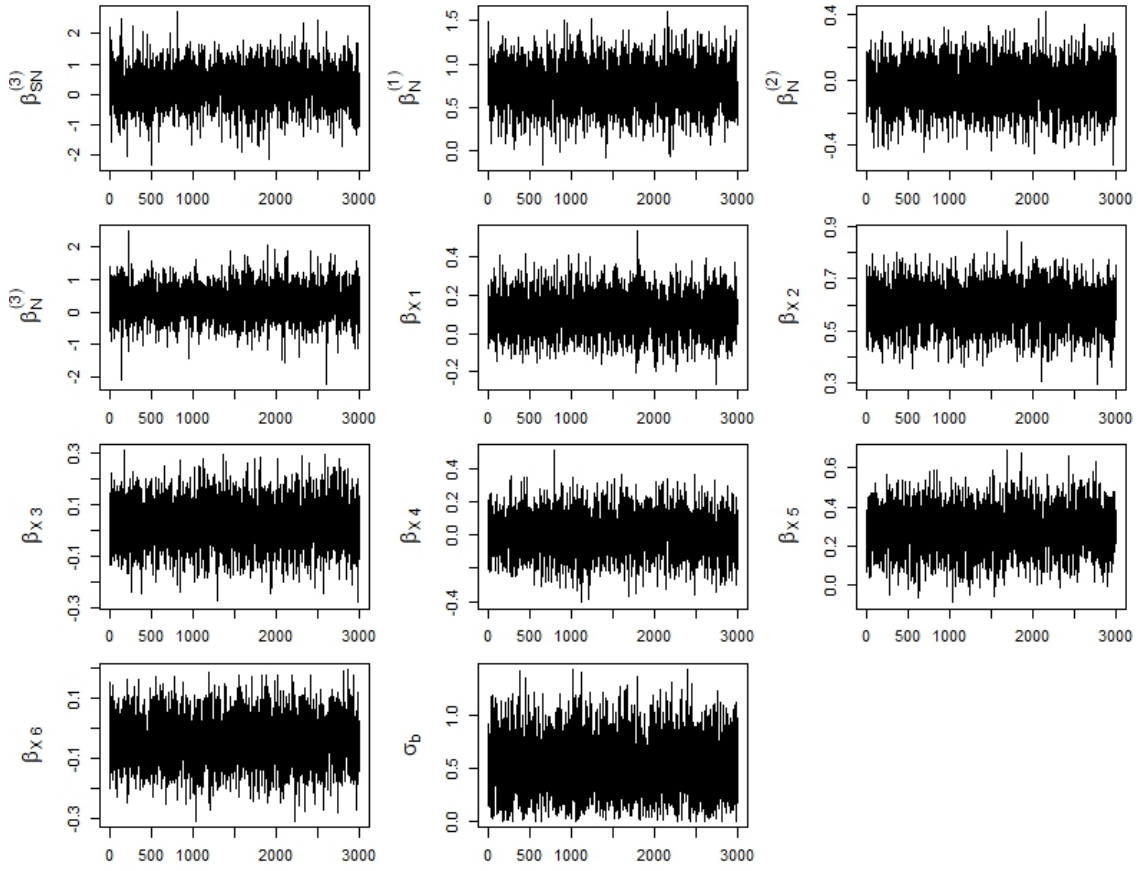


FIGURE C.3: Trace plots of the potential outcomes model parameters (part II)



Bibliography

- Adebayo, E. F., Uthman, O. A., Wiysonge, C. S., Stern, E. A., Lamont, K. T., & Ataguba, J. E. (2015). A systematic review of factors that affect uptake of community-based health insurance in low-income and middle-income countries. *BMC health services research*, *15*(1), 1–13.
- Akacha, M., Bretz, F., & Ruberg, S. (2017). Estimands in clinical trials—broadening the perspective. *Statistics in medicine*, *36*(1), 5–19.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455.
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Barnard, J., Frangakis, C. E., Hill, J. L., & 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*(462), 299–323.
- Berndt, T. J., & Murphy, L. M. (2002). Influences of friends and friendships: Myths, truths, and research recommendations. *Advances in child development and behavior*, *30*, 275–310.
- Bielicki, J. A., Stöhr, W., Barratt, S., Dunn, D., Naufal, N., Roland, D., Sturgeon, K., Finn, A., Rodriguez-Ruiz, J. P., Malhotra-Kumar, S., et al. (2021). Effect of amoxicillin dose and treatment duration on the need for antibiotic re-treatment in children with community-acquired pneumonia: The cap-it randomized clinical trial. *JAMA*, *326*(17), 1713–1724.
- Biggeri, M., Nannini, M., & Putoto, G. (2018). Assessing the feasibility of community health insurance in uganda: A mixed-methods exploratory analysis. *Social science & medicine*, *200*, 145–155.
- Brown, C. D., & Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, *80*(1), 24–38.
- Burke, J. F., Sussman, J. B., Kent, D. M., & Hayward, R. A. (2015). Three simple rules to ensure reasonably credible subgroup analyses. *Bmj*, *351*.

- Burzykowski, T., Buyse, M., & Molenberghs, G. (2005). *The evaluation of surrogate endpoints* (Vol. 427). Springer.
- Cellini, R., & Cuccia, T. (2018). How free admittance affects charged visits to museums: An analysis of the Italian case. *Oxford Economic Papers*, 70(3), 680–698.
- Chemin, M. (2018). Informal groups and health insurance take-up evidence from a field experiment. *World Development*, 101, 54–72.
- Cochran, W. G. (1963). *Sampling techniques*. 2nd edition. John Wiley & Sons.
- Cochran, W. G. (1972). Observational studies. *Statistical papers in honor of George W. Snedecor*, 77–90.
- Copas, J. B., & Li, H. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1), 55–95.
- Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1), 173–203.
- Cox, D. (1958). *Planning of experiments*. Wiley.
- Cro, S., Kahan, B. C., Rehal, S., Ster, A. C., Carpenter, J. R., White, I. R., & Cornelius, V. R. (2022). Evaluating how clear the questions being investigated in randomised trials are: Systematic review of estimands. *bmj*, 378.
- Del Prete, D., Forastiere, L., & Sciabolazza, V. L. (2020). Causal inference on networks under continuous treatment interference. *arXiv preprint arXiv:2004.13459*.
- Ding, P., & Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3), 757–777.
- Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27(3), 368–377.
- Dodd, M., Fielding, K., Carpenter, J. R., Thompson, J. A., & Elbourne, D. (2022). Statistical methods for non-adherence in non-inferiority trials: useful and used? A systematic review. *BMJ open*, 12(1), e052656.
- Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413), 9–17.
- Elliott, M. R., Raghunathan, T. E., & Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, 11(2), 353–372.
- European Medicines Agency. (2020). ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials [<https://www.ema.europa.eu/en/documents/scientific->

- [guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf](#)].
- Feller, A., Mealli, F., & Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, 42(6), 726–758.
- Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7), 799–815.
- Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923), 554.
- Food and Drug Administration. (2016). Non-inferiority clinical trials to establish effectiveness guidance for industry. *US: Department of Health and Human Services*.
- Forastiere, L., Airoidi, E. M., & Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534), 901–918.
- Forastiere, L., Lattarulo, P., Mariani, M., Mealli, F., & Razzolini, L. (2021). Exploring encouragement, treatment, and spillover effects using principal stratification, with application to a field experiment on teens' museum attendance. *Journal of Business & Economic Statistics*, 39(1), 244–258.
- Forastiere, L., Mealli, F., & VanderWeele, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using bayesian principal stratification. *Journal of the American Statistical Association*, 111(514), 510–525.
- Forastiere, L., Mealli, F., Wu, A., & Airoidi, E. (2018). Estimating causal effects under interference using bayesian generalized propensity scores. *arXiv preprint arXiv:1807.11038*.
- Frangakis, C. E. (2006). Causal effects in the presence of non compliance: A latent variable interpretation (comment on article by Forcina). *Metron*, 64, 282–288.
- Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21–29.
- Frangakis, C. E., Rubin, D. B., & Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, 3(2), 147–164.
- Frumento, P., Mealli, F., Pacini, B., & Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498), 450–466.

- Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., & Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in medicine*, 28(7), 1108–1130.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515–534.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.
- Gilbert, P. B., & Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4), 1146–1154.
- Goldsmith-Pinkham, P., & Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3), 253–264.
- Gustafson, P. (2010). Bayesian inference for partially identified models. *The international journal of biostatistics*, 6(2).
- Hernán, M. A. (2021). Methods of public health research—strengthening causal inference from observational data. *New England Journal of Medicine*, 385(15), 1345–1348.
- Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 21(4), 601–627.
- Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1), 69–88.
- Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1), 1593–1623.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396), 945–960.
- Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475), 901–910.
- Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 832–842.
- Hughes, K., & Moscardo, G. (2019). For me or not for me? exploring young adults' museum representations. *Leisure Sciences*, 41(6), 516–534.

- Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with “truncation-by-death”. *Statistics & probability letters*, 78(2), 144–149.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2), 126–132.
- Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–475.
- Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The annals of statistics*, 305–327.
- Imbens, G. W., & Rubin, D. B. (2016). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Taylor & Francis.
- Jiang, Z., & Ding, P. (2021). Identification of causal effects within principal strata using auxiliary variables. *Statistical Science*, 36(4), 493–508.
- Jiang, Z., Ding, P., & Geng, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4), 829–848.
- Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481), 101–111.
- Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in medicine*, 28(23), 2857–2875.
- Jütting, J. P. (2004). Do community-based health insurance schemes improve poor people’s access to health care? evidence from rural senegal. *World development*, 32(2), 273–288.
- Kang, H., & Imbens, G. (2016). Peer encouragement designs in causal inference with partial interference and identification of local average network effects. *arXiv preprint arXiv:1609.04464*.
- Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statistical modelling*, 3(3), 193–213.
- Kisida, B., Greene, J. P., & Bowen, D. H. (2014). Creating cultural consumers: The dynamics of cultural capital acquisition. *Sociology of Education*, 87(4), 281–295.
- Kruk, M. E., Goldmann, E., & Galea, S. (2009). Borrowing and selling to pay for health care in low-and middle-income countries. *Health Affairs*, 28(4), 1056–1066.
- Lattarulo, P., Mariani, M., & Razzolini, L. (2017). Nudging museums attendance: A field experiment with high school teens. *Journal of Cultural Economics*, 41(3), 259–277.

- Leive, A., & Xu, K. (2008). Coping with out-of-pocket health payments: Empirical evidence from 15 african countries. *Bulletin of the World Health Organization*, 86(11), 849–856C.
- Li, F., Tian, Z., Tian, Z., & Li, F. (2022). A note on identification of causal effects in cluster randomized trials with post-randomization selection bias. *Communications in Statistics-Theory and Methods*, 1–13.
- Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 948–963.
- Lipkovich, I., Ratitch, B., Qu, Y., Zhang, X., Shan, M., & Mallinckrodt, C. (2022). Using principal stratification in analysis of clinical trials. *Statistics in Medicine*, 41(19), 3837–3877.
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
- Lou, Y., Jones, M. P., & Sun, W. (2019). Estimation of causal effects in clinical endpoint bioequivalence studies in the presence of intercurrent events: Non-compliance and missing data. *Journal of biopharmaceutical statistics*, 29(1), 151–173.
- Manna, R., & Palumbo, R. (2018). What makes a museum attractive to young people? evidence from italy. *International Journal of Tourism Research*, 20(4), 508–517.
- Marks, M., Mitjà, O., Bottomley, C., Kwakye, C., Houine, W., Bauri, M., Adwere, P., Abdulai, A. A., Dua, F., Boateng, L., et al. (2018). Comparative efficacy of low-dose versus standard-dose azithromycin for patients with yaws: A randomised non-inferiority trial in ghana and papua new guinea. *The Lancet Global Health*, 6(4), e401–e410.
- Matilde Sanchez, M., & Chen, X. (2006). Choosing the analysis population in non-inferiority studies: Per protocol or intent-to-treat. *Statistics in medicine*, 25(7), 1169–1181.
- Mattei, A., Li, F., & Mealli, F. (2013). Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, 7(4), 2336–2360.
- Mattei, A., & Mealli, F. (2007). Application of the principal stratification approach to the faenza randomized experiment on breast self-examination. *Biometrics*, 63(2), 437–446.
- Mattei, A., & Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5), 729–752.

- Mattei, A., Mealli, F., & Ding, P. (2020). Assessing causal effects in the presence of treatment switching through principal stratification. *arXiv preprint arXiv:2002.11989*.
- Mattei, A., Mealli, F., & Pacini, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics*, 70(2), 278–288.
- Mauri, L., & D’Agostino Sr, R. B. (2017). Challenges in the design and interpretation of noninferiority trials. *New England Journal of Medicine*, 377(14), 1357–1367.
- Mealli, F., & Mattei, A. (2012). A refreshing account of principal stratification. *The international journal of biostatistics*, 8(1).
- Mealli, F., & Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108(503), 1120–1131.
- Mealli, F., & Rubin, D. B. (2002). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services and Outcomes Research Methodology*, 3(3), 225–232.
- Meng, X.-L. (1994). Posterior predictive p -values. *The annals of statistics*, 22(3), 1142–1160.
- Mladovsky, P., Soors, W., Ndiaye, P., Ndiaye, A., & Criel, B. (2014). Can social capital help explain enrolment (or lack thereof) in community-based health insurance? results of an exploratory mixed methods study from senegal. *Social science & medicine*, 101, 18–27.
- Morris, S. S., Flores, R., Olinto, P., & Medina, J. M. (2004). Monetary incentives in primary health care and effects on use and coverage of preventive health care interventions in rural honduras: Cluster randomised trial. *The Lancet*, 364(9450), 2030–2037.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11), 2074–2102.
- Mulvenna, P., Nankivell, M., Barton, R., Faivre-Finn, C., Wilson, P., McColl, E., Moore, B., Brisbane, I., Ardron, D., Holt, T., et al. (2016). Dexamethasone and supportive care with or without whole brain radiotherapy in treating patients with non-small cell lung cancer with brain metastases unsuitable for resection or stereotactic radiotherapy (quartz): Results from a phase 3, non-inferiority, randomised trial. *The Lancet*, 388(10055), 2004–2014.
- Nagelkerke, N., Fidler, V., Bernsen, R., & Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in medicine*, 19(14), 1849–1864.

- Nannini, M., Biggeri, M., & Putoto, G. (2021). Financial protection and coping strategies in rural uganda: An impact evaluation of community-based zero-interest healthcare loans. *Health Policy and Planning*, 36(7), 1090–1102.
- Ogburn, E. L., Sofrygin, O., Diaz, I., & Van Der Laan, M. J. (2017). Causal inference for social network data. *arXiv preprint arXiv:1705.08527*.
- Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical papers*, 51(1), 111–126.
- Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6), 1609–1623.
- Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, 2001, 411–420.
- Piaggio, G., Elbourne, D. R., Pocock, S. J., Evans, S. J. W., & Altman, D. G. (2012). Reporting of noninferiority and equivalence randomized trials: Extension of the CONSORT 2010 statement. *JAMA*, 308(24), 2594–2604.
- Qu, Y., Fu, H., Luo, J., & Ruberg, S. J. (2020). A general framework for treatment effect estimators considering patient adherence. *Statistics in Biopharmaceutical Research*, 12(1), 1–18.
- Revicki, D. A., & Frank, L. (1999). Pharmacoeconomic evaluation in the real world: Effectiveness versus efficacy studies. *Pharmacoeconomics*, 15, 423–434.
- Righini, M., Le Gal, G., Aujesky, D., Roy, P.-M., Sanchez, O., Verschuren, F., Rutschmann, O., Nonent, M., Cornuz, J., Thys, F., et al. (2008). Diagnosis of pulmonary embolism by multidetector ct alone or combined with venous ultrasonography of the leg: A randomised non-inferiority trial. *The Lancet*, 371(9621), 1343–1352.
- Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143–155.
- Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of internal medicine*, 115(11), 901–905.
- Rosenbaum, P. R. (2010). *Design of observational studies* (Vol. 10). Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2), 212–218.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.
- Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *The Proceedings of the social statistics section of the American Statistical Association*, 233, 239.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1), 1–26.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, (6), 34–58.
- Rubin, D. B. (1979). Discussion of “conditional independence in statistical theory” by ap dawid. *J. Roy. Statist. Soc. Ser. B*, 41, 27–28.
- Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371), 591–593.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.
- Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3), 279–292.
- Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with “censoring” due to death. *Statistical Science*, 299–309.
- Schochet, P. Z., & Burghardt, J. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Evaluation Review*, 31(2), 95–120.
- Schumi, J., & Wittes, J. T. (2011). Through the looking glass: Understanding non-inferiority. *Trials*, 12(1), 1–12.
- Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476), 1398–1407.
- Sommerfeld, J., Sanon, M., Kouyate, B. A., & Sauerborn, R. (2002). Informal risk-sharing arrangements (irsas) in rural burkina faso: Lessons for the development of community-based insurance (cbi). *The International journal of health planning and management*, 17(2), 147–163.
- Stan Development Team. (2014). Stan: A c++ library for probability and sampling. Online: <http://mc-stan.org>.
- Stuart, E. A., & Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Statistical methods in medical research*, 24(6), 657–674.
- Suda, K. J., Hurley, A. M., McKibbin, T., & Moroney, S. E. M. (2011). Publication of noninferiority clinical trials: Changes over a 20-year interval. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 31(9), 833–839.
- Ten Have, T. R., & Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research*, 21(1), 77–107.

- Turkova, A., Moore, C. L., Butler, K., Compagnucci, A., Saïdi, Y., Musiime, V., Nanduudu, A., Kaudha, E., Cressey, T. R., Chalermpanmetagul, S., et al. (2018). Weekends-off efavirenz-based antiretroviral therapy in hiv-infected children, adolescents and young adults (breather): Extended follow-up results of a randomised, open-label, non-inferiority trial. *PloS one*, 13(4), e0196239.
- van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), 1–26.
- VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*, 78(17), 2957–2962.
- VanderWeele, T. J., & An, W. (2013). Social networks and causal inference. *Handbook of causal analysis for social research*, 353–374.
- VanderWeele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 42–52.
- VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the e-value. *Annals of internal medicine*, 167(4), 268–274.
- VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4), 457–468.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved r for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2), 667–718.
- Wagstaff, A., Bilger, M., Sajaia, Z., & Lokshin, M. (2011). *Health equity and financial protection: Streamlined analysis with adept software*. World Bank Publications.
- Wagstaff, A., Flores, G., Hsu, J., Smitz, M.-F., Chepynoga, K., Buisman, L. R., van Wilgenburg, K., & Eozenou, P. (2018). Progress on catastrophic health spending in 133 countries: A retrospective observational study. *The Lancet Global Health*, 6(2), e169–e179.
- Zhang, J. L., & Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics*, 28(4), 353–368.
- Zhang, J. L., Rubin, D. B., & Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485), 166–176.

-
- Zigler, C. M., & Belin, T. R. (2012). A bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics*, 68(3), 922–932.