DOTTORATO DI RICERCA IN STATISTICA APPLICATA

Ciclo: XXVII

Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: A Bayesian Principal Stratification Approach

Laura Forastiere

Director of Graduate Studies: Fabio Corradi
Advisor: Fabrizia Mealli
Co-Advisor: Alessandra Mattei

Contents

Preface

Cause-and-effect questions are the motivation for much research in many branches of science and public policy. For example, one might be interested in determining whether poverty causes early death, a job training programs improves prisoners behavior after release or whether legalization increases drug consumption. The art of making a causal claim about the relationship between two factors is central to how we view and react to the world around us, to our decision making, and to the advancement of science. We care about causal inference because, ultimately, we want to intervene to improve our lives, and interventions can be targeted on adding known causes of beneficial outcomes (or removing known causes of adverse outcomes).

Causality has been a pivotal concept in the history of philosophy since the time of the Ancient Greeks. After the middle ages, where western philosophers were still debating on the interaction between the primary cause, identified as God, and secondary causes, given by human decisions and laws of nature, in the sixteenth century the concept of 'homo faber', underlying humanism and the following enlightenment movements as well as eastern philosophies, laid the foundations of intervention-based causality. From the Buddhist point of view, our actions are the principal cause that will lead to happiness or suffering for ourselves or others. In the modern age, this concept has been interiorized and individuals as well as social institutions are continuously involved in a careful decision-making process of selecting optimal actions, in accordance to their final objectives. To this end, an understanding of the causal effects of any possible action and the ability of predicting the optimal one is crucial. This requires an emphasis on the effect of causes rather than on the causes of the effects.

The one field most suited to address such problems is the field of statistics, which makes use of inductive procedures to draw inferences from the observed world. However, since the time of Hume, many have questioned whether there is any metaphys-

ical meaning of causality, or valid inferences based upon it. In the eighteenth century Hume began the modern tradition of regularity models by defining causation in terms of repeated "conjunctions" of events. Hume argued that the labeling of two particular events as being causally related rested on an untestable metaphysical assumption. Under this framework, causation was defined purely in terms of empirical criteria, rather than unobservable assumptions. Due to its well-known difficulties, regularity models of causation have largely been abandoned in favor of counterfactual models. Rather than defining causality purely in reference to observable events, counterfactual models define causation in terms of a comparison of observable and unobservable events. Linguistically, counterfactual statements are most naturally expressed using subjunctive conditional statements such as "if India had not been poor, mortality rates would have been lower". Thus, the counterfactual approach to causality begins with the idea that some of the information required for inferring causal relationships is and will always be unobserved, and therefore some assumptions must be made. In stark contrast to the regularity approach of Hume, the idea of counterfactual causation is fundamentally separate from the tools used to infer it. As a result, philosophers like David Lewis (1973) could write about the meaning of causality with little discussion of how it might be inferred. It was statisticians, beginning with Jerzy Neyman (1923) and Ronald A. Fisher (1918; 1925), who began to clarify the conditions under which causal inferences were possible if causation was fundamentally a 'missing data problem'.

The gold standard approach to answering causal questions is to conduct a controlled experiment in which treatments/exposures are allocated at random, all subjects are perfectly compliant, and all the relevant data are collected and measured without error. Randomized experiments first appeared in psychology and in education. Later, they became popular in others fields thank to Neyman and Fisher. In the real world, however, such experiments rarely attain this ideal status (with the presence of non-compliance, non-response, missing data, measurement error...),

and for most important questions, an experiment would not even be ethically, practically, or economically feasible. In these situations, as Hume anticipated, moving from measuring an association to inferring a causal link is not trivial.

Fisher, who was also a smoker, testified before Congress that the correlation between smoking and lung cancer could not prove that the former caused the latter. His claim was that one cannot substantiate causal conclusions from associations alone, even at the population level, as implied by the slogan 'Correlation does not imply causation', with the now-standard appendix, 'But it sure is a hint'.

This has led some researchers to dismiss the search for causes as something that is outside the realm of science. Until very recently, in fact, the dominant methodology has been based almost exclusively on statistical analysis which, traditionally, has excluded causal vocabulary both from its mathematical language and from its educational program. The aim of standard analysis, typified by regression and other estimation techniques, is to infer parameters of a distribution from samples drawn from that population. With the help of such parameters, one can infer associations among variables, estimates the likelihood of past and future events, as well as update such likelihood in light of new evidence or new measurements. These tasks are managed well by statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer aspects of the data generating process. With the help of such aspects, one can deduce not only the likelihood of events under static conditions, but also the dynamics of events under changing conditions, for example, changes induced by treatments or external interventions. This capability include predicting the effects of interventions, (e.g., treatments or policy decisions) and spontaneous changes, (e.g., epidemics or natural disasters), identifying the cause of reported events, and assessing responsibility and attribution. This distinction implies that causal and statistical concepts do not mix. Statistic deals with static conditions and observed phenomena, while causal analysis deals with changing conditions. There is nothing in distribution function

that would tell us how that distribution would differ if external conditions were to change because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. The joint distribution of symptoms and diseases cannot tell us whether curing the former would or would not cure the latter. Even the theory of stochastic processes, which provides probabilistic characterization of certain dynamic phenomena, assumes a fixed density function over time-indexed variables. The additional information needed for making such predictions must be provided by causal assumptions, which identify relationships that remain invariant when external conditions change. The role of these assumptions is to recover the unobserved information on the dynamics of events if external conditions were to change from their observed status to a different hypothesized status.

Over the last thirty years a formal statistical language has been developed in which causal effects can be unambiguously defined, and the assumptions needed for their estimation clearly stated. A statistical framework for causal inference that has received increasing attention in recent years is the one based on potential outcomes. It is rooted in the statistical work on randomized experiments by Fisher (1918, 1925) and Neyman (1923), as extended to nonrandomized studies and to other modes of inference by Rubin (1974, 1976, 1977, 1978, 1990) and subsequently by others. This framework was called 'Rubin's Causal Model' (RCM) by Holland (1986). The RCM allows the direct handling of complications, such as noncompliance with assigned treatment (which bridges experiments and the econometric instrumental variables methods, Angrist et al. 1996). Thank to the clarity it brings in questions of causality, in the late 1980s and 1990s this framework has become increasingly popular in many fields including statistics (Holland, 1986; Rubin, 1974, 2006; Rosenbaum, 2002), medicine (Christakis & Iwashyna, 2003; Rubin, 1997), economics (Bjorklund & Moffitt, 1987; Heckman & Hotz, 1989; Manski, 1990; Manski et al., 1992; Angrist & Imbens, 1995; Abadie & Imbens, 2006), political science (Bowers & Hansen, 2005;

Imai, 2005), sociology (Winship & Morgan, 1999; Smith, 1997) and even law (Rubin, 2001).

The RCM has two essential parts and one optional part. The first parts defines causal effects using the concept of potential outcomes. Assume that there are just two levels of treatment, denoted by 0, the control, and 1,the active treatment.. The starting essential feature of the approach is to define a causal effect as the comparison of the potential outcomes on the same unit measured at the same time: Y (0), the value of the outcome variable Y if the unit is exposed to treatment 0, and Y (1), the value of Y if exposed to the active treatment 1. The key problem of causal inference is that, for any unit, only one of these two potential outcomes can be observed, namely the one corresponding to the treatment actually received, and the potential outcome under the other treatment is missing. Thus, causal inference becomes a problem of inference with missing data.

This definition of causal effects as the comparison of potential outcomes is frequently used in contemporary culture, for example, as revealed by movies. Most of us have probably seen the film 'Sliding Doors?, with Gwyneth Paltrow as Helen Quilley. The film reveals two parallel story lines: what happens when Helen makes it through the "sliding doors" onto the train and what happens when she misses the train. When she squeezes through the sliding doors and catches the train, she meets a charming man called James and gets home to find her boyfriend Gerry in bed with another woman. In the other reality, Helen misses the train, gets mugged, goes to hospital and eventually arrives home to find Gerry alone in the shower. The stream of events that would occur under the two scenarios are Helen?s potential outcomes. Looking forward into the future at the moment of the sliding doors, both could potentially occur. Whichever narrative will take place, will be the 'actual' world or the 'observed' outcome. The one that does not occur would be termed 'counterfactual' or the 'missing' outcome.

The second part of the RCM concerns the definition of a model for the assignment

mechanism, a stochastic rule for assigning treatments to units and thereby for revealing Y (0) or Y (1) for each unit. The assignment mechanism can depend on other measurements; if these other measurements are observed values, then the assignment mechanism is ignorable and the assignment can be treated as if it were random; if the given observed values involve missing values, possibly even missing Ys, then it is nonignorable. All forms of statistical inference for causal effects, whether Bayesian or frequentist, require the positing of an assignment mechanism. The realization that the primacy of the assignment mechanism holds true for observational data no less than for experimental, is due to Donald Rubin. This insight has been turned into a motto: 'no causation without manipulation'. The third optional part of the framework is the use of Bayesian posterior predictive inference for causal effects.

In order to define for each unit of one and only one potential outcome for each treatment level, a key assumption is the stable unit treatment value assumption (STUVA, Rubin (1978, 1980, 1990)), which requires that the potential outcomes for any given unit be invariant with respect to the treatment assignment as well as to the treatment status of other units. The second part of this assumption is what Cox referred to as the absence of 'interference between different units' (Cox, 1958), meaning that the outcome for a given unit is unaffected by the treatment status of other units. This phenomenon is common in social settings where people communicate, compete, or spread disease. Such interference between units may pose a threat to the inference of causal treatment effects. Interference implies that treatment effects are not comparisons of two potential responses that a unit may exhibit, one under treatment and the other under control, but instead are inherently more complex. Sometimes this interference is a nuisance, in which case we might design the study perhaps by isolating experimental units from one another, so interference does not occur. Although this is good advice in many settings, it is highly impractical or not logically possible in many common situations. Also, the effect of the treatment received by other units, called 'spillover effect', may be of intrinsic interest. Treat-

ments may be applied to people in an existing network, and we may wish to study how effects transmit to peers in the network. For instance, it might be of interest to know the extent to which an individual is protected as a result of receiving a vaccine and the extent to which the individual is protected due to others in the same cluster receiving the vaccine. One design that facilitates causal inference in the presence of interference is a two-stage randomized experiment in which specific clusters are randomized to having a certain proportion of the cluster treated and then, within each cluster, once the proportion is determined, individuals are randomized to receive the treatment or not. However, detecting interference between units represents a generally challenging problem in observational studies as well as in single-stage randomized experiments, when the two-stage design is infeasible due to practical reasons or when interference is not considered in the design phase, and even in two-stage randomized experiments when non-compliance arises.

Sometimes, interference is confused with another quite different issue that also may arise in contexts that produce interference: statistical dependence produced by pretreatment clustering. There is dependence between outcomes of different units when there are common unmeasured factors that may affect a cluster of individuals, but also if there is reciprocal influence between outcomes of different subjects at different times. For example, people in the same family may tend to exhibit similar responses to a viral infection because of shared genes; this is clustering. In this work we will deal with both issues, interference and clustering.

Although impact assessment, i.e. the assessment of the effect of an intervention, is certainly of primary importance in many substantive contexts, it is often of both scientific and practical interest to explain why and how an intervention works. Answering such questions will not only enhance the understanding of causal mechanisms behind the intervention, but may also enable policymakers to prescribe better policy alternatives, so that the key elements of programs can be supported, and the key problems in programs that fail to reach their goals can be repaired.

For instance, a job training may prevent recidivism not only through employment but also trough a change in motivation, in the level of personal skills or in commitment to conventional social bonds. An understanding of these processes can help designing the various components of the program.

Average causal effects do not themselves provide information about the reasons why the interventions have the effects they do. These reasons are left as an unopened black box. An explanation of causal mechanisms can help analysts overcome the "black box" problem. During the past decade, such intriguing analysis have received considerable attention mostly in social and health sciences. Nonetheless, mediation analysis can be an arduous task, given that it magnifies the almost inevitable selectivity and omitted variable biases that plague research. Recently a number of statisticians have taken up the challenge. Robins and Greenland (1992) formalized this type of analysis in causal terms, and a number of articles have appeared in more recent years (e.g. Pearl, 2001; Petersen et al., 2006; Imai et al., 2010a; Ten Have & Joffe, 2012). In causal mediation analysis the definition of causal mechanisms hinges on particular potential outcomes that are based on a joint hypothetical manipulation of both the intervention and the intermediate variable. Due to this joint hypothetical manipulation - that is generally never observable - the lack of information on these quantities requires additional identifying assumptions, such as sequential ignorability.

Exploration of causal mechanisms can be crucial in clustered encouragement designs (CED). Encouragement design studies arise frequently when the treatment cannot be enforced because of ethical or practical constrains and an encouragement intervention (information campaigns, incentives...) is conceived with the purpose of increasing the uptake of the treatment of interest (drugs, job trainings, vaccines, preventive measures...). By design, encouragements always entail the complication of non-compliance and whether an individual takes the treatment or not under alternative encouragement conditions depends on observed and unobserved charac-

teristics. Encouragements can also give rise to a variety of mechanisms, particularly when encouragement is assigned at cluster level. In fact, social interactions among units in the same cluster can result in mechanisms of interference, that is one subject's outcome is affected not only by treatment he himself received but also by the treatment received by the other subjects belonging to the same social group (e.g., neighbourhood or village). Because of the presence of this and other mechanisms, estimation of the effect of the treatment in CED typically cannot appeal to the common assumptions used in non-compliance settings, i.e., exclusion restrictions, which precisely rule out any mechanism where assignment affects the outcome not through the intermediate variable, i.e., the treatment. Knowledge of these alternative pathways is not only relevant for ensuring the unbiased estimation of treatment effects but also often of substantive importance. Disentangling the effect of encouragement through spillover effects from that through the enhancement of the treatment would give a better insight into the intervention and it could be compelling for planning the scaling-up phase of the program. Given the difficulty, there have been few, if any, attempts to probe mechanisms in clustered encouragement designs. Building on previous works on CEDs and non-compliance, we use the Principal Stratification (PS) framework to define stratum-specific causal effects, that is, effects for specific latent subpopulations, defined by the joint potential compliance statuses under both encouragement conditions. The Principal Stratification approach was first introduced to generalize the Instrumental Variable (IV) method used to estimate the effect of a treatment in non-compliance settings (Imbens & Angrist, 1994; Angrist et al., 1996; Imbens & Rubin, 1997). PS was then proposed by Frangakis & Rubin (2002) as a general framework for the evaluation of treatment effects while adjusting for post-treatment variables.

We show how the latter stratum-specific causal effects provide flexible homogeneity assumptions under which an extrapolation across principal strata allows to disentangle the effects. To face identification issues, estimation of causal estimands

can be performed with Bayesian inferential methods using hierarchical models to account for clustering.

Outline of the Thesis

In this work the above mentioned subjects are elaborated in the particular context of clustered encouragement designs. The thesis is divided into two chapters. The objective of the first chapter is to disentangle two causal mechanisms through which the clustered encouragement exerts its effect on the individual outcome of interest: the one through the individual uptake of the treatment that is promoted by the encouragement, and the one through other processes, including through interference by the treatment received by the unit's neighbors or through other behavioral changes that can result from the encouragement. To this purpose, we present a new framework that defines each causal mechanism within individual principal strata, that is, latent subpopulations characterized by the potential values of the individual treatment receipt under alternative situations defined by the clustered encouragement assignment. In order to accomplish identification of the stratum-specific mechanisms, we proposed novel homogeneity assumptions that allow a limited extrapolation of the missing information across a subset of principal strata. Bayesian estimation is obtained by means of an imputation approach, where each step of the algorithm follows from the hypothesized assumptions. We illustrate the proposed methodology analyzing a cluster randomized experiment implemented in Zambia and designed to evaluate the impact on malaria prevalence of an agricultural loan program intended to increase the bed net coverage.

The second chapter is devoted to the separation from all the other mechanisms of the effect of the clustered encouragement through the treatment received by all the units belonging to the same cluster, that is, the effect through a mechanism of interference. Therefore, the objective here is to disentangle three causal mechanisms: the one through the individual uptake of the treatment, the one through interference by the treatment received by the neighbors, and another separate one

through other behavioral changes that can follow from the encouragement both at individual and cluster level. Building on the framework presented in the previous chapter, we propose an extension that further develops the concept of principal stratification to take into account both individual and neighborhood compliance, that is the behavior in terms of potential treatment uptake under the alternative encouragement conditions of both the unit itself and the other subjects of the same cluster. With this auxiliary neighborhood principal stratification, identification of the three causal mechanisms is achieved by additional homogeneity assumptions, that are similar in flavor to those proposed in chapter one but will also take into account the neighborhood compliance. An extended bayesian imputation algorithm attains the estimation of such causal estimands. To assess the frequentist performance of our bayesian estimation procedure, we performed a simulation study that is based on a real randomized experiment, whose aim was to determine the extent to which mobile immunization camps can help boosting immunization coverage in rural India.

Chapter 1

Disentangling Causal Mechanisms using Individual Principal Strata

1.1 Introduction

The main purpose of clinical trials and impact evaluations, as well as social or epidemiological studies, is to provide evidence to guide the implementation of policies and programs or the development of control measures and prevention procedures for specific target populations. Evidence-based practice, as an approach for decision-making grounded in experiential evidence from the field and relevant contextual information, has gained considerable interest and influence over the last decades in the fields of economics, psychology, political, social and health sciences. With mediation analysis, defined as the analysis of the mechanisms through which an intervention or exposure has an effect on the outcome of interest, research has gone far beyond providing evidence of overall effects. This type of analysis is mostly still confined in scientific investigation studies to test competing theories such as behavioral, epidemiological or economic models but evidence on how the effect of the exposure or treatment is actually accomplished is rarely turned into practice. A deep understanding of underlying mechanisms could be used by psychologists, social workers, health service managers or policymakers to better design their intervention in order to achieve the expected development goals. Improvements could include tailoring and focusing on particular successful components of the interventions, differentially targeting the beneficiaries and/or seeking additional components or substitute inter-

ventions that might alternatively affect the intermediate variables, i.e., variables that appear in the causal process relating the intervention and the outcome. "Unpacking the black box" by disentangling the different mechanisms involved is certainly a fascinating field of research but doubtless an arduous task. Sometimes the more you try to unravel a tangle the more you end up tying new knots. It all depends on how tight the knots are, the knowledge you have of the structure and how deep you are willing to dig. If the structure of the tangle is not completely known, assumptions have to be made before starting to pull the needles.

Modern causal inference approaches to mediation analysis, grounded in the potential outcomes framework (Rubin, 1974, 1978), have garnered tremendous support among both researchers and practitioners, although they still require strong and untestable assumptions. Indeed, since the first attempts to exploration of causal mechanisms (Baron & Kenny, 1986; MacKinnon et al., 2002), researchers have provided a more formal framework, based on causal effects whose definitions depend on hypothetical interventions on the intermediate variables. These estimands are known in the literature as 'direct' (or net) and 'indirect' (or mediational) effects, which essentially refer to the effect of the exposure or intervention on the outcome respectively not through or through a change in the intermediate variable (Robins & Greenland, 1992; Pearl, 2001). Because of their definitions, they involve quantities, sometimes named *a priori counterfactuals*, that cannot be estimated from the observed data without strong and untestable assumptions. Most of the approaches to mediation analysis hinge on *sequential ignorability* assumption (Imai et al., 2010a,b; Hafeman & VanderWeele, 2011) that, in addition to the intervention being randomly assigned, requires unconfoundedness of both the assignment and the intermediate variable, given the baseline covariates. The key assumption of unconfoundedness of the intermediate variable is highly controversial, given that is often untenable in many clinical and behavioral studies. Several authors have tried to address the problem through different techniques such as instrumental variables

(Ten Have et al., 2007; Dunn & Bentall, 2007; Albert, 2008; Small, 2012), sensitivity analysis (Imai et al., 2010b; VanderWeele, 2010a), or *Principal Stratification* (PS) (Frangakis & Rubin, 2002). The former method uses baseline covariates interacted with random assignment as instrumental variables, trading sequential ignorability with alternative assumptions such as homogeneous effects across all individuals. In the latter approach causal effects of the intervention are defined within principal strata, latent sub-populations defined by potential values of the intermediate variable. These stratum-specific effects, named Principal Causal Effects (PCE), have the property to always have a causal interpretation (provided that principal strata are unaffected by the assignment).

Principal Stratification has been introduced in the context of mediation analysis primarily as a way to highlight the limitations of standard approaches, in terms of questionable assumptions and conceptual issues (Mealli & Rubin, 2003; Rubin, 2004; Mealli & Mattei, 2012). Its use to address these limitations has been then proposed by Hill and colleagues (2002) and subsequently applied with Bayesian estimation techniques by Gallop and colleagues (2009), Elliott and colleagues 2010, and Page (2012), and finally improved by Mattei & Mealli (2011) who developed an augmented design to ease identification and estimation. Further comparisons of identifying assumptions and estimation procedures between PS and other mediation methods can be found in the literature (e.g. see Jo (2008) for a comparison with structural equation models (SEM), or Lynch et al. (2008) and Ten Have & Joffe (2012) for a comparison with g-estimation approach). Vanderweele (2008) clarifies the relationship between these two different type of effects defined by PS approach and by the standard approach based on hypothetical interventions on the intermediate variable. He shows how the effect of the assignment (Principal Causal Effect, PCE) for those whose treatment uptake does not depend on the assignment, referred to as Principal Strata Direct Effects (PSDE), can be interpreted as only "direct" effects, whereas for the rest of the population PCE can be thought as a mixture of

both "direct" and "indirect" effects. The use of PS is still an ongoing debate because of these conceptual issues and because it has been argued that it provides an estimate of the "direct" effects solely in a specific subgroup and little information on "indirect" effects (Pearl, 2011; Mealli & Mattei, 2012; VanderWeele, 2012).

Oftentimes, the so called Principal Strata Direct Effects (PSDE), that is the effect of the assignment for those whose intermediate variable does not depend on the assignment, are "naively" interpreted as "direct" effects for the whole population, implicitly making some kind of homogeneity assumptions across all principal strata. This approach leads to the statement saying that if there is no effect of the intervention for the strata where the intermediate variable is constant then all the effect of the intervention is through a change in this variable (see e.g. Elliott et al. (2010)) Homogeneity assumptions needed for the identification of causal effect defined with a priori counterfactuals on the basis of PS have been briefly discussed by Jo (2008), in terms of constancy of coefficients of SEM, by Page (2012) and by Flores & Flores-Lagunes (2009a), who also provided a set of weak monotonicity assumptions to derive nonparametric bound for the effects.

Interesting questions concerning mechanisms can be raised in a typical non-compliance setting, where the treatment itself mediates the effect of the assignment. When compliance to treatment assignment is not perfect, sometimes assignment is itself source of alternative behaviors that would affect the outcome even without involving a change in the treatment received. A particular design that can be viewed as a randomized controlled trial with non-compliance is the encouragement design. Encouragements are used when a treatment cannot be enforced for ethical or practical reasons. Treatments can be therapeutic drugs or programs, preventive measures (vaccines, condoms, bed nets...), protective or risky behaviors (drug or alcohol abuse,...). It is evident how many of the mentioned factors are not manipulable because inherently subject to self-selection or because of ethical or practical constrains. When a treatment cannot be randomly assigned, as the best practice

for clinical trials and impact evaluation would require. In these situations, encouragements, such as different conditions of offering or promoting the treatment, can be used as ex-ante instruments to induce an exogenous variation of the uptake of the treatment (Bradlow, 1998). Alternatively it can also be the case that the target treatment is an exposure or intervention that has already been evaluated in previous experimental or observational studies, providing evidence of its beneficial or detrimental effect on the outcome of interest, but its use or disuse cannot be imposed in the population. In these other circumstances encouragement interventions can be conceived to foster a behavioral change of the target population, that is to increase the probability of adoption of a beneficial treatment or decrease the likelihood of a negative behavior. In this case, encouragements can take the form of incentives, additional information, different strategies for treatment supply or public policies in general.

Hirano et al. (2000) were the first to apply the Principal Stratification approach to encouragement designs to estimate intention-to-treat effects within principal strata, i.e. PCE, with and without exclusion restriction assumptions (Imbens & Angrist, 1994; Angrist et al., 1996; Imbens & Rubin, 1997). Oftentimes, in fact, the encouragement is itself source of alternative behaviors that would affect the outcome even without involving a change in the treatment received. When the encouragement is solely an "instrument" to induce the target treatment and the aim of the study is to evaluate the treatment effect, even if the presence of alternative pathways would prevent the estimation of the effect of the treatment, we will show how disentangling the total effect of the encouragement would give insight into the dynamic of the process. On the other hand, even when the major interest relies on the effect of the encouragement on the tratment uptake and in turns on the outcome, investigating the underlying mechanisms through which the encouragement program achieves its goal is important for both descriptive and prescriptive reasons. Indeed such analysis would primarily enable assessing whether the intervention is working

the way we expect it to, that is by changing the behavior in terms of treatment uptake, with the treatment having an effect on the outcome, and secondly it would allow probing for other potential mechanisms of the cluster intervention. Based on this analysis, we believe that future interventions might more efficiently impact the outcome of interest, through tweaking the encouragement programs or policies and tailoring them to specific sub-populations with targeted components.

Here we consider *Cluster Randomized Encouragement Designs* (CED), where encouragement is randomized at the level of a cluster of subjects (e.g villages or communities) because of the specific structure of a community-based intervention (e.g. information campaigns, immunization camps, prevention measures...) or because of particular constrains, but compliance is at the individual level. CEDs with individual non-compliance can be found relatively frequently in many field experiments (Sommer & Zeger, 1991; McDonald et al., 1992; Hirano et al., 2000; Morris et al., 2004; among others). Frangakis, Rubin & Zhou (2002) extended previous work with PS to account for clustering using Bayesian hierarchical models for inference. To the best of our knowledge no previous work has attempted to apply concepts of mediation analysis to general non-compliance settings, with the treatment being the intermediate variable, and certainly not to the very common clustered encouragement designs. CEDs are intriguing because they can give rise to many different mechanisms that it is worthwhile to investigate. In fact, not only their relationship with the outcome depends on a change in the treatment uptake but also most of the times encouragements, incorporating sensitization towards the problem related to the outcome of interest, lead to an overall behavioral change and other actions that can substantially affect the outcome. Furthermore, since the encouragement is randomized at the cluster level, social interactions occurring among people living or working in the same environment give rise to mechanisms of what in the literature is referred to as *interference* or *spillover effects* (Sobel, 2006; Hong & Raudenbush, 2006; Hudgens & Halloran, 2008; Tchetgen Tchetgen & VanderWeele, 2012). Specif-

ically, neighbors' behaviors adopted as a result of the encouragement assignment, concerning the treatment uptake as well as other preventive or risky measures, affect not only their own outcomes but also those of the entire community.

By way of example, let us mention Conditional Cash Transfers (CCT), programs that have been extensively adopted in the last decade in the field of education, especially in Latin America (e.g. Mexico's Progresa, Schultz, 2004; Nicaragua, Maluccio, 2010; Honduras, Galiani & McEwan, 2013: Malawi, Baird et al., 2003) with the purpose of boosting schooling. CCT provide cash transfers to poor families, but their receipt is conditional on children attending school. Oftentimes these programs are assigned to all the poor household belonging to randomly selected municipalities or villages. This setting is a good example of compliance to encouragement. Not all the children whose families are offered the transfers would go regularly to school and even without additional money some children in control clusters would go to school anyway, with compliance depending on many socio-economic factors. The offering of cash transfers not only achieves its purpose of keeping children at school but it can also make families more aware of the education problem and make them support their children in their studies, it can motivate more effort or it can result in children getting higher grades at school because of a better psychosocial environment in the house. Furthermore, in the field of education usually social interactions cannot be neglected. In fact children attending school would share their textbook or information learnt with their peers, and also peer influence in motivation or emotional competency may occur. Thus, the presence of CCT program in a cluster may affect all types of children, attending or not attending school with or without conditional cash transfers.

VanderWeele et al. (2013) have already attempted to disentangle spillover effects in cluster randomized trials. Nevertheless, we argue that their identifying assumptions, which essentially extend sequential ignorability assumptions to accommodate cluster-level assignment and spillovers, are too stringent and rarely apply to the

extreme case of CEDs, where compliance behavior depend on the overall individual decision-making system.

The work presented in this chapter makes three contributions to the literature. First, we conceptualize the mediating role of the treatment variable in clustered encouragement designs using definitions of effects based both on hypothetical interventions on the treatment uptake and on principal strata. Second, we provide two alternative sets of homogeneity assumptions that enable to extrapolate information across principal strata and use the estimated PCE to recover the effects involving a priori counterfactuals. We discuss the flexibility of these assumptions and make clear what specific causal effects can be identified by each of them. Throughout the article, the reference to the application will be useful to outline possible ways to assess their plausibility. Third, building on previous work (specifically Frangakis, Rubin & Zhou, 2002), we incorporate an imputation-based procedure for the estimation of these intervention-based causal effects under the required assumptions.

This chapter is organized as follows. Section 1.2 describes the motivating study that we will use to illustrate the methodology. Section 1.3 provides notation and setup. In Section 1.4 we introduce the Principal Stratification approach and define a new class of causal estimands that adapt to the context of CED the notion of mechanisms based on a priori counterfactuals. Section 1.5 presents our innovative structural assumptions deriving the identification results. Section 1.6 concerns the models we will consider in the bayesian inference laid out in Section 1.7 together with the new imputation-based procedure. In Section 1.8 the methods are applied to KAHS study. We conclude in Section 1.9 with some discussion.

1.2 Motivational Study: Katete Agriculture and Health Study (KAHS)

The proposed methodology is motivated by the Katete Agriculture and Health Study (KAHS) implemented in the Katete District, a rural area with highly endemic malaria in Zambia's Eastern Province (Fink & Masiye, 2012). From a list of 256 clusters, corresponding to small rural settlements of about 250 households each, the study was restricted to 49 non-contiguous clusters, with a minimum distance of 3 km between each other. The 49 clusters were randomly assigned to one of three arms: 15 were assigned to the control group, 15 to a free net distribution and the other 19 to a subsidized bed net loan program. The purpose of the two 'encouragement' interventions was to increase bed nets coverage and ultimately reduce malaria prevalence. Here, we use a subset of the original data from the first and the third arms.

The target population of the study comprised rural farmers, known to be a population group at high risk of malaria. In each cluster, 11 farmer households were randomly selected from a complete listing of all farmers working with Dunavant Cotton, the partner organization of the program. All the households enrolled in the study, in all the three arms, were surveyed twice, once prior to the rainy season and a second time five months later. All the 11 households selected in the clusters assigned to the third arm, after the baseline interview, were allowed to obtain bed nets at a subsidized price, with repayments due at the end of the harvesting season with a crop sale deduction system. However, not all households offered the subsidies took advantage of them ordering new bed nets, whereas in the control clusters families could also buy new bed nets from local markets. Fink & Masiye (2012) evaluated the average effect of offering the agricultural loan program on the household prevalence of malaria with an intent-to-treat analysis.

There has been an extensive effort over the past decade to show the effectiveness of bed nets uptake in reducing malaria morbidity (Alonso et al., 1993; D'Alessandro,

1995; Nevill et al., 1996). Relying on these results, in the last years studies in this field usually focus on the evaluation of strategies to improve coverage. However few studies attempt to understand how these strategies work and whether their merit goes beyond the increase in bed net uptake. One of the underlying mechanisms that can occur in large-scale interventions is interference. Given the minimum distance of 3 km between clusters, any concerns of interference between cluster can be reasonably ruled out. On the contrary interference within clusters is likely to take place. Bed nets usage yields protection from malaria infection not only for subjects sleeping under them but also for individuals living in the same area. In the literature this effect is referred to as *mass community effect*. First and foremost, bed nets reduce the reservoir of infection by preventing the physically protected individuals from being infected. This effect is analogous to the *contagion effect* in vaccine trials (VanderWeele et al., 2012). In addition bed nets commonly used in the last two decades are insecticide-treated nets (ITNs). As a matter of fact the bed nets distributed in the program as well as those sold in the local markets in Zambia are treated with insecticides. ITNs yield an additional mass effect by affecting the vector of transmissions in three ways. First, insecticides kill adult mosquitos infected with malaria parasites reducing the probability of a person in the community being bitten by an infected mosquito. Second, mass coverage shortens the lifespan of the mosquitos and lowers the possibility for maturation of the parasites resulting in a reduction of the proportion of mosquitos that become infective. Third insecticides repel mosquitos. It has been argued that the repellant effect of the insecticides can be either harmful or beneficial for those who do not sleep under the nets. In fact, mosquitoes could be diverted to neighboring houses lacking nets. However this fear, plausible at low coverage, has been largely allayed especially if the coverage is high. On the contrary, a massive presence of bed nets might divert certain species of mosquitoes from human to animal biting, thereby reducing human-to-human transmission. These three components are analogues to the *infectiousness effect* (VanderWeele et al., 2012).

These components of the mass community effect of bed nets have already been assessed by several researchers in trials where free distribution of bed nets was randomized at cluster level and a comparison of malaria outcomes was carried out between households belonging to villages assigned to the intervention arm but who did not receive any net and households belonging to the control arm. Information on the distance between treated and untreated households was also used for the scope (Binka et al., 1998; Howard et al., 2000; Hawley et al., 2003). Nevertheless, none of the encouragement studies have tried to investigate the extent to which interference of the actual bed nets uptake or behavioral changes in the neighborhood plays a role for those who are assigned to receive new bed nets. The purpose of our analysis of the KAHS study is to investigate the different mechanisms through which the offer of agricultural loans had an effect, by analyzing the heterogeneous effect across different compliance behaviors about new bed nets purchase.

1.3 Notations and Definitions

In this section we will give formal definitions of the aforementioned effects in the potential outcomes framework (Rubin, 1974, 2005). The setting consist of $j = 1,\dots,J$ clusters and $i = 1,\dots,N_j$ units in each cluster with a total of N units uniquely denoted by the pair of indices ij. Let A_j denote a binary cluster encouragement assignment, so that $A_j = 1$ if cluster j is assigned to the encouragement program and $A_j = 0$ otherwise. Let $M_{ij} \in \{0,1\}$ and $Y_{ij} \in \mathcal{Y}$ denote the treatment received and outcome variables for unit i in cluster j. Let also introduce a vector of covariates, $\mathbf{C}_{ij} = \left(\mathbf{X}_{ij}, \mathbf{V}_j, h_i(\mathbf{X}_{-ij})\right) \in \mathcal{C}$, where $\mathbf{X}_{ij}$ is a vector of covariates of unit i in cluster j, $\mathbf{V}_j$ is a vector of cluster-specific characteristics and $h_i(\mathbf{X}_{-ij})$ is a function of the vector of covariates of all the units living next to unit i. Finally let $\mathbf{A}$, $\mathbf{M}$ and $\mathbf{Y}$ be the $(J\times 1)$-dimensional vector of encouragement assignments and the $(N\times 1)$-dimensional vectors of treatment received and outcomes, respectively.

In the KAHS study farmer households are the units of analysis and the clusters

of settlements are the units of assignment to either the agricultural loan program (A_j=1) or control (A_j=0). The treatment concerns the purchase of new bed nets between the baseline and the follow-up survey. To simplify the methodology, the analysis is based on a binary treatment variable, being $M_{ij} = 1$ if household i in cluster j has bought at least one more bed net and $M_{ij} = 0$ if no purchase has been carried out. In terms of the outcome, let Y_{ij} be the proportion of reported cases of malaria during the month prior to the follow-up interview in each household i belonging to cluster j. Note that throughout this thesis we will use the term "individual" to refer to the lowest level of the analysis, which in this case are households.

We now introduce notation for the primitive potential outcomes. Let $M_{ij}(\mathbf{A})$ denote the potential purchase of at least one bed net a household i would have decided to carry out under assignment vector $\mathbf{A}$. Similarly let $Y_{ij}(\mathbf{A},\mathbf{M})$ denote the potential outcome that household i in cluster j would have experienced if $\mathbf{A}$ and $\mathbf{M}$ were the vectors of assignments and treatments received in the whole population.

Assumption 1. *Cluster-level SUTVA for the encouragement assignment*
Cluster-level Stable Unit Treatment Value Assumption (SUTVA) for the encouragement assignment consists of two parts:

(i) An individual's potential outcomes and potential values of the intermediate variable do not vary with encouragements assigned to clusters other than the individual own cluster, i.e. $M_{ij}(\mathbf{A}) \equiv M_{ij}(A_j)$ and $Y_{ij}(\mathbf{A},\mathbf{M}) \equiv Y_{ij}(A_j,\mathbf{M}_j)$, where $\mathbf{M}_j$ is the vector of dimensions $N_J \times 1$ of treatment received by individuals of cluster j.

(ii) For each cluster there are no different versions of each encouragement level. Formally:

$$\text{if } A_j = A_j' \text{ then } M_{ij}(A_j) = M_{ij}(A_j')$$

$$\text{and if } A_j = A_j' \text{ and } \mathbf{M}_j = \mathbf{M}_j' \text{ then } Y_{ij}(A_j,\mathbf{M}_j) = Y_{ij}(A_j',\mathbf{M}_j')$$

Cluster-level SUTVA is an extension of the individual-level SUTVA introduced by

Rubin (1978, 1980, 1990) to settings with cluster-level assignments and individual-level intermediate variable. For further discussion on cluster-level SUTVA version see VanderWeele (2008). Yet it is worth noting that part (i) of the assumption requires that the outcome Y_{ij} of individual i in cluster j does not vary with the encouragement conditions or treatments received in other clusters. However the previous assumption does not rule out the possibility of spillover effects of the intermediate variable within clusters, that is Y_{ij} can be affected by the treatment received by other units of the same cluster j. Under cluster-level SUTVA we can use the notation $M_{ij}(A_j)$ and $Y_{ij}(A_j, \mathbf{M}_j)$.

Note that the only observable potential outcome is the one where, if A_j were set to a, the treatment received by all the units in cluster j were left to the value it would take under encouragement condition a, that is $Y_{ij}\big(a, \mathbf{M}_j(a)\big)$. Throughout we will use the notation $Y_{ij}(a)$ for potential outcomes of this type.

Based on these potential outcomes, the overall average effect of the cluster encouragement intervention on the individual outcome, referred to as *Intent-to-Treat Effect* (ITT), within each level $\mathbf{c}$ of baseline covariates, is defined as the following contrast:

$$ITT(\mathbf{c}) := \mathrm{E}\big[Y_{ij}(1) \mid \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid \mathbf{C}_{ij} = \mathbf{c}\big] \tag{1.3.1}$$

In the sequel, in order to be able to shed light on the heterogeneity of the effects, we will define all causal estimands as average effects within levels of the baseline covariates $\mathbf{C}_{ij}$.

1.4 Principal Stratification Approach

Principal stratification has been first introduced by Frangakis and Rubin (2002), in order to address post-treatment complications in an experimental setting. Its use in mediation analysis has been proposed as a way to relax the sequential ignorability assumption but still being able to yield valid causal inference of what VanderWeele (2008) called the principal strata direct effects (Gallop et al., 2009; Elliott et al.,

2010; Page, 2012; Mattei & Mealli, 2011).

The units under study can be stratified in subpopulations, the so-called *Principal Strata*, defined according to the potential values of the actual treatment received:

$$S^{m_0 m_1} := \{i : M_{ij}(0) = m_0, M_{ij}(1) = m_1\} \tag{1.4.1}$$

Since only one of the two potential values is observed, these four subpopulations are latent, in the sense that in general it is not possible to identify the specific subpopulation a unit i belongs to. Let S_{ij} be the indicator of the latent group to which subject i belongs. When both A_j and M_{ij} are binary there are 4 strata $S_{ij} \in \{S^{00}, S^{11}, S^{01}, S^{10}\}$, often referred in the literature on compliance as *never-takers*, *always-takers*, *compliers*, and *defiers*. Strata membership can also be referred to as compliance status.

In the bed nets application household can be divided in principal strata based on the behavior in terms of bed nets uptake under both encouragement conditions. Never-takers are the families who would not buy a new bed net neither if assigned nor if not assigned to receive subsidies, always takers are those who would buy new bed nets anyway, compliers those families who would buy new bed nets only if they were offered subsidies and defiers would be those who would not buy new bed nets with subsidies but would carry out the purchase at full price. We argue that this last category is not plausible in this setting and thus we will make the following monotonicity assumption.

Assumption 2. *Monotonicity of Compliance*

Monotonicity of encouragement assignment on treatment receipt requires

$$M_{ij}(0) \leq M_{ij}(1) \quad \forall i, j$$

This assumption rules out the presence of defiers. Indeed it conveys that there is no unit who would take the treatment if not encouraged to do so but would

not if encouraged. This restricted pattern of compliance behavior enables the identification of the conditional distribution of compliance status. In fact, if we let $\pi_{m_0 m_1}(\mathbf{c}) := P(S_{ij} = S^{m_0 m_1} \mid \mathbf{C}_{ij} = \mathbf{c})$ denote the probability of belonging to stratum $S^{m_0 m_1}$ conditional on baseline covariates, the monotonicity assumption implies the following result $\forall \mathbf{c} \in \mathscr{C}$:

$$\begin{aligned} \pi_{10}(\mathbf{c}) &= 0 \\ \pi_{11}(\mathbf{c}) &= P(M_{ij}(0) = 1 \mid \mathbf{C}_{ij} = \mathbf{c}) \\ \pi_{00}(\mathbf{c}) &= P(M_{ij}(1) = 0 \mid \mathbf{C}_{ij} = \mathbf{c}) \\ \pi_{01}(\mathbf{c}) &= 1 - \pi_{11}(\mathbf{c}) - \pi_{00}(\mathbf{c}) \end{aligned} \tag{1.4.2}$$

In what follows we will maintain this assumption. As mentioned previously in the KAHS study this assumption is plausible because there should not plausibly be any reason to buy a bed net at a full price but not with subsidies.

1.4.1 Principal Causal Effects

The overall effect of the cluster encouragement within each principal stratum and within levels of baseline covariates is named *principal causal effect* (PCE) and is defined as:

$$PCE(m_0, m_1, \mathbf{c}) := \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{1.4.3}$$

ITT is then a weighted average of PCEs, with weights given by the probability of belonging to each principal stratum:

$$ITT(\mathbf{c}) = \sum_{m_0 m_1} PCE(m_0, m_1, \mathbf{c}) \cdot \pi_{m_0 m_1}(\mathbf{c}) \tag{1.4.4}$$

In principal strata where the treatment receipt is unaffected by the encouragement, i.e. never-takers and always-takers, principal causal effect, $PCE(m, m, \mathbf{c})$ with

$m \in \{0,1\}$, are called *dissociative causal effect* ($DCE(m,\mathbf{c})$).

DCEs include all the mechanisms that do not involve a change in the treatment received. In particular they are a combination of two different types of effects: *pure encouragement effects*, that is effects of the cluster encouragement through modification in the environment or behavioral changes, other than the one regarding the treatment receipt, of both the unit itself or its neighbors (Frangakis, Rubin & Zhou, 2002), and effects due to mechanisms of *interference* due to a change in the treatment uptake of other inhabitants of the same cluster. Several behavioral changes often occur when the encouragement to take a beneficial treatment (or not take a risk behavior), is provided by information campaigns that will either increase the awareness of the problem or just act as reminders. In this case the additional information received would also encourage to make use of other measures to limit risk of infection. Behavioral changes can lead to an effect on the outcome either by themselves or in the way they vary the effect of the treatment on the outcome. When clusters are the level of randomization, these behavioral changes can also be at cluster level, such as structural interventions in the community. Interventions designed to boost the use of bed nets often comprise different components that are responsible of different mechanisms leading to malaria reduction. First, encouragements, such as subsidies, could influence the usage as well as the quantity of new bed nets; second, an awareness-raising component could lead to a better usage of old bed nets as well as the uptake of other preventive measures such as repellents or mosquito screens for windows and doors; third, another component could be a village cleaning or disinfestation.

The difference between the two dissociative effects for never-takers and always-takers can be substantial. On the one hand this can be due to the possible interaction between encouragement and treatment, that is, a change on the effect of A_j on Y_{ij} depending on the treatment uptake M_{ij}, on the other hand the different inherent characteristics of the two strata can influence the way the encouragement has an

effect on their outcome.

Estimation of the latter effect within levels of covariates $\mathbf{C}_{ij}$ would allow to identify the individual as well as cluster characteristics of the units that do not get any benefit from the cluster intervention if they don't take the treatment, neither through interference nor through other mechanisms. In the phase of scaling up the intervention to other communities, alternative targeted strategies can be applied to people with these characteristics and with a higher probability of being never-takers, e.g., free distribution of bed nets or higher discounts. Moreover, estimation of the effect for always-takers will provide us with a better understanding of the relevance of the encouragement and also whether the encouragement itself has a beneficial effect even for this sub-population.

As far as compliers are concerned, $PCE(0,1,\mathbf{c})$ is a combination of all the aforementioned mechanisms as well as the effect of the encouragement involving a change in the individual treatment uptake.

1.4.2 Individual Treatment Mediated Effect and Net Encouragement Effect

In order to disentangle for the whole population the two different types of causal mechanisms, through or not through a change in the individual treatment uptake, it is necessary to introduce quantities based on hypothetical interventions on the intermediate variable. Let us decompose $\mathbf{M}_j$ into $\mathbf{M}_j = [M_{ij}, \mathbf{M}_{-ij}]$, where $\mathbf{M}_{-ij}$ denotes the vector of treatment taken by all the individuals in cluster j, except for unit i, and let $\mathbf{M}_{-ij}(a)$ be its potential value under $A_j = a$. We can then rewrite the potential outcomes $Y_{ij}(A_j, \mathbf{M}_j)$, already defined in section 1.3, as $Y_{ij}(A_j, M_{ij}, \mathbf{M}_{-ij})$. Let us now consider a particular intervention on the intermediate variables that would set $\mathbf{M}_j = [M_{ij}, \mathbf{M}_{-ij}] = [m, \mathbf{M}_{-ij}(a)]$. Among the 2^{N_j+1} potential outcomes that can be conceived for each unit, based on a joint intervention on the encouragement and on the treatment receipt, we will focus solely on 4 of them, precisely the ones of the form $Y_{ij}(a, m, \mathbf{M}_{-ij}(a))$, denoting the outcome that unit i in cluster j would

have experienced if cluster j were assigned to the encouragement status $A_j = a$, the treatment received by unit ij were set to $M_{ij} = m$ and all the other individuals in the cluster could take the treatment they would have taken under the encouragement status that has been set to a. Since the third term in the potential outcome is a function of the encouragement condition, we will use the simplified notation: $Y_{ij}(a, m) \equiv Y_{ij}(a, m, \mathbf{M}_{-ij}(a))$. A peculiar case occurs when M_{ij} is set to the value it would take under encouragement $\tilde{a}$, i.e. $Y_{ij}\big(a, M_{ij}(\tilde{a})\big)$. As mentioned, potential outcomes of this form require that we conceive, together with the clustered encouragement intervention, an additional intervention that is able to set the treatment received by each subject to a specific value, without having any effect on the outcome. For instance, the joint intervention underlying the potential outcome $Y_{ij}(1, 0)$ is conceivable if there were a rationing, that is the number of bed nets available in the program were less than the number that households belonging to the villages were the program was implemented could potentially request. We then can think of an intervention that offers subsidized bed nets to household ij, but at the same time creates the condition for which that household finds nets out of stock, assuming no secondary consequences.

Potential outcomes of this type, whenever they can be deemed well-defined, allow the definition of causal estimands that decompose the overall encouragement effect into causal mechanisms, through or not through a change in the individual treatment uptake: *Individual Treatment Mediated Effect* (iTME) and *Net Encouragement Effect* (NEE). Note that in this article no attempt will be made to disentangle spillover effects from pure encouragement effects. This being said, we can give formal definition of the two main casual mechanisms of interest, within principal strata.

We define *Net Encouragement Effect* (NEE) within principal stratum $S^{m_0 m_1}$ as the following contrast:

$$NEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) := \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(0, M_{ij}(\tilde{a})\big) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{1.4.5}$$

In words, it is the difference between potential outcomes under the two encouragement conditions intervening to keep the individual treatment received by unit ij, M_{ij}, fixed at the value it would take under $A_j = \tilde{a}$, averaged over all units belonging to the principal stratum $S^{m_0 m_1}$ and with values of covariates $\mathbf{C}_{ij} = \mathbf{c}$. This quantity represents the effect of the encouragement on the outcome net of the effect of the treatment uptake. By definition, NEEs are a combination of spillover effects by intermediate variables of other subjects belonging to the same cluster and other mechanisms that do not involve a change in the individual treatment uptake. In the KAHS study, $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ indicates the average, over all units with $\mathbf{C}_{ij} = \mathbf{c}$ and belonging to principal stratum $S^{m_0 m_1}$, of the effect of offering subsidies to the 11 farmer households enrolled in the study and belonging to the same cluster on the risk of malaria for one of these units, not through the change in the number of bed nets owned by the household itself and, specifically, if we intervened to keep the binary indicator of bed nets purchase of this household to what it would have been under the clustered encouragement status $A_j = \tilde{a}$.

Likewise, the *Individual Treatment Mediated Effect* (iTME), for each encouragement condition $A_j = a$, is given by the following expression:

$$iTME^{a}(m_0, m_1, \mathbf{c}) := \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(0)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \tag{1.4.6}$$

In words, it is the average difference of the potential outcomes, within each principal stratum and within each level of the covariates, resulting from an intervention that varies the actual treatment for each unit i in cluster j, M_{ij}, from the one that this unit would have received having assigned cluster j to the active encouragement condition, $A_j = 1$, to the one that it would have received under the control encouragement condition, $A_j = 0$, keeping the encouragement status fixed at a. Precisely this quantity captures to what extent the encouragement achieves its aim through its main characteristics, i.e. an increase or reduction of the treatment uptake in

the population. The definition of the quantities NEE and iTME is not new in the literature of mediation analysis. Indeed, they correspond to the natural direct and indirect effects (Robins & Greenland, 1992; Pearl, 2001) within principal strata (VanderWeele, 2008; Mealli & Mattei, 2012). VanderWeele (2010b) also provided expressions for these effects when a treatment is administered at cluster level and the intermediate variable is measured at individual level. The change in the terminology is due, in our view, to a better fit to the setting of clustered encouragement designs, where the terms direct and indirect would be confusing.

Let us focus now on the strata $S^{mm} = \{i : M_{ij}(0) = M_{ij}(1) = m\}$, with $m = \{0,1\}$, where the individual treatment received, M_{ij}, does not depend on the encouragement intervention A_j, namely never-taker (m=0) and always-takers (m=1). Within these two strata the individual treatment mediated effect is canceled out and the dissociative causal effect equals both net encouragement effects:

$$DCE(m,\mathbf{c}) \equiv NEE^0(m,m,\mathbf{c}) = NEE^1(m,m,\mathbf{c}) \tag{1.4.7}$$

Proof. The proof is carried out bearing in mind that in the strata of the type S^{mm} the two potential values of the intermediate variable, $M_{ij}(1)$ and $M_{ij}(1)$, coincide.

$$\begin{aligned}
DCE(m,\mathbf{c}) &= \mathrm{E}\left[Y_{ij}(1) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}(0) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(1)\right) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0)\right) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0)\right) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0)\right) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= NEE^0(m,m,\mathbf{c})
\end{aligned}$$

With similar manipulations we yield the second result:

$$\begin{aligned}
DCE(m,\mathbf{c}) &= \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(1)\big) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(0, M_{ij}(0)\big) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(1)\big) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(0, M_{ij}(1)\big) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= NEE^{1}(m, m, \mathbf{c})
\end{aligned}$$

□

In contrast, the overall effect of the clustered encouragement for compliers decomposes into the net encouragement effect and the individual treatment mediated effect:

$$PCE(0, 1, \mathbf{c}) = NEE^{1-a}(0, 1, \mathbf{c}) + iTME^{(a)}(0, 1, \mathbf{c}) \tag{1.4.8}$$

Proof.

$$\begin{aligned}
PCE(0,1,\mathbf{c}) &= \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(1)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(0, M_{ij}(0)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(1)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(a, M_{ij}(1-a)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad + \mathrm{E}\big[Y_{ij}\big(a, M_{ij}(1-a)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(0, M_{ij}(0)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= NEE^{1-a}(0,1,\mathbf{c}) + iTME^{a}(0,1,\mathbf{c})
\end{aligned}$$

□

In our example, $iTME^{a}(0,1,\mathbf{c})$ represents the average effect of the agricultural loan program on the proportion of malaria cases experienced by each complier household with $\mathbf{C}_{ij} = \mathbf{c}$ through an increase in the number of bed nets owned by the household itself, under the clustered encouragement status $A_j = a$.

The choice of one or the other decomposition, indexed by a, depends on the particular interest, the application, a descriptive or perspective approach and future

interventions under assessment. As we can see compliers are the only units to actually exhibit a non zero iTME besides a possible difference between the two NEEs. The reason of the difference between the two effects $NEE^0(0,1,\mathbf{c})$ and $NEE^1(0,1,\mathbf{c})$, as well as between $iTME^0(0,1,\mathbf{c})$ and $iTME^1(0,1,\mathbf{c})$, can be found in the interaction between the encouragement and the individual treatment uptake, so that the effect of the clustered encouragement might depend on whether the subject takes the treatment and conversely the effect of the treatment on the outcome varies with the presence of the encouragement.

In any case, a conceptual point has to be made. In this application the effects $NEE^1(0,1,\mathbf{c})$ $iTME^0(0,1,\mathbf{c})$ are problematic because they involve the potential outcome $Y_{ij}(0, M_{ij}(1))$, which for compliers is equal to $Y_{ij}(0,1)$. This quantity is not well-defined because it would require an intervention that sets M_{ij} to 1, namely that makes a complier household ij buy at least one new bed net, while each household in cluster j, including ij, is not assigned to the loan program. Since the purchase of bed nets is a treatment that cannot be enforced such intervention is hard to conceive and it would rather be another kind of encouragement that would affect the actual number of bed nets bought as well as lead to other mechanisms. On the contrary, $NEE^0(0,1,\mathbf{c})$ $iTME^1(0,1,\mathbf{c})$ involve the potential outcome $Y_{ij}(1, M_{ij}(0))$, which is equal to $Y_{ij}(1,0)$ for compliers. This quantity hinges on an intervention that sets M_{ij} to 0, namely that precludes the purchase of any new bed net for a complier household ij, while each household in cluster j, including ij, is assigned to the loan program. In a way this might be easier to conceptualize if we think on the rationing intervention described earlier.

In light of these considerations, the scope of our analysis will be to disentangle net encouragement effect and individual treatment mediated effect for compliers in the form of $NEE^0(0,1,\mathbf{c})$ and $iTME^1(0,1,\mathbf{c})$, and estimate dissociative causal effects for always-takers and never-takers.

We can now derive population effects. The population net encouragement effect,

averaged over subgroups of the population with the same level of covariates, is given by the weighted sum of the net encouragement effect of all the strata:

$$\begin{aligned} NEE^0(\mathbf{c}) &= \sum_{(m_0,m_1)} NEE^0(m_0,m_1,\mathbf{c})\pi_{m_0m_1}(\mathbf{c}) \\ &= \sum_m DCE(m,\mathbf{c})\pi_{mm}(\mathbf{c}) + NEE^0(0,1,\mathbf{c})\pi_{m_0m_1}(\mathbf{c}) \end{aligned} \tag{1.4.9}$$

Conversely, the population intermediate treatment mediated effect, averaged over subgroups of the population with the same level of covariates, results from intermediate treatment mediated effect for compliers, scaled by the conditional probability of belonging to this principal stratum:

$$iTME^1(\mathbf{c}) = iTME^1(0,1,\mathbf{c})\pi_{01}(\mathbf{c}) \tag{1.4.10}$$

As we will see more in details in section 1.5.2, by virtue of the particular behavior of compliers, having $M_{ij}(0) = 0$ and $M_{ij}(1) = 1$, we can interpret their individual treatment mediated effect for compliers as the average causal effect of the receipt of treatment within this subpopulation. This makes it clear that the individual treatment mediated effect, being a product of two quantities, represents both the impact of the encouragement on the treatment take-up ($\pi_{01}(\mathbf{c})$) and the treatment effect on the outcome (iTME).

1.5 Identifying assumptions for causal mechanisms

Throughout we will make the following assumption:

Assumption 3. *Unconfoundedness of the clustered encouragement assignment*
Conditional on a set of covariates $\mathbf{C}_{ij}$, the encouragement status of each cluster, A_j, is independent of all the potential outcomes and the potential values of the treatment received:

$$\{Y_{ij}(a), M_{ij}(\tilde{a})\} \perp\!\!\!\perp A_j \mid \mathbf{C}_{ij} = \mathbf{c} \quad \forall \mathbf{c} \in \mathscr{C}, a = \tilde{a} = \{0,1\} \text{ and } \forall i, j$$

When the encouragement is randomized, unconfoundedness of the encouragement assignment holds without conditioning on covariates. This is actually the case in the KAHS study. We will henceforth maintain this assumption. It is worth to remark that assumption (3) implies that the encouragement is also unconfounded within principal strata and levels of baseline covariates, that is $Y_{ij}(a) \perp\!\!\!\perp A_j \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}$.

If strata memberships were known, this unconfoundedness assumption would allow to identify principal causal effects comparing the outcome under the two encouragement conditions of individuals with the same values of compliance status and covariates. Unfortunately we do not in general know which individuals are in which principal stratum. Individuals with the same observed value of the intermediate variables are in general mixtures of different principal strata. The monotonicity assumption (2) allows to identify the compliance status of some particular units: those units who are assigned to the control group and take the treatment are identified as always-takers, and, similarly, those who do not take the treatment under the encouragement are identified as never-takers. In randomized experiments with non-compliance, exclusion restriction assumptions is commonly assumed to point-identify (i.e., consistently estimate) principal causal effect for compliers. Exclusion restriction basically rules out the presence of net effects, therefore it cannot be invoked in these settings where these are effects of interest. Nevertheless, the use of Bayesian inference circumvents this identifiability problem because, even when causal estimands are intrinsically not fully identified, posterior distributions are always proper when proper priors are assumed. Weak identifiability is reflected though in the flatness of the posterior distribution. We will explain in more details the Bayesian inference procedures in the following sections. Once principal causal effects have been estimated, a full assessment of causal mechanisms, as defined in the previous section, requires a last step, which is the decomposition of $PCE(\mathbf{0},\mathbf{1},\mathbf{c})$ for compliers into $NEE^{\mathbf{0}}(\mathbf{0},\mathbf{1},\mathbf{c})$ and $iTME^{\mathbf{1}}(\mathbf{0},\mathbf{1},\mathbf{c})$.

NEEs and iTMEs involve potential outcomes of the form $Y_{ij}(a,m)$ and, in particular, causal estimands of interest in the malaria application are defined based on comparisons between $Y_{ij}\big(1,M_{ij}(0)\big)$, $Y_{ij}\big(0,M_{ij}(0)\big)$ and $Y_{ij}\big(1,M_{ij}(1)\big)$. Information about potential outcomes of this form for each unit is not in general in the data. In one specific experiment, where only the encouragement assignment is randomized, only one of all these possible potential outcome is ultimately observed, namely $Y_{ij}\big(A_j,M_{ij}(A_j)\big)$, where A_j is the encouragement status assigned to cluster j. Potential outcomes of the type $Y_{ij}\big(a,m\big)$, with m set to a particular value for all units or to $M_{ij}(\tilde{a})$, are observable only if $m \equiv M_{ij}(a)$, which occurs if the treatment receipt for unit ij is actually set to $M_{ij}(a)$ or if it is set to $M_{ij}(\tilde{a})$, with $\tilde{a} \neq a$, but for this unit $M_{ij}(0) \equiv M_{ij}(1)$, i.e. the unit is a never-taker or an always-taker. In these two cases the potential outcome is in the data and its expression collapses in $Y_{ij}(a)$. On the contrary, potential outcomes can never be not even potentially observed for units with $M_{ij}(a) \neq m$, hence in this case they are called *a priori counterfactuals* (Rubin, 2004). Therefore here the only problematic counterfactuals that is never observable is $Y_{ij}\big(1,M_{ij}(0)\big)$ for compliers. In fact, for never-takers and always-takers $Y_{ij}\big(1,M_{ij}(0)\big)$ is observable when $A_j = 1$, provided that $Y_{ij}\big(1,M_{ij}(0)\big) \equiv Y_{ij}\big(1,M_{ij}(1)\big)$, which is also the reason why dissociative causal effects coincide with net encouragement effects. Estimation of a priori counterfactuals would require an extrapolation from other individuals in the data. Assumptions allowing this kind of extrapolation are required. As previously discussed, the definition of causal effects such as iTME and NEE, is conceptually based on the existence of a possible intervention on the intermediate variable. If such intervention is at least conceivable then the intermediate variable M_{ij} can be considered as another assignment and we can define a multivariate assignment mechanism $p\big(A_j,M_{ij}|\mathbf{C}_{ij},Y_{ij}(0,0),Y_{ij}(0,1),Y_{ij}(1,0),Y_{ij}(1,1)\big)$. Intuitively the possibility of an extrapolation of the information across groups of individuals depends on the extent to which A_j and M_{ij} are independent of the values of the potential outcomes.

In the mediation literature it has widely been shown that non-parametric identification of population mediated and non-mediated effects, as the one defined with a priori counterfactuals, can be obtained under sequential ignorability assumptions (see Ten Have & Joffe (2012) for a review of the different specifications), which would translate in the setting of cluster randomized trials in the unconfoundedness of the cluster-specific intervention and the unconfoundedness of the intermediate variable conditional on the observed cluster-specific intervention and baseline covariates (VanderWeele, 2010b). The validity of these assumptions actually allows to extrapolate information on a priori counterfactuals from values of the observed outcome of other units. Sequential ignorability consists of two assumptions. We report here their expression in the setting of cluster-level interventions.

Assumption 4. Unconfoundedness of the encouragement assignment
Conditional on a set of covariates $\mathbf{C}_{ij}$, the encouragement status of each cluster, A_j, is independent of all the potential outcomes and the potential values of the treatment received:

$$\left\{Y_{ij}(a,m), M_{ij}(\tilde{a})\right\} \perp\!\!\!\perp A_j \mid \mathbf{C}_{ij} = \mathbf{c}, m \qquad \forall \mathbf{c} \in \mathscr{C}, m, a, \tilde{a} = \{0,1\} \quad \text{and } \forall i, j$$

This assumption is an extension of unconfoundedness assumption 3 and is satisfied when the encouragement is randomized.

Assumption 5. Conditional unconfoundedness of the treatment receipt
Conditional unconfoundedness of the treatment receipt requires that, after conditioning for a set covariates $\mathbf{C}_{ij}$ and the encouragement assignment, potential outcomes are independent of the potential values of the intermediate variable:

$$Y_{ij}(a,m) \perp\!\!\!\perp M_{ij}(\tilde{a}) \mid A_j = \tilde{a}, \mathbf{C}_{ij} = \mathbf{c} \qquad \forall \mathbf{c} \in \mathscr{C}, m, a, \tilde{a} = \{0,1\} \quad \text{and } \forall i, j$$

Essentially, assumption (3) rules out the presence of unmeasured confounders of

the relationships of A_j with M_{ij} and Y_{ij}, while assumption (5) prohibits unmeasured confounders of the relationships between A_j and Y_{ij} as well as measured or unmeasured confounders of the same relationships affected by the encouragement A_j. We should distinguish two types of counterfactuals: when $a = \tilde{a}$ and when $a \neq \tilde{a}$. In the former case, as already stressed, the potential outcome is potentially observable and its expression collapses in $Y_{ij}(a)$. In this case identification results depend on the sole assumption of unconfoundedness of the encouragement assignment. When this assumption hold its average in subgroups of the population within levels of covariate, $E[Y_{ij}(a) \mid \mathbf{C}_{ij} = \mathbf{c}]$ can be estimated by the mean of the observed outcomes of individuals under encouragement status $A_j = a$, $E[Y_{ij} \mid A_j = a, \mathbf{C}_{ij} = \mathbf{c}]$. On the contrary, counterfactuals of the type $Y_{ij}(a, M_{ij}(\tilde{a}))$, with $a \neq \tilde{a}$ can be identified only if the two sequential ignorability assumptions are satisfied and the identification expression is as follows:

$$\mathrm{E}\big[Y_{ij}\big(a, M_{ij}(\tilde{a})\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big] = \sum_{m=0}^{1} \mathrm{E}\big[Y_{ij} \mid A_j = a, M_{ij} = m, \mathbf{C}_{ij} = \mathbf{c}\big] \times P\big(M_{ij} = m \mid A_j = \tilde{a}, \mathbf{C}_{ij} = \mathbf{c}\big) \tag{1.5.1}$$

For the proof see Pearl (2001, 2011) and Imai (2010b).

The critical feature of evaluating causal mechanisms in cluster randomized encouragement designs (CED) is that even if the experiment randomizes the encouragement, the intermediate variable, i.e., the actual treatment received, is instead self-selected by individuals. Consequently, unconfoundedness of the intermediate variable required by the sequential ignorability assumption is unlikely to hold, even conditioning on observed covariates, because of possible unmeasured factors confounding the relation between M_{ij} and Y_{ij}. In fact, in our empirical study, household's decision of carrying out the purchase of new bed nets depends on observed but, presumably, also on unobserved characteristics. Here we propose the use of Principal Stratification approach to , primarily, estimate the overall effect of the clustered encouragement for each principal stratum and, subsequently, recover the individual treatment mediated effect and the net encouragement effect for all strata,

without relying on sequential ignorability assumptions.

1.5.1 Homogeneity Assumptions

When the sequential ignorability assumption does not hold, information about a prior counterfactuals for compliers cannot be extrapolated across strata and thus principal causal effect for this sub-population cannot be decomposed into the two causal mechanisms of interest. Here we provide two alternative homogeneity assumptions that enable us to make use of the information available in the strata S^{mm}, with $m = 0, 1$, where all potential outcomes are observable, to estimate a priori counterfactuals in other strata. Essentially, these assumptions concern solely the missing information and allow only the extrapolation that is strictly needed across strata with a similar compliance behavior at least under one encouragement condition, in contrast with the stronger assumption of sequential ignorability that enables a greater extrapolation across strata. For the sake of clarity, here we focus on identification of the effects of interest for the application of bed nets, that is, $NEE^0(0,1,\mathbf{c})$ and $iTME^1(0,1,\mathbf{c})$, where the only a priori counterfactual is $Y_{ij}\big(1, M_{ij}(0)\big)$ for compliers. In the appendix, we provide a generalization of these homogeneity assumptions, for identification of $NEE^{\tilde{a}}(0,1,\mathbf{c})$ and $iTME^{1-\tilde{a}}(0,1,\mathbf{c})$, with $\tilde{a} \in \{0,1\}$, and not restricted to the monotonicity assumption. The proofs of the theorems are reported in the appendix for the general case.

Assumption 6. *Stochastic Homogeneity of the Counterfactual across Never-Takers and Compliers*

Stochastic homogeneity of the counterfactual $Y_{ij}\big(1, M_{ij}(0)\big)$ across never-takers and compliers is said to be assumed if the following conditional independence holds:

$$Y_{ij}(1,0) \perp\!\!\!\perp M_{ij}(1) \mid M_{ij}(0) = 0, \mathbf{C}_{ij} = \mathbf{c} \qquad \forall \mathbf{c} \in \mathscr{C} \text{ and } \forall i, j$$

Assumption (6) coveys the idea that the distribution of the counterfactual $Y_{ij}\big(1, M_{ij}(0)\big)$, which corresponds to $Y_{ij}(1,0)$ for never-takers and compliers, is that same for these two principal strata, conditioning on baseline covariates.This allows to estimate the a priori counterfactual $Y_{ij}\big(1, M_{ij}(0)\big)$ for compliers using the information on $Y_{ij}(1,0)$ provided by never-takers assigned to $A_j = 1$, for whom we observe $Y_{ij}(1)$. This assumption is neither testable nor can find support in the data. If never-takers and compliers share the same conditional distribution of the potential outcome $Y_{ij}\big(0, M_{ij}(0)\big)$, we could assume that it is also true when encouragement is set to the opposite condition. However, this is neither a sufficient nor a necessary condition.

Theorem 1. *If assumption 6 holds, the net encouragement effect for compliers within levels of covariates, $NEE^0(0,1,\mathbf{c})$, is given by:*

$$NEE^0(0,1,\mathbf{c}) = \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{00}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big]$$

Often assumption (6) with $a \neq \tilde{a}$ cannot be supported, especially when the data do not provide evidence on the equality of the distribution of $Y_{ij}(0)$ for never-takers and compliers, even within the same levels of covariates. For example in KAHS study never-takers and always-takers can be substantially different households. Therefore, we will provide an alternative assumption that might be more reasonable in some applications.

Assumption 7. *Homogeneity of Mean Difference between Counterfactuals for Never-takers and Compliers:*

$$\mathrm{E}\big[Y_{ij}(1,0) - Y_{ij}(0,0) \mid M_{ij}(0) = 0, M_{ij}(1), \mathbf{C}_{ij} = \mathbf{c}\big] = \mathrm{E}\big[Y_{ij}(1,0) - Y_{ij}(0,0) \mid M_{ij}(0) = 0, \mathbf{C}_{ij} = \mathbf{c}\big] \quad \forall \mathbf{c} \in \mathscr{C}$$

Assumption 7 states that the average difference of potential outcomes under the two encouragement conditions and intervening to set the treatment receipt of each unit

to 0, is the same for all those with $M_{ij}(0) = 0$, that is those who would not take the treatment if A_j were set to 0, i.e. never-takers and compliers, and is independent of the potential treatment receipt under the opposite encouragement status, $M_{ij}(1)$. In KAHS study this means that households that would not buy any new bed net without loans, would have the same average effect of the offer of the program to their cluster on the reduction of risk of infection, if we intervened to keep their number of bed nets bought at follow-up fixed at 0, regardless of their behavior under the control condition. Given this assumption we are able to introduce the following theorem:

Theorem 2. *If assumption 7 is satisfied, the net encouragement effect for compliers,* $NEE^0(0,1,\mathbf{c})$*, within levels of covariates, can be extrapolated from the dissociative causal effect for never-takers:*

$$NEE^0(0,1,\mathbf{c}) \equiv DCE(0,\mathbf{c})$$

The effect of the encouragement is the same for never-takers and compliers, intervening to set M_{ij} to 0 or in other words to prevent any purchase of new bed nets. Assumption (7) allows then to estimate $NEE^0(0,1,\mathbf{c})$ for compliers and hence $NEE^0(\mathbf{c})$ in the entire population.

Assumptions (6) and (7) provide the possibility of a generalization of the potential outcome $Y_{ij}\big(1, M_{ij}(0)\big)$ or the net encouragement effect NEE^0 from never-takers to compliers, as stated by the theorems 1 and 2. As a fair consequence, these assumptions also yield identification of the individual treatment mediated effect in the latter principal stratum, $iTME^1(0,1,\mathbf{c})$.

Corollary 1. *If assumption* (6) *holds the individual treatment mediated effect for compliers,* $iTME^1(0,1,\mathbf{c})$*, within levels of covariates, is given by:*

$$iTME^1(0,1,\mathbf{c}) = PCE(0,1,\mathbf{c}) - \Big(\mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{00}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big]\Big)$$

If assumption (7) *holds the individual treatment mediated effect for compliers,* $iTME^1(0,1,\mathbf{c})$*, within levels of covariates, is given by:*

$$iTME^1(0,1,\mathbf{c}) = PCE(0,1,\mathbf{c}) - DCE(0,\mathbf{c})$$

1.5.2 Average Treatment Effect

In a canonical non-compliance setting the main goal is to estimate the average treatment effect (ATE), i.e., the average effect of the non-randomized treatment on the outcome. The average treatment effect in the entire population, within levels of covariates, can be defined as the following difference:

$$\begin{aligned} ATE^a(\mathbf{c}) &:= \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid \mathbf{C}_{ij} = \mathbf{c}\big] \\ &= \sum_{(m_0=m_1)} \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{m_0m_1}, \mathbf{C}_{ij} = \mathbf{c}\big]\pi_{m_0m_1}(\mathbf{c}) \\ &\quad + \sum_{(m_0\neq m_1)} \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{m_0m_1}, \mathbf{C}_{ij} = \mathbf{c}\big]\pi_{m0m_1}(\mathbf{c}) \end{aligned} \tag{1.5.2}$$

where the last expression simply expands the definition taking an average of the specific average treatment effects within the different principal strata. Referring to the definition in (1.5.2) we have to make two main considerations. First, we can see that the average treatment effects in general depends on the specific value of a we consider for the encouragement condition, while we compare the two scenarios where the treatment is or is not taken. The possible difference between $\mathrm{ATE}^0(\mathbf{c})$ and $ATE^1(\mathbf{c})$ is due to the interaction between the encouragement and the individual treatment uptake on the outcome. In clustered encouragements it can also be due to the interaction of the individual treatment uptake with other behavioral changes in other subjects in the same cluster. Second, unfortunately the empirical data do not provide any information on the treatment effect for principal strata where the treatment uptake is unaffected by the encouragement assignment if $M_{ij}(0) = M_{ij}(1) = 0$ because there is no individual information on the counterfactual $Y_{ij}(a,1)$ and vice versa for the symmetric stratum. The only strata where we could learn something

about the treatment effect are those where $M_{ij}(0) \neq M_{ij}(1)$.

Let us define the complier average causal effect (CACE), i.e. the average treatment effect for compliers, within levels of covariates, as follows:

$$CACE^{a}(\mathbf{c}) := \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{1.5.3}$$

Because of non-compliance the treatment is not randomized. Instrumental variable methods use the effect of the assignment on the the treatment receipt to recover the average treatment effect from the intention-to-treat analysis. Typically, these methods appeal to exclusion restriction assumptions, which substantially rule out the presence of net effects. Formally, the exclusion restriction assumption for a stratum $S^{m_0 m_1}$ states that $Y_{ij}(a,m) = Y_{ij}(a,m') \;\; \forall i,j : S_{ij} = S^{m_0 m_1}$, which implies the same equality in terms of the mean outcome and thus zero net effects for this principal stratum. Assumptions of exclusion restriction for always-takers and never-takers jointly with monotonicity of compliance result in the point identification of the principal causal effect for compliers, whereas exclusion restriction for compliers enables to interpret it as the average treatment effect for this sub-population, also known as compliers average causal effect (CACE). For this same reason, when exclusion restriction for compliers applies, CACE can be written in terms of principal causal effect: $\mathrm{E}\big[Y_{ij}(1) - Y_{ij}(0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big]$.

Nevertheless, when exclusion restriction assumptions are violated, if assumption (6) or (7) hold, the resulting identification of the individual treatment mediated effect for compliers $iTME^{1}(0,1,\mathbf{c})$ will also yield identification of $CACE^{1}(\mathbf{c})$, given the following equality:

$$CACE^{1}(\mathbf{c}) \equiv iTME^{1}(0,1,\mathbf{c}) \tag{1.5.4}$$

Proof.

$$\begin{aligned} CACE^{a}(\mathbf{c}) &= \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] \\ &= \mathrm{E}\big[Y_{ij}\big(a, M_{ij}(1)\big) - Y_{ij}\big(a, M_{ij}(0)\big) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\big] = iTME^{a}(0,1,\mathbf{c}) \end{aligned}$$

□

When we are evaluating a new treatment that cannot be randomized and we use the encouragement as an instrument, the effect of primary interest is $CACE^0(\mathbf{c})$. In that case assumptions similar to (6) and (7) are needed (see section A 1 in the appendix for a generalization of the assumptions). Alternatively, when the treatment effect has already been assessed in previous experiments, that is $CACE^0(\mathbf{c})$ is already known, and an encouragement, designed to increase or decrease its uptake, is the intervention of interest, the estimated $CACE^1(\mathbf{c})$ will give insight into how the encouragement itself changes the effect of the treatment on the outcome. This is the case when the treatment is the purchase of new bed nets.

1.6 Hierarchical Models for Cluster Interventions

In this section we describe the models used for our analysis: a model for the outcome and a model for the principal strata membership. Because of the cluster-level randomization the use of the hierarchical framework is needed. In cluster randomized trials, when the unit of intervention is a community or a group of individuals, we cannot ignore correlation among individuals arising from common environmental factors and even reciprocal influence. Failure in taking this correlation into account may lead to wrong inference conclusions in terms of standard errors. In our setting, individuals living in the same community are likely to show resemblance not only in terms of outcomes, but also in terms of individual treatment uptake. In fact, individual compliance in participating to the program offered in the community may be related not only to individual characteristics, but also to the cluster environment and not least to reciprocal peer influence. Further, the level of resemblance in outcomes may vary across different individual strata. Correlation in cluster randomized trials with individual non-compliance has been intensively studied by Jo (2008), after Frangakis, Rubin & Zhou (2002), who were the first authors to accommodate in their analysis correlation in both outcome and non-compliance status. Here we

extend the model framework used by Frangakis, Rubin & Zhou (2002).

Potential Outcome Model

We will report here the model used to analyze the particular application of KAHS study. We want to emphasize, though, that the general framework presented here can be also used for all kinds of outcome models. In our malaria example, the outcome of interest, Y_{ij}, is the proportion of malaria cases that household i in cluster j has experienced in the month prior to the follow-up interview. Therefore, we assume a relative binomial distribution for the potential outcomes of the form $Y_{ij}(a)$

$$Y_{ij}(a)|S_{ij},\mathbf{C}_{ij} \sim \frac{Bin\left(n_{ij},p_{ij}\right)}{n_{ij}} \tag{1.6.1}$$

and we specify a hierarchical generalized linear model for the probability $p_{ij} = p_{ij}\left(a,S_{ij},\mathbf{C}_{ij}\right)$, as a function of the encouragement $A_j = a$, the principal stratum S_{ij} and the vector of covariates $\mathbf{C}_{ij}$:

$$\begin{gathered} g\left(p_{ij}\left(a,S_{ij},\mathbf{C}_{ij}\right)\right) = \boldsymbol{\beta}^{S_{ij}T}\mathbf{Z}_{ij}^{Yf} + \boldsymbol{b}_j^T\mathbf{Z}_{ij}^{Yr} = \boldsymbol{\beta}_0^{S_{ij}T}\mathbf{C}'_{ij} + \boldsymbol{\beta}_1^{S_{ij}T}\mathbf{C}'_{ij}a + b_{0j} + \mathbf{b}_{1j}^T\mathbf{X}_{ij} \\ \mathbf{b}_j \sim N(\mathbf{0},\Sigma_b) \end{gathered} \tag{1.6.2}$$

where $\mathbf{C}'_{ij} = (1,\mathbf{C}_{ij})$, $g(\cdot)$ is a link function, $\boldsymbol{\beta}^{S_{ij}}$ are the fixed effects for each principal stratum and $\mathbf{b}_j$ are the random effects, with variable vectors $\mathbf{Z}_{ij}^{Yf} = \left[1,\mathbf{C}_{ij},a,\mathbf{C}_{ij}a\right]$ and $\mathbf{Z}_{ij}^{Yr} = \left[1,\mathbf{X}_{ij}\right]$ respectively, allowing for random intercepts and random individual covariates slopes. We also assume that the two potential outcomes $Y_{ij}(0)$ and $Y_{ij}(1)$ are independent, given the covariates and strata membership.

Principal Strata Model

Principal strata membership can also be modeled by a hierarchical generalized linear model to take into account cluster correlation in individual treatment:

$$g\Big(P\big(S_{ij}=S^{m_0m_1}|\mathbf{C}_{ij}\big)\Big)=\boldsymbol{\alpha}^T\mathbf{Z}_{ij}^{Sf}+\mathbf{a}_j^T\mathbf{Z}_{ij}^{Sr}=\boldsymbol{\alpha}^T\mathbf{C}'_{ij}+\mathrm{a}_{0j}+\mathbf{a}_{1j}^T\mathbf{X}_{ij}$$
$$\mathbf{a}_j\sim N(\mathbf{0},\Sigma_a) \tag{1.6.3}$$

where $g(\cdot)$ is the link function, $\alpha^{S_{ij}}$ are the fixed effects and $\mathbf{a}_j$ are the random effects, with variable vectors $\mathbf{Z}_{ij}^{Sf}=\big[1,\mathbf{C}_{ij}\big]$ and $\mathbf{Z}_{ij}^{Sr}=\big[1,\mathbf{X}_{ij}\big]$ respectively, assuming covariate $\mathbf{C}_{ij}$ to be predictors of strata membership.

Here we follow the approach used in Frangakis, Rubin & Zhou (2002) and Barnard et al. (2003), who modeled the strata membership using an *Ordinal Probit Model*. In general in an ordinal probit model for an ordinal outcome with L categories the probability of belonging to a category lower than l is modeled as $P(Y_i\le l)|\ \mathbf{C}_{ij})=\Phi\big(\boldsymbol{\alpha}_l\ \mathbf{C}_{ij})\big)$, with $l=1,\ldots,L-1$, so that the probability of belonging to the category l ends up being $P(Y_i=l|\ \mathbf{C}_{ij})=\big(P(Y_i\le l+1)|\ \mathbf{C}_{ij})\big)\big(1-\Phi\big(\boldsymbol{\alpha}_l\ \mathbf{C}_{ij})\big)\big)$. The function $\Phi(\cdot)$ is the standard normal cumulative distribution function.

According to this parametrization here we illustrate the ordinal probit model for S_{ij} when monotonicity is assumed, so that we end up with three strata with two linked probit models, the first modeling membership in the never-taker stratum and the second modeling membership in the complier stratum conditional on not being a never-taker. In our setting of custer-based intervention we extend the above model to an *Ordinal Mixed Probit Model*, parameterized as:

$$\begin{aligned}
&\Psi_n(\mathbf{C}_{ij},\boldsymbol{\alpha},\mathbf{a})=P(S_{ij}=S^{00})|\mathbf{C}_{ij})=1-\Phi\big(\boldsymbol{\alpha}_n^TZ_{ij}^{Sf}+\mathbf{a}_{nj}^TZ_{ij}^{Sr}\big)\\
&\Psi_c(\mathbf{C}_{ij},\boldsymbol{\alpha},\mathbf{a})=P(S_{ij}=S^{01})|\mathbf{C}_{ij})=\big(1-\Psi_n(\mathbf{C}_{ij},\boldsymbol{\alpha},\mathbf{a})\big)\big(1-\Phi\big(\boldsymbol{\alpha}_c^TZ_{ij}^{Sf}+\mathbf{a}_{cj}^TZ_{ij}^{Sr}\big)\big)\\
&\Psi_a(\mathbf{C}_{ij},\boldsymbol{\alpha},\mathbf{a})=P(S_{ij}=S^{11})|\mathbf{C}_{ij})=1-\Psi_n(\mathbf{C}_{ij},\boldsymbol{\alpha},\mathbf{a})-\Psi_c(\mathbf{C}_{ij},\boldsymbol{\alpha},\mathbf{a})
\end{aligned} \tag{1.6.4}$$

with $\boldsymbol{\alpha}=\big(\boldsymbol{\alpha}_n,\boldsymbol{\alpha}_c\big)$ and $\mathbf{a}=\big(\mathbf{a}_n=\big(\mathbf{a}_{n1},\ldots,\mathbf{a}_{nJ}\big),\mathbf{a}_c=\big(\mathbf{a}_{c1},\ldots,\mathbf{a}_{cJ}\big)\big)$ and

$$\mathbf{a}_{nj}\sim N\big(\mathbf{0},\Sigma_{a_n}\big)\quad \mathbf{a}_{cj}\sim N\big(\mathbf{0},\Sigma_{a_c}\big)$$

The above model has an equivalent formulation as a latent-variable model. In this formulation the two probit models are represented as arising from two underlying

continuous random variables S_{ij}^{n} and S_{ij}^{c}:

$$S_{ij} = \begin{cases} S^{00} & \text{if} \quad S_{ij}^{n} \equiv \boldsymbol{\alpha}_{n}^{T} Z_{ij}^{Sf} + \mathbf{a}_{nj}^{T} Z_{ij}^{Sr} + V_{ij} \leq 0 \\ S^{01} & \text{if} \quad S_{ij}^{n} \geq 0 \text{ and } S_{ij}^{c} \equiv \boldsymbol{\alpha}_{c}^{T} Z_{ij}^{Sf} + \mathbf{a}_{cj}^{T} Z_{ij}^{Sr} + U_{ij} \leq 0 \\ S^{11} & \text{if} \quad S_{ij}^{n} \geq 0 \text{ and } S_{ij}^{c} \geq 0 \end{cases} \tag{1.6.5}$$

where U_{ij} and V_{ij} are independently distributed as $N(0,1)$. The latter formulation is going to facilitate computation later.

1.7 Bayesian Inference

Let A_{j}^{obs} be the observed encouragement assigned to cluster j. Assuming that all the potentially observable information for each cluster is in the random vector $\left(A_j, \mathbf{C}_j, \mathbf{M}_j^{obs}, \mathbf{M}_j^{mis}, \mathbf{Y}_j^{obs}, \mathbf{Y}_j^{mis}\right)$, where each vector with subscript j contains the corresponding variable for all the units in cluster j, whereas we denote with superscript *obs* and *mis*, respectively, the observed and missing but observable potential outcomes, that is: $\mathbf{Y}_j^{obs} \equiv \mathbf{Y}_j(A_j)$, $\mathbf{Y}_j^{mis} \equiv \mathbf{Y}_j(1-A_j)$, $\mathbf{M}_j^{obs} \equiv \mathbf{M}_j(A_j)$ and $\mathbf{M}_j^{mis} \equiv \mathbf{M}_j(1-A_j)$. As extensively discussed, counterfactuals of the form $Y_{ij}\left(a, M_{ij}(\tilde{a})\right)$ are never observable unless $M_{ij}(\tilde{a}) \equiv M_{ij}(a)$. Under assumptions (6) or (7) presented above, all the causal estimands depend solely on the observable potential outcomes Y_{ij}^{obs} and Y_{ij}^{mis} of individuals belonging to each principal stratum. Therefore we can assume that all the missing information required for each cluster is contained in the vectors $\left(\mathbf{M}_j^{mis}, \mathbf{Y}_j^{mis}\right)$.

In particular, Bayesian inference for causal estimands, which are functions of $\left(\mathbf{M}^{obs}, \mathbf{M}^{mis}, \mathbf{Y}^{obs}, \mathbf{Y}^{mis}, \mathbf{C}\right)$, follows from their joint posterior predictive distribution, that is their conditional distribution given the observed data, which can be written as the product of independently identically distributed random variables conditional on a generic parameter $\boldsymbol{\theta}$ (de Finetti, 1974). Let $\boldsymbol{\theta}$ denote the vector of parameters

of the models described above:

$$\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{b}, \mathbf{a}, \Sigma_b, \Sigma_a\right)$$

where we have collected each set of parameters such that $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{S^{00}}, \boldsymbol{\beta}^{S^{11}}, \boldsymbol{\beta}^{S^{01}}\right)$, $\boldsymbol{b} = \left(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_J\right)$, and $\Sigma_a = \left(\Sigma_{a_n}, \Sigma_{a_c}\right)$.

The posterior distribution of $\boldsymbol{\theta}$ can be written from the joint distribution, mentioned above, marginalized over the missing values:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}, \mathbf{M}^{obs},\ \mathbf{C}\ ,\ \mathbf{A}) \propto p(\boldsymbol{\theta}) \int\int \prod_{j=1}^{J} p\left(\mathbf{Y}_j^{obs}, \mathbf{M}_j^{obs}, \mathbf{Y}_j^{mis}, \mathbf{M}_j^{mis},\ \mathbf{C}_j \mid \boldsymbol{\theta}\right) d\mathbf{Y}_j^{mis} d\mathbf{M}_j^{mis} \tag{1.7.1}$$

which is a result of randomization of assignment $\mathbf{A}$ (assumption 3) and the independence between clusters (assumption 1) and where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters $\boldsymbol{\theta}$. The difficulty in the integration over $\mathbf{M}_j^{mis}$ leads us to consider the joint posterior of $\left(\boldsymbol{\theta}, \mathbf{M}^{mis}\right)$, or alternatively the joint posterior of $\left(\boldsymbol{\theta}, \mathbf{S}\right)$:

$$p(\boldsymbol{\theta}, \mathbf{S} \mid \mathbf{Y}^{obs},\ \mathbf{C}\ ,\ \mathbf{A}) \propto p(\boldsymbol{\theta}) \prod_{j=1}^{J} p\left(\mathbf{Y}_j^{obs}, \mathbf{S}_j,\ \mathbf{C}_j, \mathbf{A} \mid \boldsymbol{\theta}\right) \tag{1.7.2}$$

which follows from the assumed independence between the potential outcomes.

The second term in (1.7.2) is the complete-data likelihood function, which results in the likelihood function of a finite mixture model with known membership, unlike the observed likelihood where the strata membership is unknown. The complete-data likelihood function, namely $\mathscr{L}(\boldsymbol{\theta}; \mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C}, \mathbf{A}) := p(\mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C}, \mathbf{A} \mid \boldsymbol{\theta})$, can be factorized in $p(\mathbf{Y}^{obs} \mid \mathbf{S}, \mathbf{C}\ \boldsymbol{\theta}) p(\mathbf{S} \mid \mathbf{C},\ \boldsymbol{\theta}) p(\mathbf{C} \mid \boldsymbol{\theta})$. We will assume throughout that the vector of random effects $\mathbf{b}_j$ accounts for all the unmeasured common factors affecting the outcome of all the units in cluster j, as well as unmeasured individual post-intermediate variables of every unit in the cluster affecting not only the unit's final outcome but also his neighbors', including the unit's outcome measured at pre-

vious time points or other behavioral characteristics. As a consequence, we make the assumption of independence between units' potential outcomes, conditioning on $\mathbf{b}_j$. As a result we have the further factorization of the the complete-data likelihood function:

$$\mathscr{L}(\boldsymbol{\theta};\mathbf{Y}^{obs},\mathbf{S},\mathbf{C},\mathbf{A}) = \prod_{j=1}^{\mathrm{J}}\prod_{i=1}^{N_j} p\big(Y_{ij}|\, A_j,\mathbf{S},\mathbf{C},\ \boldsymbol{\theta}\big) \times P\big(\mathbf{S}\,|\,\mathbf{C},\boldsymbol{\theta}\big)p(\mathbf{C}\,|\,\boldsymbol{\theta}) \tag{1.7.3}$$

where assumption of consistency (1) has been used to express the distribution of the observed potential outcome in terms of the distribution of the observed values.

Letting $\delta_{ij}(S^{m_0m_1}) = \delta(S^{m_0m_1}, S_{ij})$ be 1 if $S_{ij} = S^{m_0m_1}$ and 0 otherwise, we can write:

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta};\mathbf{Y}^{obs},\mathbf{S},\mathbf{C},\mathbf{A}) = \prod_{j=1}^{\mathrm{J}}\prod_{i=1}^{N_j}\sum_{m_0m_1}\delta_{ij}(S^{m_0m_1})p\big(Y_{ij}|\, A_j, S_{ij}=S^{m_0m_1},\mathbf{C}_{ij},\ \boldsymbol{\theta}\big)\\ \times P\big(S_{ij}=S^{m_0m_1}\,|\,\mathbf{C}_{ij},\boldsymbol{\theta}\big)p(\mathbf{C}_{ij}\,|\,\boldsymbol{\theta})\end{aligned} \tag{1.7.4}$$

The two models involved in the likelihood for Y_{ij} and S_{ij} have already been defined in (1.6.1) and (1.6.4) respectively. The complete-data likelihood allows the full conditional distributions $p(\boldsymbol{\theta}\,|\,\mathbf{Y}^{obs},\mathbf{S},\mathbf{C}\ ,\ \mathbf{A})$ and $p(\mathbf{S}\,|\,\mathbf{Y}^{obs},\ \mathbf{C}\ ,\ \mathbf{A},\boldsymbol{\theta})$ to be analytically tractable. Therefore, the joint posterior distribution of $\big(\boldsymbol{\theta},\mathbf{S}\big)$ motivates a two-stage Gibbs-sampling strategy that first samples the missing strata memberships S_{ij}, thereby allowing assessment of the distributions of Y_{ij} conditional on the complete data consisting of subpopulations without mixture components. This approach is well known as the *Data Augmentation* scheme (Tanner & Wong, 1987). See the appendix A 4 for the detailed Gibbs-Sampling procedure.

1.7.1 Prior Specification

Here we describe our prior distribution $p(\boldsymbol{\theta})$. We assume an independence structure expressed in the following factorization of the prior:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}) \prod_j p(\mathbf{b}_j \mid \Sigma_b) p(\Sigma_b) p(\boldsymbol{\alpha}_n) p(\boldsymbol{\alpha}_c) \prod_j p(\mathbf{a}_{nj} \mid \Sigma_{a_n}) p(\Sigma_{a_n}) p(\mathbf{a}_{cj} \mid \Sigma_{a_c}) p(\Sigma_{a_c}) \tag{1.7.5}$$

where Σ_{a_n} and Σ_{a_c} are the submatrices of Σ_a corresponding to the covariance matrices of vectors $\mathbf{a}_n$ and $\mathbf{a}_c$, thought independent. It follows that the random effects $\mathbf{a}_{nj}$, $\mathbf{a}_{cj}$ and $\mathbf{b}_j$ are independent across groups as well as of the coefficients of each probit model and of the model for Y_{ij}. We have chosen to use proper but diffuse priors similar, in order to be relatively noninformative and to ensure fast convergence. Accordingly, we posit a normal prior distribution for the coefficients of the outcome model. The fixed effects can be jointly modeled as

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0}\right) \tag{1.7.6}$$

whereas the random effects are modeled independently for each cluster

$$\mathbf{b}_j \mid \Sigma_b \sim N(\mathbf{0}, \Sigma_b) \tag{1.7.7}$$

with the covariance matrices following an inverse-Wishart distribution:

$$\Sigma_b \sim IW\left(\eta_0^b, \eta_0^b S_0^b\right) \tag{1.7.8}$$

Typical hyper-parameters can be: $\boldsymbol{\mu}_{\beta 0} = \mathbf{0}$, $\Lambda_{\beta 0} = \xi^b$ I, where ξ^b is a scaling parameter, $\eta_0^b = |\mathbf{b}_j|$ and S_0^b is a preliminary estimates of Σ_b.

The parameters of the models for the principal strata follow the same patterns, although property of conjugacy can here be satisfied. Thus, for the two vectors of fixed effects of both models, we choose a prior normal distribution

$$\boldsymbol{\alpha}_n \sim N\left(\boldsymbol{\mu}_{\alpha 0}^n, \Lambda_{\alpha 0}^n\right) \quad \boldsymbol{\alpha}_c \sim N\left(\boldsymbol{\mu}_{\alpha 0}^c, \Lambda_{\alpha 0}^c\right) \tag{1.7.9}$$

as well as for the random effects

$$\mathbf{a}_{nj} \mid \Sigma_{a_n} \sim N\left(\mathbf{0}, \Sigma_{a_n}\right) \quad \mathbf{a}_{cj} \mid \Sigma_{a_c} \sim N\left(\mathbf{0}, \Sigma_{a_c}\right) \tag{1.7.10}$$

with an inverse-Wishart prior for covariances matrices

$$\Sigma_{a_n} \sim IW\left(\eta_0^n, \eta_0^n S_0^n\right) \quad \Sigma_{a_c} \sim IW\left(\eta_0^c, \eta_0^c S_0^c\right) \tag{1.7.11}$$

with the following possible choices for the hyper-parameters: $\boldsymbol{\mu}_{\alpha 0}^n = \boldsymbol{\mu}_{\alpha 0}^c = \mathbf{0}$, $\Lambda_{\alpha 0}^n = \Lambda_{\alpha 0}^c = \xi$ I, $\eta_0^n = |\mathrm{a}_{nj}|$, $\eta_0^c = |\mathrm{a}_{cj}|$ and S_0^n and S_0^c are preliminary estimates of Σ_{a_n} and Σ_{a_c} respectively.

1.7.2 Imputation Approach for Finite Population Effects

We introduce now a bayesian procedure for the estimation of the effects in the finite study population. For the sake of simplicity, we will describe the procedure only for the estimation of the effects of interest for the motivating application, although a similar procedure could be used in future applications for the other effects. We define individual effects as the difference of the corresponding potential outcomes for each unit in the study. Thus, the intent-to-treat effect, the net encouragement effect and the individual treatment mediated effect for unit i in cluster j take the following expressions: $ITT_{ij} := Y_{ij}(1) - Y_{ij}(0)$, $NEE_{ij}^0 := Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))$ and $iTME_{ij}^1 := Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))$. For each unit, one of the two potential outcomes involved in the intent-to-treat effect is observed, $Y_{ij}^{obs} = Y_{ij}(A_j^{obs})$, whereas for NEE and iTME all potential outcomes can be missing and one can be a priori counterfactual. Relying on one of the two homogeneity assumptions, we show how estimation of the finite population effects can be accomplished. Let $\mathcal{O}$ be the collection of observed outcomes, observed intermediated variables, encouragement conditions and covariates in the entire population: $\mathcal{O} = \left\{\mathbf{Y}^{obs}, \mathbf{M}^{obs}, \mathbf{A}^{obs}, \mathbf{C}\right\}$.

Bayesian simulation-based approach enables to simulate from the posterior distri-

butions of the causal estimands. In a model-based imputation approach to causal inference missing information for each unit is imputed using its predictive posterior distribution and causal estimands, as function of the observed and missing infomation, are computed resulting in a draw from their posterior distribution. Let $f_{m_0 m_1}(a \mid \mathbf{c})$ denote the predictive posterior distribution of the potential outcome $Y_{ij}(a)$:

$$f_{m_0 m_1}(a \mid \mathbf{c}) = p\big(Y_{ij}(a) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c},\ \mathscr{O}\big) \tag{1.7.12}$$

We can easily generate replicates of $Y_{ij}(a)$ from the posterior predictive distribution of each principal stratum by adding a simple step within the MCMC using the conditional distribution of the potential outcomes, evaluated at parameter values $\boldsymbol{\theta}^k$:

$$f_{m_0 m_1}(a \mid \mathbf{c}, \boldsymbol{\theta}^k) = p\big(Y_{ij}(a) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}, \boldsymbol{\theta}^k\big) \tag{1.7.13}$$

This result follows from $f_{m_0 m_1}(a, \mathbf{c}) = \int p\big(Y_{ij}(a) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}, \boldsymbol{\theta}\big) p(\theta \mid \mathscr{O}) d\boldsymbol{\theta}$. At each iteration $k = 1, \ldots, K$ of the MCMC, samples from the posterior distribution of *PCE* for each principal stratum $S^{m_0 m_1}$ are drawn as follows:

1. For units belonging to $S^{m_0 m_1}$ at iteration k, missing potential outcomes, $Y_{ij}^{mis} = Y_{ij}(1 - A_j^{obs})$, are imputed from their conditional distribution:

$$Y_{ij}^{k,mis} \sim f_{m_0 m_1}\big(1 - A_j^{obs} \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k\big) \qquad \forall i, j : S_{ij}^k = S^{m_0 m_1}$$

2. PCE within each principal stratum $S^{m_0 m_1}$ is computed as:

$$\widehat{PCE}^k(m_0, m_1, \mathbf{c}) = \frac{1}{|\mathscr{S}_c^{m_0 m_1}|} \sum_{i,j \in \mathscr{S}_c^{m_0 m_1}} \big(2A_j^{obs} - 1\big)\big(Y_{ij}^{obs} - Y_{ij}^{k,mis}\big)$$

where $\mathscr{S}_c^{m_0 m_1} = \{i, j : S_{ij}^k = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\}$. If the number of covariates is large and/or they are continuous we might want to categorize some of them and/or consider groups $\mathscr{S}_c^{m_0 m_1}$ defined based on few covariates for which a subgroup analysis might be of interest.

Let us now turn to the analysis of mechanisms. As depicted in (1.4.7), for principal strata of the type S^{mm}, i.e., never-takers and always-takers, there is no effect through a change in the treatment received and principal causal effects are called dissociative causal effects, $\widehat{DCE}(m, \mathbf{c}) = \widehat{PCE}(m_0, m_1, \mathbf{c})$, as they are entirely net encouragement effects. On the contrary for the stratum S^{01} of compliers, which is, under monotonicity, the only stratum where the treatment is affected by the encouragement, the overall effect of the encouragement comprises both individual treatment effect from the net encouragement effect. With sequential ignorability (5) not holding, disentangling these two effects for this stratum can be accomplished under one of the two assumptions (6) or (7). In general we can separate the derivation of $NEE^0(0, 1, \mathbf{c})$ into two three steps. The first two steps involve, respectively, the counterfactual $Y_{ij}(0) = Y_{ij}(0, M_{ij}(0))$ and $Y_{ij}(1, M_{ij}(0))$, whereas the third step concerns the mean difference.

3. For each unit being a complier at iteration k, the potential outcome $Y_{ij}^{\,k}(0)$ is derived as follows: if assumption (6) holds, $Y_{ij}^{\,k}(0)$ is simply taken from $Y_{ij}^{\,obs}$ or $Y_{ij}^{\,mis}$, depending on A_j^{obs}; if assumption (7) holds, in order to follow the identification result in theorem 2, $Y_{ij}^{\,k}(0)$ is imputed from the conditional distribution of $Y_{ij}(0)$ for never-takers, given his values of covariates $\mathbf{C}_{ij}$:

$$Y_{ij}^{\,k}(0): \begin{cases} \text{3a. } \textit{if assumption 6: } Y_{ij}^{\,k}(0) = Y_{ij}^{\,obs} \cdot (1 - A_j^{obs}) + Y_{ij}^{\,k,mis} \cdot A_j^{obs} \\ \text{3b. } \textit{if assumption 7: } Y_{ij}^{\,k}(0) \sim f_{00}(0 \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k) \end{cases} \qquad \forall i, j : S_{ij}^k = S^{01}$$

4. For each unit being a complier at iteration k , $Y_{ij}^{\,k}\big(1, M_{ij}(0)\big)$ is imputed from the conditional distribution of $Y_{ij}(1)$ for principal stratum S^{00}, i.e. never-takers, given his values of covariates $\mathbf{C}_{ij}$:

$$Y_{ij}^{\,k}\big(1, M_{ij}(0)\big) \sim f_{00}(1 \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k) \qquad \forall i, j : S_{ij}^k = S^{01}$$

5. $NEE^{k,0}$ for compliers is computed by taking the average, within levels of co-

variates, of the difference between the two imputed potential outcomes:

$$\widehat{NEE}^{k,0}(0,1,\mathbf{c})=\frac{1}{|\mathscr{S}_c^{01}|}\sum_{i,j:S_{ij}^k=\mathscr{S}_c^{01}}\left(Y_{ij}{}^k\left(1,M_{ij}(0)\right)-Y_{ij}{}^k(0)\right)$$

Again subgroup analysis based on covariates might require some restrictions.

Estimation of individual treatment effects requires a last step: subtracting the estimated net encouragement effects from the principal causal effects for compliers:

6. $$\widehat{iTME}^{k,1}(0,1,\mathbf{c})=\widehat{PCE}^{k}(0,1,\mathbf{c})-\widehat{NEE}^{k,0}(0,1,\mathbf{c})$$

These steps, for either assumption, are resulting in draws from the posterior distribution of the causal estimands. Finally, point estimates are derived as summary statistics of these distributions, such as the mean or the median.

1.8 Application to KAHS Study

We now show in details the application of the methodology presented in the previous sessions to the KAHS study. With regards to the choice of covariates, let $\mathbf{C}_{ij}$ be a collection of baseline covariates, that a preliminary analysis has shown to be useful for predicting strata membership. In particular, these are the number of household members ($C_{1_{ij}}$), an education characteristic, being the maximum grade reached by any member of the households ($C_{2_{ij}}$), the number of bed nets per sleeping space ($C_{3_{ij}}$, labeled *household baseline coverage*), the number of sleeping spaces per household member ($C_{4_{ij}}$), and finally the proportion of members that have been sick with malaria during the year prior to the baseline survey ($C_{5_{ij}}$). We also included a neighbors' characteristic, being the average number of bed nets per sleeping space owned at baseline by all the remaining households of the cluster ($C_{6_{ij}}$, labeled *neighborhood baseline coverage*). Cluster covariates ($\mathbf{V}_j$) are not considered.

As far as priors specification is concerned, priors hyper-parameters for the fixed effects of the principal strata model are set as follows: $\boldsymbol{\mu}_{\alpha 0}^n=\boldsymbol{\mu}_{\alpha 0}^c=\mathbf{0}$, $\Lambda_{\alpha 0}^n=\Lambda_{\alpha 0}^c=$

10 I. Furthermore, we argue that random intercepts suffice to explain within cluster correlation of the compliance status, meaning that, while the overall principal strata distribution might vary across clusters, the extent to which compliance status for each household depend on baseline covariates is sufficiently constant. Accordingly, we set to zero every random slope, i.e. $\mathbf{a}_{n1j} = \mathbf{a}_{c1j} = 0$. As a consequence, random effects are assumed to follow a uni-dimensional normal distribution, $a_{n0j} \mid \sigma^2_{a_n} \sim N\left(0, \sigma^2_{a_n}\right)$ and $a_{c0j} \mid \sigma^2_{a_c} \sim N\left(0, \sigma^2_{a_c}\right)$, and the conjugate prior distribution of their variances reduces from inverse-Wishart to inverse-gamma, $\sigma^2_{a_n} \sim IG\left(\eta^n_0, s^n_0\right)$ and $\sigma^2_{a_c} \sim IG\left(\eta^c_0, s^c_0\right)$, where we set $\eta^n_0 = \eta^c_0 = 0.01$ and $s^n_0 = s^c_0 = 0.01$.

As already said, we posit a binomial distribution for the potential outcomes, with probability $p_{ij}(a, S_{ij}, \mathbf{C}_{ij})$ being a function of the principal stratum, the encouragement condition and baseline covariates, as modeled in (1.6.2). We adopt a logit link $g(\cdot)$. In the outcome model we consider a subset of $\mathbf{C}_{ij}$ given by all the covariates used in the strata model excluding the number of household members. Moreover, we are particularly interested in probing the heterogeneity of the effect of the encouragement on malaria risk between different levels of household bed net coverage at baseline. Thus, we consider only the interaction term corresponding to the variable of interest, namely $C_{3_{ij}} a$, while all the other interaction coefficients are set to zero: $\beta^{S_{ij}}_{11} = \beta^{S_{ij}}_{12} = \beta^{S_{ij}}_{14} = \beta^{S_{ij}}_{15} = \beta^{S_{ij}}_{16} = 0, \forall S_{ij} \in \{S^{00}, S^{01}, S^{11}\}$. In addition, we let the coefficients for baseline covariates to be the same across strata, with the exception for the covariate that is also present in the interaction term: $\beta^{S^{00}}_{0k} = \beta^{S^{01}}_{0k} = \beta^{S^{11}}_{0k} = \beta_{0k}$, with $k = 1, 2, 4, 5$.

As with the principal strata model, between clusters variation is taken into account by the inclusion of random intercepts, with the argument that the dependance of the outcome from covariates should not vary consistently across clusters and also that the small sample size does not enable to explore the variation of the effects between clusters. Random intercepts are also deemed constant for all principal strata. Hence, the potential outcome model is characterized by the following constraint: $\mathbf{b}_{1j} = 0, \forall j$.

Finally, the model in (1.6.2) for the probability of the binomial potential outcome, can be rewritten as follows:

$$logit\left(p_{ij}\left(a, S_{ij}, \mathbf{C}_{ij}\right)\right) = \beta_{00}^{S_{ij}} + \beta_{02}C_{2_{ij}} + \beta_{03}^{S_{ij}}C_{3_{ij}} + \beta_{04}C_{4_{ij}} + \beta_{05}C_{5_{ij}} + \beta_{06}C_{6_{ij}} + \beta_{10}^{S_{ij}}a + \beta_{11}^{S_{ij}}C_{3_{ij}}a + b_{0j}$$
$$b_{0j} \sim N(0, \sigma_b^2) \tag{1.8.1}$$

The choice for the prior distributions follows the specification outlined in section 1.7.1. Specifically, we postulate a multivariate normal prior for $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0}\right)$, with hyper-parameters $\boldsymbol{\mu}_{\beta 0} = \mathbf{0}$ and $\Lambda_{\beta 0} = 10\,\mathrm{I}$, and an inverse-gamma distribution for the variance of the random intercept, $\sigma_b^2 \sim IG\left(\eta_0^b, s_0^b\right)$, setting $\eta_0^b = 0.01$ and $s_0^b = 0.01$.

Table 1 gives some basic informations of our data and summary statistics of the baseline covariates, the intermediate variable and the outcome. As we can see, the randomization of the assignment leads to the baseline covariates being closely balanced in the two subgroups defined by assignment. The lack of perfect balance for some of them is handled by covariates adjustment. In the intervention arm, 44% of the households did not buy new bed nets; these must be never-takers and the remaining buyers households must be either always-takers or compliers. Similarly in the control arm, 41% of the households did buy new bed nets after the baseline survey; these must be always-takers and the remaining non-buyers households must be either never-takers or compliers. Accordingly, as a result of the monotonicity assumption 2, based on the set of equalities in (1.4.2) and method of moments estimators, probabilities of belonging to each principal stratum are estimated to be 0.15, 0.44 and 0.41 for compliers, never-takers and always-takers respectively. The last row in table 1 provides an ITT analysis, indicating that the encouragement intervention result in a $44.9\%(= 0.0476/0.1060 \times 100)$ reduction of the risk of contacting malaria. The between arms difference of -0.0479 in the mean proportion of malaria cases, among the households who do not buy new bed nets, suggests that the encouragement itself has a beneficial effect, regardless of the effect through the purchase of new nets. However this observed difference cannot

Table 1: *Summary statistics of the baseline covariates, the intermediate variable and the outcome.*

	Control Assignment $A=0$		Encouragement Assignment $A=1$		Difference between assignments	
Clusters	15		19		-	-
Households	161		195		-	-
Household Members C_1	5.4660	(0.1640)	6.2205	(0.2103)	0.7547	(0.2631)
Education C_2	5.7826	(0.3505)	6.4410	(0.3286)	0.6584	(0.4734)
Household Baseline Coverage C_3	0.4569	(0.0914)	0.5646	(0.0425)	0.1076	(0.0992)
Sleeping Spaces per Member C_4	0.4532	(0.1612)	0.4913	(0.0192)	0.0380	(0.0248)
Malaria Risk (Baseline) C_5	0.3533	(0.0295)	0.3195	(0.0338)	-0.0338	(0.0442)
Neighborhood Baseline Coverage C_6	0.4569	(0.0914)	0.5646	(0.0425)	0.1076	(0.0992)
Bed Net Purchase, $(P(M \mid A))$,						
$M=0$	0.5901	(0.0497)	0.4410	(0.0355)	-0.1490	(0.0600)
$M=1$	0.4099	(0.0497)	0.5590	(0.0355)	0.1490	(0.0600)
Malaria Risk (Follow-up) $(E[Y \mid A,M])$						
among bed nets non-buyers $(M=0)$	0.1213	(0.0295)	0.0734	(0.0171)	-0.0479	(0.0336)
among bed nets buyers $(M=1)$	0.0840	(0.0199)	0.0466	(0.0181)	-0.0374	(0.0228)
All	0.1060	(0.0215)	0.0584	(0.0111)	-0.0476	(0.0241)

Estimates of population means with their standard errors (in parenthesis), based on the method of moments, are reported. The second and third blocks of rows concern the intermediate variable and the outcome. Due to their bernoulli and binomial distributions, estimated means are also estimates of the probability of buying new bed nets and the probability of infection, respectively.

be interpreted causally because of the different compliance types involved in such contrast, due to the intermediate variable not being randomized. An analysis based on principal stratification could, to a certain extent, overcome this problem. In order to disentangle net encouragement effects and individual treatment mediated effects for compliers, we can argue that in the KAHS study we cannot rely on assumption (6) of stochastic homogeneity of counterfactuals. Indeed, in such a study concerning malaria, prevention behavior is difficult to predict by observed characteristics and the risk of infection from malaria depends on many different observed and unobserved factors. Therefore, we believe that, for each household,

the distribution of the potential proportion of malaria cases under the loan program assigned to the whole cluster and intervening to set M_{ij} to 0 or, in other words, if somehow we prevented the purchase of any new bed net for that household, is arguably not shared by never-takers and compliers. On the contrary, it can be more reasonable to assume homogeneity of the mean difference between counterfactuals, as stated by assumption (7), which translates into homogeneity between never-takers and compliers of the effect of the clustered encouragement when the household could not make any new purchase of bed nets (theorem 2).

Furthermore, we can hypothesize that in KAHS most of the effect of the encouragement intervention on malaria risk for always-takers can be explained by an increased number of bed nets bought under the loan program. This non-negative dissociative effect $DCE(1, \mathbf{c})$ is due to the particular choice of the binary intermediate variable that only distinguishes the purchase of zero versus at least one new new bed net at follow-up. Conversely, since no additional awareness campaign and no village interventions was provided to clusters assigned to the agricultural loan program, we can assume that for never-takers the effect of their cluster being assigned to the loan program, i.e. $DCE(0, \mathbf{c})$, is mostly due to spillovers of the purchase of new bed nets by other households belonging to the same cluster. This assumption only affects the interpretation of the estimated effects but does not alter the analysis.

1.8.1 Results

We will first focus on the characterization of principal strata. In table 2 we report posterior means and 95% intervals for the coefficients of the two latent variable models in (1.6.5), jointly used to characterize the strata membership. We can see that the only covariates that really matter in the prediction of compliance status are those related to the household baseline coverage and household living space, that is $C_{3_{ij}}$ and $C_{4_{ij}}$. In particular, there is evidence that the probability of being a never-taker increases with the number of bed nets per sleeping space (mean and 95% interval for α_{n4}: -2.227 [-2.775,-2.038]) and so does the probability of being

Table 2: *Estimated Parameters for Principal Strata Model*

	S^n *Model*		S^c *Model*	
	Mean	95% Interval	Mean	95% Interval
Household Members C_1	0.061	[-0.014, 0.087]	0.016	[-0.161, 0.073]
Education C_2	0.053	[-0.008, 0.074]	-0.090	[-0.410, 0.013]
Household Baseline Coverage C_3	-2.227	[-2.775,-2.038]	-2.289	[-4.540,-1.634]
Sleeping Spaces per Member C_4	-0.856	[-1.709,-0.568]	-0.691	[-3.270, 0.184]
Malaria Risk (Baseline) C_5	0.080	[-0.520, 0.280]	1.260	[-0.350, 1.827]
Neighborhood Baseline Coverage C_6	0.938	[0.149, 1.210]	1.032	[-1.347, 1.805]
Random Intercept Variance, σ_a^2	0.096	[0.017, 0.125]	1.588	[0.035, 1.597]

a complier (mean and 95% interval for α_{c4}: -2.289 [-4.540,-1.634]). The number of sleeping spaces per household member has a similar pattern since this covariate gives information on room sharing in the house and thus the need of bed nets per sleeping space. This result is not surprising because overall it means that households with higher coverage are less likely to buy new bed nets. Another expected result is that, holding household coverage and the other covariates fixed, neighborhood baseline coverage reduces the probability of being a never-taker, probably because of a peer influence. The role of the remaining covariates is less evident. Nonetheless, in our analysis some of these covariates have shown to be helpful in predicting compliance status in our finite study population.

For this reason and also to overcome the difficulty in the interpretation of the coefficients, due to the structure of the ordinal probit model, we have derived estimates of the sample mean of the covariates within each principal stratum, i.e., $\overline{C}_{hm_0m_1} = \sum_{i,j:S_{ij}=S^{m_0m_1}} C_{h_{ij}}/|S^{m_0m_1}|$, $\forall k \in \{1,\ldots,5\}$ and $\forall m_0, m_1 \in \{0,1\}$. Posterior distributions of the sample means are averaged over all possible vectors of $\mathbf{S}$ and $\boldsymbol{\theta}$ from their joint posterior distribution. Means and 95% intervals of these distributions are shown in table 3. Results confirm the interpretation of coefficients given above, that is, never-takers have on average higher coverage and, on the contrary, always-takers are those with a smaller number of bed nets per sleeping spaces and also a greater

number of members per room. In addition, there is evidence that compliers in our study have a greater highest grade in the family, whereas always-takers have on average a greater proportion of malaria cases in the year prior to the baseline survey. The latter result, together with the low household coverage, can explain most of the compliance behavior of the always-takers. Finally, the mean of neighborhood baseline coverage within principal strata, averaged over the remaining covariates, seems to be lower for always-takers with no evidence of a difference between never-takers and compliers.

Table 2 also reports estimates of the between-cluster variation in regard to compliance status. The estimated variance of the random intercept a_{n0j} included in the model for the conditional probability of being a never-taker versus being an always-taker or a complier (model for S_{ij}^{n}) is estimated to be 0.096 (95% quintiles: $[0.017, 0.125]$) reflecting in an intra-class correlation of 0.088. Similarly the estimated variance of the random intercept a_{c0j} of the model for the conditional probability of being an always-taker versus a complier (model for S_{ij}^{c}), conditional on not being never-takers, is estimated to be 1.588 (95% quintiles: $[0.035, 1.597]$), reflecting in an intra-class correlation of 0.614. These results can be interpreted saying that the proportion of never-takers does not differ substantially across clusters conditional on covariates, whereas the proportion of always-takers and compliers does.

The left column of Table 4 shows posterior principal strata rates, in the overall population and within three coverage categories defined by household baseline coverage:

$$\widetilde{C}_{4_{ij}} = \begin{cases} \textit{Low Coverage} & \text{if } C_{4_{ij}} \leq 0.4 \\ \textit{Medium Coverage} & \text{if } 0.4 < C_{4_{ij}} \leq 0.8 \\ \textit{High Coverage} & \text{if } C_{4_{ij}} > 0.8 \end{cases}$$

The overall probabilities of compliance status, given by the bayesian procedure, approximately match the aforementioned method of moments estimates.

A deeper characterization of principal strata is provided by the distribution of

Table 3: *Distribution of Covariates within Principal Strata*

	Compliers		Never-Takers		Always-Takers	
	Mean	95% Interval	Mean	95% Interval	Mean	95% Interval
Household Members C_1	6.409	[5.483,7.296]	5.443	[5.376,5.527]	6.192	[5.956,6.438]
Education C_2	7.212	[6.247,8.228]	5.969	[5.873,6.110]	5.972	[5.678,6.227]
Household Baseline Coverage C_3	0.483	[0.338,0.623]	0.786	[0.760,0.827]	0.218	[0.169,0.270]
Sleeping Spaces per Member C_4	0.494	[0.422,0.562]	0.502	[0.491,0.509]	0.434	[0.415,0.455]
Malaria Risk (Baseline) C_5	0.267	[0.179,0.358]	0.297	[0.279,0.307]	0.400	[0.379,0.425]
Neighborhood Baseline Coverage C_6	0.529	[0.446,0.612]	0.545	[0.525,0.564]	0.476	[0.450,0.496]

potential outcomes. The right column of Table 4 summarizes the predictive posterior distribution of potential malaria rates without encouragement, i.e. $\overline{Y}(0) = \frac{1}{N}\sum_{ij} Y_{ij}(0)$, by principal strata and by coverage categories $\widetilde{C}_{4_{ij}}$. Several important results merit attention here. First, we can see that, among never-takers, there is no evidence of a reduction of risk with an increase of coverage. This unexpected result must be due to other unmeasured factors affecting the relationship between bed nets coverage at baseline and malaria risk without encouragement, as well as compliance status. For example, never-takers with low coverage at baseline are likely to be households at lower risk, because of housing conditions, environmental factors or protective behaviors, such as the use of house spraying or windows screens. Conversely, for always-takers 95% intervals get wider as coverage augments due to the small proportion of always-takers in higher levels, hence no conclusion can be drawn on the difference between subgroups defined by coverage. For compliers posterior means seem to decrease with $\widetilde{C}_{4_{ij}}$, but still intervals makes this pattern consistent to random fluctuation. The second and more important point concerns a comparison between principal strata. At all coverage levels, compliers are those households who would have a considerably higher risk of malaria infection if not encouraged to buy new nets with loans, with an overall mean risk of 32.1% against 6% for never-taker and 8.4% for always-takers. This result can be somewhat surprising, but we can give some intuitive explanations. For never-takers, the low risk of contracting malaria

compared to the other principal strata might be due to better housing conditions, as well as a greater use, at least in 2010, of other preventive measures such as windows screens and preventive behaviors such as keeping doors and windows closed at night, being indoors after sunset or removing possible breeding sites in the house. For always-takers, we should remember that this is the sub-population that would buy new bed nets even without the encouragement, therefore their actual household coverage at follow-up has increased compared to the one at baseline. This can be one of the reasons for their low risk, probably together with the take-up of similar preventive behaviors to the ones used by never-takers. On the contrary, compliers seem to be the sub-population most at risk of malaria, at all levels of baseline coverage, when not encouraged and hence the number of ben nets owned does not increase. The reason can be, besides the use of less preventive measures and more risky behaviors, the presence of higher risk factors, such as livestock animals, co-morbidities, pregnancies, house damage, as well as presumably, for those with medium high coverage, old bed nets in bad physical integrity which makes them ineffective and no longer impregnated with insecticides.

In any case, the different mean potential outcome under control encouragement between principal strata supports our hypothesis of assumption 6 of partial stochastic homogeneity of counterfactuals being implausible.

Table 5 concerns the estimated effects defined in section 1.4.1, that is principal causal effects PCE, net encouragement effects NEE^0 and individual treatment effects iTME^1, by principal strata and by coverage levelS $\widetilde{C}_{4_{ij}}$. Estimates are based on imputations from the predictive posterior distributions of potential outcomes, as outlined in section 1.7.2. Results are based on 45000 iterations, combining three chains, each run for 25000 iterations, with a burn-in of 10000 iterations. To check for convergence, for each effect we computed the potential scale reduction factor (Gelman and Rubin, 1992), giving a maximum value of 1.04, suggesting no evidence against convergence.

Table 4: *Principal Strata Rates and Malaria Rates by principal strata*

	Principal Strata Rates $P(S_{ij} = S^{m_0 m_1})$		Malaria Rates $\overline{Y}(0)$	
Principal Strata	Mean (SD)	95% Interval	Mean (SD)	95% Interval
NEVER-TAKERS				
Low Coverage	0.198 (0.023)	[0.147,0.231]	0.042 (0.019)	[0.006,0.082]
Medium Coverage	0.520 (0.024)	[0.459,0.561]	0.051 (0.022)	[0.013,0.103]
High Coverage	0.788 (0.016)	[0.745,0.804]	0.074 (0.036)	[0.025,0.167]
All	0.456 (0.017)	[0.418,0.483]	0.060 (0.026)	[0.020,0.123]
ALWAYS-TAKERS				
Low Coverage	0.671 (0.031)	[0.603,0.724]	0.084 (0.011)	[0.065,0.109]
Medium Coverage	0.283 (0.044)	[0.204,0.367]	0.092 (0.053)	[0.029,0.221]
High Coverage	0.094 (0.031)	[0.039,0.157]	0.063 (0.075)	[0.000,0.275]
All	0.399 (0.027)	[0.345,0.452]	0.084 (0.020)	[0.057,0.135]
COMPLIERS				
Low Coverage	0.130 (0.045)	[0.051,0.224]	0.348 (0.110)	[0.177,0.574]
Medium Coverage	0.197 (0.058)	[0.082,0.306]	0.331 (0.117)	[0.165,0.592]
High Coverage	0.118 (0.039)	[0.039,0.186]	0.290 (0.131)	[0.123,0.633]
All	0.145 (0.038)	[0.073,0.219]	0.321 (0.099)	[0.179,0.545]

Reported results are means, standard deviations and 95% intervals of the posterior distribution of strata membership rates, and the posterior predictive distribution of malaria rates by principal strata under encouragement status $A_j = 0$, i.e. $\overline{Y}(0)$. Both distributions are averaged over $\mathbf{C}_{ij}$(or just over C_{1ij}, C_{2ij}, C_{4ij} and C_{5ij} when results are presented within household baseline coverage categories $\widetilde{C}_{4ij}$), the clusters and $\boldsymbol{\theta}$.

Consider principal causal effects, presented in the last block of columns. The estimated PCE for compliers is a reduction of malaria risk of 17.2% (posterior mean), with similar estimates at every level of household baseline coverage. As expected, this total effect, being the sum of $NEE^0(0,1)$ and $iTME^1(0,1)$, is much larger than PCEs in the other principal strata. The estimated PCE for always-takers, i.e., $DCE(1)$ is a reduction of the risk of infection of 6.2% (posterior mean). 95% intervals provide a strong evidence of a beneficial effect of the encouragement for both compliers and always-takers. These effects are slightly less pronounced if we look at posterior medians.

For never-takers, instead, we find a negligible effect of the encouragement, i.e., $DCE(0,\mathbf{c})$, for all levels of coverage. Proportion of malaria cases at baseline and

Table 5: *Estimated Effects within Principal Strata and by coverage levels*

	NEE^0				iTME^1				PCE			
Principal Strata	Mean	Median	(SD)	95% Interval	Mean	Median	(SD)	95% Interval	Mean	Median	(SD)	95% Interval
NEVER-TAKERS			DCE(0)									
Low Coverage	0.025	0.023	(0.039)	[-0.043, 0.114]		–			0.025	0.023	(0.039)	[-0.043, 0.114]
Medium Coverage	0.011	0.014	(0.034)	[-0.058, 0.079]		–			0.011	0.014	(0.034)	[-0.058, 0.079]
High Coverage	0.010	0.017	(0.053)	[-0.106, 0.118]		–			0.010	0.017	(0.053)	[-0.106, 0.118]
All	0.014	0.018	(0.041)	[-0.072, 0.097]		–			0.014	0.018	(0.041)	[-0.072, 0.097]
ALWAYS-TAKERS			DCE(1)									
Low Coverage	-0.062	-0.064	(0.017)	[-0.095,-0.026]		–			-0.062	-0.064	(0.017)	[-0.095,-0.026]
Medium Coverage	-0.066	-0.061	(0.048)	[-0.186,-0.002]		–			-0.066	-0.061	(0.048)	[-0.186,-0.002]
High Coverage	-0.050	-0.034	(0.073)	[-0.258, 0.033]		–			-0.050	-0.034	(0.073)	[-0.258, 0.033]
All	-0.062	-0.062	(0.023)	[-0.118,-0.023]		–			-0.062	-0.062	(0.023)	[-0.118,-0.023]
COMPLIERS						CACE^1						
Low Coverage	0.014	0.014	(0.044)	[-0.073, 0.104]	-0.208	-0.200	(0.132)	[-0.470,0.027]	-0.194	-0.183	(0.128)	[-0.448, 0.029]
Medium Coverage	0.015	0.019	(0.050)	[-0.091, 0.110]	-0.170	-0.157	(0.144)	[-0.473, 0.078]	-0.155	-0.138	(0.141)	[-0.456, 0.079]
High Coverage	0.015	0.022	(0.064)	[-0.125, 0.138]	-0.191	-0.175	(0.157)	[-0.553, 0.071]	-0.176	-0.152	(0.147)	[-0.530, 0.047]
All	0.014	0.018	(0.041)	[-0.072, 0.091]	-0.186	-0.178	(0.125)	[-0.452,0.030]	-0.172	-0.159	(0.123)	[-0.435, 0.032]
ALL											ITT	
Low Coverage	-0.028	-0.026	(0.016)	[-0.062,-0.006]	-0.021	-0.019	(0.015)	[-0.058, 0.002]	-0.050	-0.049	(0.023)	[-0.097,-0.014]
Medium Coverage	-0.006	-0.004	(0.024)	[-0.059, 0.043]	-0.023	-0.017	(0.024)	[-0.082, 0.011]	-0.030	-0.022	(0.030)	[-0.100, 0.019]
High Coverage	0.007	-0.005	(0.041)	[-0.078, 0.097]	-0.016	-0.012	(0.016)	[-0.055, 0.005]	-0.009	-0.007	(0.038)	[-0.092, 0.070]
All	-0.016	-0.014	(0.026)	[-0.072, 0.034]	-0.026	-0.024	(0.018)	[-0.068, 0.004]	-0.042	-0.042	(0.027)	[-0.100, 0.008]

Means, medians, standard deviations and 95% intervals of the posterior distribution of net encouragement effects NEE^0, individual treatment mediated effect iTME^1 and principal causal effects, are presented by principal strata and household baseline coverage categories $\widetilde{C}_{4ij}$. The last block of rows concerns the estimated effect in the whole population.

potential proportion under control encouragement have not suggested lack of knowledge and awareness of malaria for this subpopulation, but, on the contrary, probably never-takers are the most aware of preventive measures or in general the less at risk at least in 2010, regardless of the encouragement conditions. As said earlier, we can argue that for this principal stratum there is little effect of the encouragement itself, such as an increase in the usage of old bed nets or the undertaking of other measures. Thereby this result suggests no evidence of spillover effects for never-takers, at any coverage level.

The overall ITT, given by the average of the three principal causal effects, is estimated as a decrease in the risk of malaria of 4.2% (95% interval: $[-10\%, 0.8\%]$), which approximates the ITT estimated from the observed data. Note that 95%

posterior intervals at medium and high coverage are wider and include zero making the results consistent with random fluctuation. This is due to the high proportion of never-takers in these categories.

When it comes to disentangling the effects for compliers, $iTME^1(0,1)$ is estimated by the posterior mean as a reduction of 18.6% (95% interval: $[-45.2\%, 3.0\%]$) whereas $NEE^0(0,1)$ as a minimal increase with high uncertainty (posterior mean: 1.4%; 95% interval: $[-7.2\%, 9.1\%]$). The individual treatment effect for compliers is equivalent to the average effect of the purchase of at least one bed net, i.e. $CACE^1$.

Average net encouragement effects in the whole population, computed taking the average of the beneficial dissociative causal effects of always-takers, the negligible dissociative causal effects of never-takers and the negligible net encouragement effects of compliers, are beneficial with strong evidence only within the low coverage category with a posterior mean of -2.8%. Finally, by multiplying $iTME^1(0,1)$ by the proportion of complier, we obtain an estimate of the individual treatment effect in the population given by -2.6% (95% interval: $[-6.8\%, 0.4\%]$).

1.9 Discussion

In this chapter we provide a framework based on the principal stratification approach to investigate the different mechanisms elicited in cluster encouragement designs, through the individual treatment uptake or through other pathways, including spillover effects. We define net encouragement effects and individual treatment mediated effects within principal strata, with the latter only present among compliers when monotonicity of compliance is assumed. The core of this work concerns the proposal of homogeneity assumptions allowing to disentangle the two different effects for this subpopulation, under violation of sequential ignorability.

Principal causal effects themselves provide us with useful information on how encouragement has an impact on the outcome, within different subpopulations types defined by compliance behavior. Our analysis of the KAHS study, gives evidence

that for those households who would buy new bed nets only if agricultural loans were offered, the compliers, the offer of loans to 11 farmers living in their villages and the surroundings helps reducing the risk of contracting malaria. It also suggests that those who would purchase bed nets anyway, the always-takers, benefit from the loan program, most likely through an increase in the number of bed nets purchased due to the subsidized prize. On the contrary, it shows nonsignificant effect for never-takers, that is for those who would not buy new bed nets regardless of the encouragement. Consequently there is no evidence of spillover effects from the increased number of bed nets in the cluster, due to the encouragement, at least for this subpopulation. The slightly detrimental effect for this subpopulation, especially with low coverage, even if intervals are too wide to draw definite conclusions, suggests the importance to investigate spillover effects in large scale programs.

Furthermore, the analysis of compliance status provided by the principal stratification framework, compared with simple ITT analysis, gives insight into the extent to which encouragement enhances the treatment uptake, how different types of the population react to the encouragement and what are the characteristics of individuals that encouragement is able to reach. KAHS program evaluation has provided an interesting case study in which principal strata differ substantially by their potential risk under control intervention. Specifically, compliers would have much higher risk of infection. In any case, this analysis shows how the loan program was able to reach the subpopulation most at risk and more in need to be prompted to take on better prevention measures.

This characterization of principal strata can also help us understand whether and which homogeneity assumption is more plausible to untie the mediated and non-mediated effects among compliers, the one concerning the distribution of counterfactuals or the one involving their mean difference. Besides the availability of two possible forms of homogeneity, a further advantage of our formalization of identifying homogeneity assumptions is the flexibility of specification. In fact, although

we have focused on a particular case that is suitable for the application study, in the appendix A 1 we provide more general homogeneity assumptions represented in terms of two parameters whose specification leads to a particular assumption involving two different principal strata, always-takers or never-takers on the one hand and compliers or defiers on the other hand. Each specific assumption enables the identification of a combination of the two effects $NEE^{\tilde{a}}$ and $iTME^{1-\tilde{a}}$, with $\tilde{a}=0$ or $\tilde{a}=1$, for three principal strata, namely always-takers, never-takers and compliers or defiers. Therefore, this general formalization would allow to assess the type of effects that can be estimated according to the particular assumption that is deemed plausible in the specific setting and, on the other hand, it gives insight into the assumptions that would be required for the identification of the effects of interest. The choice about which particular homogeneity assumption holds has to be determined on a case-by-case basis, with the help of subject matter knowledge and comparison of principal strata in terms of covariates and potential outcomes. In our application we rely on homogeneity of the net encouragement treatment effect between never-takers and compliers, conditional on covariates. Based on this assumption and the application of the imputation approach for the estimation of causal mechanisms, there is no evidence of a net encouragement effect among compliers, at any coverage level. Therefore all the effect of the encouragement for this principal stratum would be through the purchase of new bed nets, resulting in a quite high beneficial effect of this treatment on the risk of malaria, when the encouragement is assigned in the same cluster ($iTME(0,1) = CACE^1$). This conclusion is important in that is shows how the solely purchase of few bed nets in a household at high risk can make a real difference. Hopefully, if the loan program were offered to more farmers in each cluster, an increased coverage in the community would exponentiate exponentially reduce malaria through beneficial spillovers. This study does not allow us to assess this hypothesis, arguably because of the small number of beneficiaries.

Final results suggest that the impact of the encouragement is mostly driven by

enhancing the purchase of bed nets in that 15% of population that otherwise would have a high risk of infection and would not prioritize prevention, the compliers, for whom the effect of new bed nets is high at every level of baseline coverage, but almost as much is given by the effect due to subsidized price through the increased number of new bed nets among those households who would carry out a new purchase anyway, i.e., the alway-takers, who constitute 41% of the population.

Since a negligible effect was found among never-takers, if resources were limited, baseline information were already available and the offer of the loan program had a cost itself even if subsidies were not used (e.g. mail service, door-to-door visits...), we may want to exclude this subpopulation from the encouragement program. The lack of knowledge of strata membership would force to exclude those units with higher probability of being never-takers. In KAHS study these are mostly those with higher coverage. Prediction errors in compliance status would result in a less beneficial effect in the population, provided that for compliers the effect of encouragement on risk reduction is estimated to be high, even at high coverage level. Maximization of a risk reduction could be computed for beneficiaries selection. Collection of more baseline features to better predict principal strata would ease this task.

Despite the expanded coverage under encouragement, the analysis shows that the risk of malaria would still be quite high for the subpopulation of compliers (with a posterior mean for $\overline{Y}(1)$ of 0.149 as can be seen from results reported in tables 4 and 5, that is $\overline{Y}(0) - PCE(0,1) = 0.321 - 0.172$). ITNs work by preventing indoor night biting and by killing mosquitos, but do not protect from risk behaviors. This alarming result suggests that interventions should be complemented with other encouragement interventions to promote alternative vector control measures, such as environmental management, windows screens or training to prevent high risk practices. In addition, the observed heterogeneity in malaria risk highlights the need, in the design phase, of a detailed characterization of behavioral, socio-economic and environmental risk factors of the target population in order to select appropriate

suites of interventions.

Our analysis of KAHS has several limitations. First, the choice of a binary intermediate variable, although it allows to shed light on a well-defined principal stratification of the population, it does not use information on actual number of bed nets. A continuous intermediate variable could also be handled in the principal stratification framework (Jin and Rubin 2008; Bartolucci and Grilli 2011; Schwartz et al. 2011), and homogeneity assumptions could be defined accordingly. Second, homogeneity of spillover effects that these assumptions imply can be problematic. When net encouragement effects incorporate spillover effects by intermediate variables, the validity of assumption (7) should raise more concerns and carefulness. Indeed, even if we believe that an increase in the intermediate variables of one neighbor affects a unit's outcome in the same way for compliers and never-takers, assumption (7) is not sustainable when the average compliance behavior in the vicinity of each unit differs between the two principal strata. A clustering of principal strata, due to mechanisms such as homophily or peer influence in the compliance behavior, would make spillover effects of the clustered encouragement intervention differ across principal strata. In our example, the estimated intra-class correlation of compliance status suggests a slight difference of the proportion of always-takers and compliers across clusters, even after conditioning on covariates. However we can assume that an increased number of bed nets in the remaining households of the clusters, due to the encouragement, is on average the same for compliers and never-takers, making assumption (7) more sustainable. Future works could focus on the estimation of spillover effects accounting for a differential distribution of potential values of the intermediate variable in the neighborhood. Another potential threat to homogeneity of spillover effects is a possible differential use of preventive measures other than bed nets in the matching strata. If this is the case, then we expect compliers to be more affected by an increased neighborhood coverage, either positively or negatively. Third, it is also possible that the offer of loans to compliers has encouraged them to

take other preventive measures together with the usage of the new bed nets. In this case the individual treatment mediated effect for compliers would be overestimated. Nonetheless, it is also possible that these farmers felt already satisfied with their new bed nets, also because the loan program in theory should not be associated with any prevention campaign. Anyhow, in this article we have emphasized the arguments that can be made in favor or against homogeneity assumptions in a challenging application with possible spillover effects and the presence of important latent features that make the distribution of potential outcomes differ substantially across principal strata. In many applications, the validity of homogeneity assumptions can be much less controversial.

Chapter 2

Disentangling Spillover Effects using Neighborhood Principal Strata

2.1 Introduction

In the previous chapter we focused on individual principal strata defined by the potential individual treatment uptake and we said that one of the two effects into which principal causal effects were decomposed, namely the net encouragement effect, can be interpreted as including the effect of the encouragement through a change in the neighbors' treatment. In this chapter we will now drill down into the details of this type of effect, by giving formal definition and introducing a novel approach that, relying on a new set of assumptions, will allow us to isolate such an effect.

When outcomes involve or depend on behavioral changes or transmittable objects such as money, information material, parasites or virus, mechanisms of psychological influence or physical transmission are likely to take place between individuals who interact with one another or simply share the same environment in the daily life. In these circumstances, the implementation of an intervention can give rise to interdependent outcomes, that is, one's outcome is affected by the treatment received by other subjects. In general this interference mechanism can be found in many applications in different fields from behavioral economics and education to psychology, social science and infectious diseases. Depending on the field of study the effect of other subjects' treatment is referred to as *spillover effect* or *peer effect*.

In this chapter we will focus on the same setting of clustered encouragement designs described in chapter 1 and we will assume a cluster-level SUTVA as defined in 1. In this scenario, one subject's outcome does not depend on the intervention assigned to other clusters but can be affected by the treatment received by other subjects belonging to the same cluster. This effect of the encouragement assigned to the neighborhood might pass through several behavioral changes of the neighbors, including the uptake of the treatment of interest or other unmeasured characteristics.

We can give several examples of clustered encouragement designs where interference mechanisms are likely to take place.

A classical example of interference comes from the field of infectious diseases. In areas where the low vaccine coverage is mainly due to the lack of local immunization services, mobile immunization camps can be used to increase vaccination rates. In a clustered randomized trial, mobile immunization camps are set up in randomly selected villages (Banarjee et al., 2010). The presence of the camp has an overall protective effect on all the inhabitants of the village. In fact not only it directly prevents those receiving vaccines from being infected but it also protects both vaccinated and unvaccinated people by reducing the number of infected from whom to contract the disease (contagion effect) and also by an alteration of the contagion mechanism (infectiousness effect). If the presence of an immunization camp is associated with an information campaign, in the villages randomized to receive the intervention, behavioral changes essential to reduce the risk of infection can also be achieved as an effect of the additional information provided, regardless of the vaccine receipt.

Another example of the same kind regards interventions to improve condom use and hence reduce the risk of STD infection. The simplest intervention is a free distribution of condoms. It has been claimed that differential distribution in a same community can negatively affect condom purchase on those who don't receive them for free, either for psychological or market reasons. For this and other practical

reasons many studies have assigned free distribution of condoms at village level. The receipt of condoms can be thought as an encouragement and the actual use of them as the individual treatment. In villages assigned to the intervention, an individual who doesn't use condoms even if given for free is less likely to contract the disease thanks to the increase condom use of the subjects belonging to his social network.

Sometimes such spillover effects can be desirable because they reinforce the effect of the intervention. Other times an increase in the treatment uptake in the neighborhood can be detrimental for a certain type subject and would reduce the effect of the clustered encouragement intervention. A quantification of these mechanisms would allow to optimize the design of the clustered intervention in a scale-up phase in order to achieve better results with less resources, by tailoring and targeting the encouragement both at individual and cluster level.

At the analysis stage, spillover effects can be a nuisance or the major effects of interest. In both cases valid statistical inference have to take into account this interdependence. In clustered encouragement schemes all individuals belonging to the same cluster share the same encouragement assignment. However, since the actual treatment received is not randomized but rather self-selected, in the same cluster different individuals end up either being exposed to the active treatment or unexposed. It is precisely this within-cluster variability of the actual treatment received that enables to investigate the relationship between one subject's outcome and the treatment received by other individuals of the same cluster. Nevertheless, passing from an assessment of association to causal conclusions would require to make explicit assumptions. This type of interference mechanism or spillover effect, together with its identifying conditions, will be the focus of this chapter.

In the past decade extensive effort has been made to give proper definitions and partitions of the effect of a treatment in the presence of interference in the framework of causal inference (Sobel, 2006; Hong & Raudenbush, 2006; Hudgens & Halloran,

2006; Rosenbaum, 2007; Manski, 2013). Tchetgen Tchetgen & VanderWeele (2011, 2012) have proposed a decomposition of spillover effects arising in vaccine trials into *contagion effect* and *infectiousness effect* and provided bounds and identification results, while Hudgens & Halloran (2006, 2008, 2012) have discussed the problem of developing causal methods for estimating these effects. At the same time, researchers have begun addressing their investigation through experimental methods and proper designs. Non-parametric identification of causal effects in the presence of interference of the treatment can be achieved with a two-stage randomized design (Duflo & Saez, 2002; Giné & Mansuri, 2011; Sinclair et al., 2012): at the first stage specific clusters, defined as to be independent, are randomized to having a certain proportion of treated individuals; subsequently at the second stage, once the proportion is determined, within each cluster individuals are randomly assigned to one of the two treatment conditions. In this work we focus on a different setting where there is reason to believe that the actual treatment received by a subject also affects the outcome of other subjects interacting with him, but the treatment cannot be randomized and groups of subjects are randomly assigned to receive or not receive an encouragement intervention.

In a previous work, VanderWeele et al. (2013) defined three different effect arising in a group randomized study when interference is given by an intermediate variable: The methodological developments that was proposed to disentangle the three effects is accomplished accomodating the framework of mediation analysis to spillover effects and it is shown that an extended version of the *sequential ignorability* assumptions yields non-parametrical identification. In an encouragement design such assumptions are generally questioned because of the self-selection of the treatment.

In the present study we draw on a similar decomposition of the effect of the clustered encouragement: an effect mediated by the individual treatment (*Individual Mediated Effect*), a spillover effect mediated by the treatment received by neighbors within the same cluster (*Spillover Mediated Effect*) and an effect of the encourage-

ment due to factors other than a change in the distribution of the treatment received in the cluster (*Pure Encouragement Effect*). Each of these effects answers different interesting questions and gives insight into a different but complimentary set of underlying consequences of the encouragement intervention. The way evidence can be incorporated in decision-making depends on the specific setting. Though, in order to get the best from the analysis and turn the results into policy decisions, it would be crucial to investigate the heterogeneity of these effects in different type of subjects. As a matter of fact, identification issues are circumvented here by the use of bayesian estimation applied to a novel principal stratification of the population, based on both individual and neighborhood potential behavior in terms of the treatment uptake under both encouragement conditions. The use of the individual principal stratification coupled with the set of identifying assumptions, as presented in the previous chapter, enables to untie in the entire population the individual mediated treatment effect from all the other factors that are involved in the impact of the encouragement intervention. The addition of a neighborhood stratification along with suitable assumptions will allow us to further isolate the spillover mediated effect.

The chapter starts with section 2.2 by introducing a second illustrative example, taken from the vaccination field, that will help explaining the concepts discussed in this part. The notation used in the previous chapter is applied to this example and other variables are also introduced to account for the neighborhood treatment status. The novel principal stratification approach, based on both the individual and the neighborhood compliance behavior, is presented in section 2.3. In section 2.4, we define the new causal mechanisms of interest within each principal stratum. The problem of identification is discussed in section 2.5, where we provide additional homogeneity assumptions, followed by identification results. The modeling details are presented in section 2.6, while section 2.7 is devoted to the Bayesian inference with a new imputation algorithm for the estimation of the defined causal mecha-

nisms. The frequentist performance of the bayesian estimation procedure is tested by means of a simulation study, developed in section 2.8. Section 2.9 concludes the thesis with a brief summary and discusses future research directions.

2.2 An illustrative Example: Notations and Definitions

In order to illustrate our approach aiming at disentangling the effect of spillover by the neighbors' treatment uptake, we will refer to a hypothetical study example concerning vaccine trials, adapted from Banarjee et al. (2010). Immunization is one of the most successful and cost-effective interventions in the past century, preventing a series of major illness affecting children. However, in developing countries coverage rates wane, vaccines continue to be underused and undervalued and vaccine-preventable diseases remain a threat to world health, killing two to three million people every year. In India, immunization services are offered free in public health facilities, but, despite many decades of efforts to immunize children against these diseases, only 44% of children aged 1-2 years have received the basic package of immunization (as defined by WHO and Unicef) and that drops to 1-2% in rural areas. Most common reasons for non vaccinations are the lack of local facilities, poor supply of vaccines in the region and unreliability of health workers but also unawareness of the need of vaccine, fear of side effects, mistrust, misconception regarding the effect of vaccines and family members busy or ill. Religion, gender and socio-economic status can be determinants of coverage inequalities. Mobile immunization camps can help increasing coverage in rural areas thanks to a better reliability of immunization services. Banarjee et al. (2010) provided randomly selected poor villages of rural Rajastan with regular monthly, well-publicized immunization camps, which offered to all children aged 0-3 years the basic package. The aim of the study was to assess the effectiveness of the intervention in terms of an increase in vaccination coverage. Suppose now that researchers also wanted to evaluate the final impact on diseases reduction three years after the the onset

of the intervention so that endpoint measures are now available. As an example, we will focus on the effect of immunization camps on tuberculosis (or TB), which remains one of the major health problems in India accounting for one million new cases every year. It is also the largest killer from a single major pathogen in adult life. Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis, which mostly affects the lungs but it can also damage other parts of the body. TB spreads from person to person through the air when a person with TB of the lungs or throat coughs, sneezes, or talk. People infected with TB bacteria can either keep them in inactive (latent) form or develop the disease, with a probability depending on his immune system and other risk factors. People with latent TB have a 10% lifetime risk that symptoms will develop later into an active infection. A person with active but untreated tuberculosis may infect 10-15 (or more) other people per year,whereas those with latent infection are not contagious. Tuberculosis is closely linked to both overcrowding and malnutrition, making it one of the principal diseases of poverty. Other diseases can also increase the risk of developing TB. These include HIV, chronic lung diseases, alcoholism and diabetes mellitus. The vaccine against tuberculosis is called BCG. It activates specific antibodies, preparing the subject to be ready to fight bacteria. In this way the vaccine does not prevent someone being infected, but it prevents the development of the disease reducing the bacterial load. It is specifically designed to defend children against TB and it has been shown that it protects them for about 15 years. There is reason to suspect the presence of interference mechanism by the vaccine received by other subjects. This might occur precisely because if the vaccinated people are less likely to develop the disease whenever infected, they are also less prone to transmit TB to others.

In this chapter we will use the same notation introduced in chapter 1. We will now translate it to the illustrative example.

We consider as units of analysis all unvaccinated and healthy children aged 0-18 months at baseline and as units of randomization villages assigned to either the

active encouragement group (A_j=1), where mobile immunization camps are set up, or to the control group (A_j=0), where children are left to the unreliable municipal immunization services. M_{ij} is an indicator of the vaccine receipt, being $M_{ij} = 1$ if child i in village j gets vaccinated against TB between baseline and follow-up (18 months later) surveys and $M_{ij} = 0$ otherwise. As for the outcome, let Y_{ij} be the bacterial load (log10 CFU) found in a sputum specimen of child ij in a laboratory test performed 3 years after the onset of the program.

In order to incorporate spillover effects in the analysis let $\mathcal{N}_{ij}$ be the neighborhood of unit ij, $\mathcal{N}_{ij} = [1j, \ldots, i-1j, i+1j, \ldots, N_j j]$, that is all units in cluster j excluding unit ij. Let $\mathbf{M}_{-ij}$ be the vector of vaccine receipt indicators of all the units in $\mathcal{N}_{ij}$, $\mathbf{M}_{-ij} = [M_{1j}, M_{2j}, ..., M_{i-1j}, M_{i+1j}, ..., M_{n_j j}]$, and let N_{ij} denote a scalar summarizing this vector and taking values [0,1]:

$$N_{ij} = G_{ij}(\mathbf{M}_{-ij})$$

where $G_{ij}(\cdot)$ is a linear functional $G_{ij} : \{0,1\}^{N_j - 1} \rightarrow [0,1]$.

For instance N_{ij} can be a weighted proportion of units under treatment in the entire cluster j or in a smaller vicinity of the unit i, $G_{ij}(\mathbf{M}_{-\mathbf{ij}}) = \sum_{k \in \mathcal{N}_{ij}} \frac{w_{ik} M_{kj}}{N_j - 1}$, with weights w_{ik} such that $w_{ik} \in [0,1]$ and $\sum_{k \in \mathcal{N}_i} w_{ik} = 1$. For instance, weights can depend on the distance between unit i and unit k in cluster j or on other characteristics. For this reason we can call this variable *neighbors' treatment* or *neighbors' vaccine receipt* in this specific example.

We now turn to the primitive potential outcomes. Banarjee et al. (2010) claim the absence of any contamination between villages from all encouragement groups being sufficiently far from each other (over 20 km). Formally this translates into cluster-level SUTVA assumption in 1. Under this assumption we defined the potential outcome $Y_{ij}(A_j, \mathbf{M}_j) \equiv Y_{ij}(A_j, M_{ij}, \mathbf{M}_{-ij})$, being in this example the potential presence of active bacteria in child ij under encouragement condition A_j and vaccine status $\mathbf{M}_j = [M_{ij}, \mathbf{M}_{-ij}]$ in village j. The latter expression of the potential outcome high-

lights the possibility of conceiving two different type of hypothetical intervention on vaccine receipt, one for child ij and the other for his neighbors. We will also assume that Y_{ij} depends on the vaccination received by the neighbors only through the summarizing variable N_{ij}, so that, in all those cases where $N_{ij} = G_{ij}\left(\mathbf{M}_{-ij}\right)$, $Y_{ij}(A_j, \mathbf{M}_j)$ will be mapped into $Y_{ij}(A_j, M_{ij}, N_{ij})$. Hereafter we will use the following notation: $Y_{ij}(a, m, n) \equiv Y_{ij}(A_j = a, M_{ij} = m, N_{ij} = n)$.

Let us turn our attention to the hypothetical interventions on the intermediate variables, M_{ij} and N_{ij}. Let $M_{ij}(a)$ denote the potential bacterial load that child ij would have experienced if the village j he belongs to were assigned to encouragement condition $A_j = a$. Similarly let $N_{ij}(a) = G_{ij}(\mathbf{M}_{-ij}(a))$ be the potential neighbor's vaccine receipt that units living next to unit ij would have received if the cluster-level encouragement intervention, A_j, were set to a. A particular intervention on the intermediate variables would set $\mathbf{M}_j = [M_{ij}, \mathbf{M}_{-ij}] = [M_{ij}((\tilde{a}), \mathbf{M}_{-ij}(a')]$, with $\tilde{a}, a' \in \{0, 1\}$. This kind of joint intervention leads to a particular potential outcome of the form $Y_{ij}\left(a, M_{ij}(\tilde{a}), N_{ij}(a')\right)$, which denotes the potential bacterial load that child i in cluster j would have experienced if A_j, were set to a, the indicator of his vaccine receipt, M_{ij}, were set to the value it would have taken under encouragement condition $A_j = \tilde{a}$ and the neighbors' vaccine receipt were set to the value it would have taken under $A_j = a'$. Counterfactuals of this type are in general not observed. Indeed the only one that can be found in the data for each unit is one of the two potential outcomes of the form $Y_{ij}\left(A_j, M_{ij}(A_j), N_{ij}(A_j)\right)$, where A_j is the encouragement status assigned to cluster j. This coincides with the previously used notation $Y_{ij}(A_j)$. We will define the effects of interest based on counterfactuals of this type and we will see later on in the chapter how the missing data problem can be solved.

In what follows we will maintain the assumptions of unconfoundedness of the encouragement assignment, which is supported by the randomized experiment.

Assumption 8. Unconfoundedness of the encouragement assignment

$$Y_{ij}(0), Y_{ij}(1), M_{ij}(0), M_{ij}(1), N_{ij}(0), N_{ij}(1) \perp\!\!\!\perp A_j \mid \mathbf{C}_{ij} = \mathbf{c} \quad \forall i, j$$

2.3 Principal Stratification Approach

In order to investigate the different mechanisms that arise in clustered encouragement designs, involving not only the individual treatment uptake but also the neighbors' treatment, we will extend the principal stratification approach introduced in the previous chapter. A basic principal stratification with respect to a post-assignment variable is a partition of the population into sets in which all units share potential values of the post-assignment variable under both assignment conditions. We will define two basic principal stratifications: one with respect to the individual treatment received M_{ij} and the other with respect to the neighbors' treatment N_{ij}. The former, which we will refer to as *individual principal stratification*, has already been outlined in chapter 1, as it was first introduced by Frangakis & Rubin (2002). Conversely, the latter, hereinafter referred to as *neighborhood principal stratification*, is new in the literature and we will see how it is useful for the evaluation of spillover effects.

Individual Principal Strata

The individual principal stratification is a partition of units in subpopulations, the so-called *individual principal strata*, defined according to the potential individual vaccine status under both encouragement conditions:

$$S^{m_0 m_1} := \{ij : M_{ij}(0) = m_0, M_{ij}(1) = m_1\} \tag{2.3.1}$$

S_{ij} be the indicator of the individual principal stratum to which subject ij belongs. We will maintain here the assumption of monotonicity (2), so that there will be only three principal strata $S_{ij} \in \{S^{00}, S^{01}, S^{11}\}$: *never-takers*, *compliers* and *always-takers*, respectively.

In the BCG vaccine example, never-takers are those children who would not get

vaccinated against TB during the 18 months after the onset of the study, regardless of the immunization service, compliers are those who would get vaccinated only with the mobile immunization camp and finally always-takers are those children who, during that period, would take a BCG vaccine, if not already received, either in local facilities or in the program mobile camps.

We can formulate some hypotheses on the characteristics of these three different subpopulations. Always-takers might live closer to the public health centers, belong to more educated families, be more aware of the importance of vaccination, or be in contact with a large number of contagious TB adults making their relatives more convinced to take the children to get vaccinated. On the contrary, never-takers and compliers might be those children who live at a greater distance from public health facilities, belong to families whose knowledge about vaccination is low and/or are in contact with fewer TB adults. However, compliers do make use of the mobile immunization camps and never-takers do not. This behavior might be explained by, for instance, relatives of never-takers being busy or ill, having a greater fear of side effects or misconception of the effects of vaccines. These subtle differences between never-takers and compliers are likely not to affect the disease outcome or at least the effect of the intervention, thus, as we will see later, either aforementioned explanation will support our homogeneity assumptions.

Neighborhood Principal Strata

We propose here an innovative approach for the investigation of spillover effects based on a neighborhood principal stratification. This partition of the population under study is based on the neighbors' treatment variable N_{ij}. Neighborhood principal strata are sets of units with the same potential values of the neighbors' treatment uptake under both encouragement conditions:

$$nS^{n_0 n_1} := \{ij : N_{ij}(0) = n_0, N_{ij}(1) = n_1\} \qquad n_0, n_1 \in [0,1] \tag{2.3.2}$$

Let nS_{ij} be the indicator of the neighborhood principal stratum. The number of all possible principal strata $nS^{n_0 n_1}$ depends on the function $G_{ij}(\cdot)$. If N_{ij} is simply the proportion of subject in cluster j taking the treatment, excluding unit ij, then for each cluster j there will be N_j^2 neighborhood principal strata. When the assumption of monotonicity holds, then the this number drops to $\frac{N_j(N_j+1)}{2}$. Furthermore, if the number of observations per cluster does not depend on the cluster, i.e. $N_j = N/J \;\; \forall j = 1,\ldots,J$, then the total number of possible neighborhood principal strata will be $\frac{N(N+J)}{2J^2}$.

Let us now consider, for every unit, three other variables, related to the neighborhood principal stratum, but each one of them accounting for the presence of one individual principal stratum in the unit's neighborhood:

$$nS_{ij}^{m_0 m_1} = G_{ij}\left(\boldsymbol{\delta}_{ij}\right) \tag{2.3.3}$$

where $\boldsymbol{\delta}_{ij}$ is a vector of indicator functions in $\mathcal{N}_{ij}$, $\boldsymbol{\delta}_{ij} = \left[\delta_{1j},\ldots,\delta_{i-1j},\delta_{i+1j},\ldots,\delta_{N_j j}\right]$, with δ_{kj} being being 1 if subject kj belongs to the individual principal stratum $S^{m_0 m_1}$ and 1 otherwise, with $k \in \mathcal{N}_{|}$

$$\delta_{kj} = \begin{cases} 1 & \text{if } S_{kj} = S^{m_0 m_1} \\ 0 & \text{otherwise} \end{cases}$$

If $G_{ij}(\cdot)$ is the proportion function, then $nS_{ij}^{m_0 m_1}$ is simply the proportion of subjects belonging to the individual principal stratum $S^{m_0 m_1}$ in $\mathcal{N}_{ij}$. Due to the properties of the function $G_{ij}(\cdot)$ we have the following constraint:

$$nS_{ij}^{00} + nS_{ij}^{11} + nS_{ij}^{01} = 1 \qquad\qquad \forall ij \tag{2.3.4}$$

It is easy to show the relation between the potential values of the neighbors' treat-

ment and these three variables:

$$N_{ij}(0) = nS_{ij}^{11} = 1 - (nS_{ij}^{00} + nS_{ij}^{01}); \qquad N_{ij}(1) = nS_{ij}^{01} + nS_{ij}^{11} = 1 - nS_{ij}^{00} \tag{2.3.5}$$

or interchangeably

$$nS_{ij}^{11} = N_{ij}(0); \qquad nS_{ij}^{01} = N_{ij}(1) - N_{ij}(0); \qquad nS_{ij}^{00} = 1 - N_{ij}(1) \tag{2.3.6}$$

2.4 Neighborhood Principal Strata Causal Effects

The intersection between an individual principal stratum $S^{m_0 m_1}$ and a neighborhood principal stratum $nS^{n_0 n_1}$ defines a superstratum that characterizes the compliance behavior to the encouragement assignment in terms of treatment uptake, both of the subject himself and of his neighbors. As we have already done in the previous chapter with individual principal strata, we can define the effect of the encouragement within each superstratum and level of covariates:

$$nPCE(m_0, m_1, n_0, n_1, \mathbf{c}) := \mathrm{E}\left[Y_{ij}(1) - Y_{ij}(0) \,|\, S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \tag{2.4.1}$$

Thanks to the relations shown in 2.3.5 and 2.3.6 we can also express the superstratum-specific principal causal effects as:

$$nPCE(m_0, m_1, n_0, n_1, \mathbf{c}) = \mathrm{E}\left[Y_{ij}(1) - Y_{ij}(0) \,|\, S_{ij} = S^{m_0 m_1}, nS_{ij}^{01} = n_1 - n_0, nS_{ij}^{11} = n_0, nS_{ij}^{00} = 1 - n_1, \mathbf{C}_{ij} = \mathbf{c}\right] \tag{2.4.2}$$

It is important to note that, given the constraint in (2.3.4), in the conditioning set only two out of the three variables nS_{ij}^{01}, nS_{ij}^{11} and nS_{ij}^{00} are sufficient.

2.4.1 Three-Way Decomposition: Individual Treatment Mediated Effect, Spillover Mediated Effect and Pure Encouragement Effect

In order to investigate how this effect is achieved we examine other contrasts of the potential outcomes that would result intervening on the values of A_j, M_{ij}, and N_{ij}. To assess the extent to which the encouragement has an effect on the outcome through a change in the individual treatment received we define, within each superstratum and level of covariates, *Individual Treatment Mediated Effect* ($iTME^a(m_0, m_1, n_0, n_1, \mathbf{c})$, with $a \in \{0,1\}$) as the contrast

$$\begin{aligned} iTME^a(m_0, m_1, n_0, n_1, \mathbf{c}) := \quad & \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1), N_{ij}(a)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ & - \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(0), N_{ij}(a)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \end{aligned} \tag{2.4.3}$$

In words, $iTME^a(m_0, m_1, n_0, n_1, \mathbf{c})$ is the average difference of the potential outcomes, within each superstratum and each level of the covariates, under two scenarios, one where we intervene to set the encouragement A_j to a, the actual treatment received by each unit i in cluster j, M_{ij}, to the value that would result having cluster j assigned to the active encouragement condition, $A_j = 1$, and the neighbor's treatment N_{ij} to the value that would occur if A_j were a, and the other where encouragement status and the neighbor's treatment are the same but we set the actual treatment received by unit ij to the value that would result having cluster j assigned to the control group, $A_j = 0$. In the vaccine study, the quantity represents the effect of the immunization camp on the TB bacterial load reported by the child ij at the end point visit, solely through a change in the vaccine uptake of the child himself.

Likewise we define as *Spillover Mediated Effect* ($sME^a(m_0, m_1, n_0, n_1, \mathbf{c})$, with $a \in \{0,1\}$) the following difference:

$$sME^{a}(m_0, m_1, n_0, n_1, \mathbf{c}) := \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1-a), N_{ij}(1)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1-a), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \quad (2.4.4)$$

It is the average effect on the outcome of unit ij, with values of covariates being $\mathbf{C}_{ij} = \mathbf{c}$ and belonging to the principal strata $S_{ij} = S^{m_0 m_1}$ and $nS_{ij} = nS^{n_0 n_1}$, of a change in the neighbors' treatment uptake N_{ij} from the value that corresponds to the control encouragement to the one corresponding to the active encouragement assignment, having the encouragement status in cluster j set to $A_j = a$ and the individual treatment receipt M_{ij} set to what it would be under the encouragement condition $A_j = 1 - a$. In the vaccine study the quantity represents the effect of the immunization camp on the TB bacterial load reported by the child ij at the end point visit, solely through a change in the vaccine received by the children located in the surrounding area, where the relevance of each of these children depend on function $G_{ij}(\cdot)$. Finally the encouragement intervention can result in a decrease or increase of the individual outcome Y_{ij} through a change in underlying cluster or individual behaviors other than in the treatment receipt. As already mentioned in the previous chapter, this can happen when encouragements incorporate information campaigns (e.g. advertisements, letters) that overall will increase the awareness or simply bring public attention to the problem. We have already referred to this effect as *Pure Encouragement Effect* (*PEE*). Formally, this quantity is encoded by the following contrast:

$$PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) := \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \quad (2.4.5)$$

In words, $PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c})$, with $\tilde{a} \in \{0, 1\}$, is the average difference, within each superstratum and level of covariates, of the potential outcomes under the two

encouragement conditions intervening to keep the individual treatment received, M_{ij}, and the neighbor's treatment, N_{ij}, fixed at the value they would take under $A_j = a$.

In the reference study, in each village belonging to the active group a social worker was responsible for informing mothers about the availability of the immunization camps and educating them about the benefits. These informative actions might be responsible of different mechanisms. A greater awareness and fear of TB infection might boost the use of preventive measures, such as warding off contact with known TB patients or sick people in general or avoiding crowded or enclosed environments. These behavioral changes and thus *PEE* will in general depend on the principal stratum the unit belongs to. If we can assume that always-takers are those who already have more knowledge on the disease, when the presence of immunization camps draws their attention to the problem we can hypothesize that they will react by intensifying their prevention strategies in a sharper way by, for instance, leading a healthier life style with balanced diet, avoidance of smocking and alcohol and good personal hygiene or by maintaining good indoor ventilation. Pure encouragement effect is also accounting for behavioral changes in the neighborhood. Since subjects belonging to different individual principal strata, will react differently to the components of the encouragement that are not responsible of a change in the treatment uptake, *PEE* will also depend on the proportion of always-takers, never-takers and compliers in the neighboring areas. Furthermore, the presence of an immunization camp in the village may also lead children or adults who already have the disease, and therefore not part of the study, but living in contact with participants, to tighten up their isolation or to take and adhere to the treatment. This effect would also be part of *PEE*.

It is easy to prove that the superstratum-specific principal causal effect can be written as the sum of the individual treatment mediated effect, the spillover mediated effect and the pure encouragement effect as follows:

$$nPCE(m_0, m_1, n_0, n_1, \mathbf{c}) = PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) + sME^{1-\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) + iTME^{1-\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) \tag{2.4.6}$$

Proof.

$$\begin{aligned}
nPCE(m_0, m_1, n_0, n_1, \mathbf{c}) =& \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(1), N_{ij}(1)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&- \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(1), N_{ij}(1)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&+ \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0), N_{ij}(1)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&- \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0), N_{ij}(1)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&+ \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&- \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&- \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
= PEE^0(m_0, m_1, n_0, n_1, \mathbf{c}) &+ sME^1(m_0, m_1, n_0, n_1, \mathbf{c}) + iTME^1(m_0, m_1, n_0, n_1, \mathbf{c})
\end{aligned}$$

The second equality is obtained by adding and subtracting two terms: the conditional expected value of $Y_{ij}\left(1, M_{ij}(0), N_{ij}(1)\right)$ and $Y_{ij}\left(1, M_{ij}(0), N_{ij}(0)\right)$. It is easy to show that the decomposition with $\tilde{a} = 1$ would be given by two different terms: the conditional expected values of $Y_{ij}\left(0, M_{ij}(1), N_{ij}(1)\right)$ and $Y_{ij}\left(0, M_{ij}(1), N_{ij}(0)\right)$. □

The choice between the two decompositions, with $\tilde{a} = \{0, 1\}$ depends on the applications and on the effects of interest for the study. Each effect has a specific meaning. For instance, sME^1 is the effect of a change in the neighbors' treatment, while there is an active encouragement condition in the cluster and the unit is left to take the treatment he would naturally take under the control encouragement status. This is an interesting quantity to investigate because it will give a sense of the spillover effect of the neighborhood while the clustered program is implemented

but the unit itself is left in the treatment condition corresponding to the control status. Moreover, potential outcomes based on hypothetical interventions on the intermediate variables involved in the effects of interest must be conceivable. We will get into the details of this problem in the following section. The choice of $\tilde{a}$ also depends on the validity of identifying assumptions that allow to recover information from the observed data on the corresponding effect. We will see what this means later in the chapter.

The four superstratum-specific causal effects in equation (2.4.6) can be marginalized over the conditional distribution of the neighborhood principal strata yielding the average effects in the individual principal stratum $S^{m_0 m_1}$, within levels of covariates:

$$\begin{aligned}
iTME^{a}(m_0, m_1, \mathbf{c}) &= \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1), N_{ij}(a)\right) - Y_{ij}\left(a, M_{ij}(0), N_{ij}(a)\right) | S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \int_{n_0}\int_{n_1} iTME^{a}(m_0, m_1, n_0, n_1, \mathbf{c}) \cdot \pi_{n_0 n_1}(m_0, m_1, \mathbf{c}) dn_0 dn_1 \\
sME^{a}(m_0, m_1, \mathbf{c}) &= \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1-a), N_{ij}(1)\right) - Y_{ij}\left(a, M_{ij}(1-a), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \int_{n_0}\int_{n_1} sME^{a}(m_0, m_1, n_0, n_1, \mathbf{c}) \cdot \pi_{n_0 n_1}(m_0, m_1, \mathbf{c}) dn_0 dn_1 \\
PEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) &= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) - Y_{ij}\left(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) | S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \int_{n_0}\int_{n_1} PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) \pi_{n_0 n_1}(m_0, m_1, \mathbf{c}) dn_0 dn_1
\end{aligned} \tag{2.4.7}$$

where $\pi_{n_0 n_1}(m_0, m_1, \mathbf{c}) := P(nS_{ij} = nS^{n_0 n_1} \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c})$ refers to the probability of belonging to the neighborhood principal stratum $nS^{n_0 n_1}$ conditional on the individual principal stratum and on baseline covariates.

2.4.2 The role of interference

When the intervention is assigned at cluster level, many different mechanisms can occur because of social interactions, such as the simple fact of living in the same area or more active interactions that can influence the behaviors of interacting individuals. In particular, in encouragement experiments, the effect of the cluster

encouragement on behavioral changes, including but not restricted to changes in the treatment uptake, in other participants living nearby, can also in turn have an effect on one's outcome. A deeper understanding of the role played by interference and how they are included in NEEs and iTMEs can help us to better interpret the results. In CEDs *interference* by the treatment uptake between units occurs when a unit's outcome depends on other subjects' treatment, violating unit-level SUTVA. In clustered encouragement designs mechanisms broadly referred to *interference* by neighbors' treatment uptake can happen through different types of processes: one's outcome may be affected by neighbors' treatment uptake either through their previous outcome (contagion effect) or through an effect on the unit's own behaviors or environmental factors or even because the effect of the neighbors' previous outcome on the subsequent unit's outcome is modified by their behavioral changes (infectiousness effect).

In the literature, the presence of interference is usually assessed through the evaluation of the effect of a change of the neighbors' treatment uptake, keeping everything else fixed. This is generally referred to as *spillover effects* and can be formalized by the causal estimand $\mathrm{E}\big[Y_{ij}(a,m,n') - Y_{ij}(a,m,n')\big]$, with $n \neq n'$. Spillover mediated effects are just a special case of this quantity with the additional interpretation of the effect of the encouragement "through" neighbors' treatment. Furthermore, if these quantities are not zero and interference is present, the difference of these spillover effects under alternative values of a and m provides additional insight into the mechanisms underlying this phenomenon. If spillover effects varies under different conditions a of the clustered encouragement, i.e. $\mathrm{E}\big[Y_{ij}(a,m,n') - Y_{ij}(a,m,n')\big] \neq \mathrm{E}\big[Y_{ij}(a',m,n') - Y_{ij}(a',m,n')\big]$ with $a' \neq a$, then interaction between this variable A_j and neighbors' treatment $\mathbf{M}_{-ij}$ on the effect on the outcome Y_{ij} is said to be present. For instance, the extent to which a unit's probability of being infected by an infectious disease can be affected by his neighbors' getting vaccinated might depend on the presence of an immunization campaign that might change preventive behaviors of

inhabitants of the treated villages. Similarly, an interaction between the individual treatment uptake M_{ij} and neighbors' treatment $\mathbf{M}_{-ij}$ on the effect on the outcome Y_{ij} is found with the inequality $\mathrm{E}\left[Y_{ij}(a,m,n')-Y_{ij}(a,m,n')\right] \neq \mathrm{E}\left[Y_{ij}(a,m',n')-Y_{ij}(a,m',n')\right]$ with $m' \neq m$. In the vaccine example, this interaction can be explained by saying that the effect of neighbors' treatment on a unit's risk of infection will most likely vary according to the vaccination status of the unit itself, since a vaccine already protects from infection with a high effectiveness on average.

Alternatively to spillover effects, to assess interference one may also be interested on the effect of the individual treatment uptake or of the clustered encouragement on the individual outcome under different pre-fixed levels of neighbors' treatment, quantified for example by $\mathrm{E}\left[Y_{ij}(a,1,n)-Y_{ij}(a,0,n)\right]$ and $\mathrm{E}\left[Y_{ij}(1,m,n)-Y_{ij}(0,m,n)\right]$, respectively. The variability of these effects with respect to n could provide some evidence on the presence of the interference. In fact, the two aforementioned interactions are responsible of this variability. Pure encouragement effects and individual treatment mediated effects follow in this category of effects. Therefore, a possible interaction between the treatment received by the neighbors, $\mathbf{M}_{-ij}$, and the treatment received by unit ij, M_{ij}, together with an interaction between A_j and M_{ij}, could partially explain the difference between individual treatment mediated effects for compliers, namely between $iTME^{a}(0,1,n_0,n_1,\mathbf{c})$ with $a=0$ or $a=1$. In fact, the effect of a vaccine on the unit who gets vaccinated will vary according to the overall coverage in the neighborhood, this being encoded by the variable N_{ij} we are intervening on to keep it at the value $N_{ij}(a)$. Likewise, the difference between the pure encouragement effects $PEE^{\tilde{a}}(m_0,m_1,n_0,n_1,\mathbf{c})$ with $\tilde{a}=0$ or $\tilde{a}=1$, as well as the difference between $nDCE(0,n,\mathbf{c})$ for never-takers and $nDCE(1,n,\mathbf{c})$ for always-takers, can be due to an interaction between the A_j and $\mathbf{M}_{-ij}$. For example, the effect of a mobile immunization camp through a change in preventive behaviors can depend on the vaccination coverage in the village, as a sign of the average sensitization to the problem of infectious diseases in the community, that is likely to have a reciprocal

influence between units.

Without an accurate analysis that specifically addresses the problem of spillover effects it can be hard to draw inference on these underlying mechanisms, because of the complicated structure of the causal pathways. Yet estimation of the pure encouragement effects, individual treatment mediated effects and spillover mediated effects would allow a researcher to assess whether the data support hypothesis on underlying mechanisms like those previously mentioned and also it could provide insight into the direction of spillover effects, if present.

2.4.3 Two-Way Decomposition: Individual Treatment Mediated Effect and Net Encouragement Effect

In this section we will show the relation between the quantities used in chapter 1 and those of the present chapter. As with the individual treatment mediated effect introduced in the previous chapter, the same notation is used to denote the average individual treatment mediated effect in $S^{m_0 m_1}$ resulting from the marginalization in (2.4.7). This is due to the fact that the potential outcome of the form $Y(a, M_{ij}(\tilde{a}), N_{ij}(a))$, where we let the neighbor's treatment take on the natural value it would exhibit under the present encouragement condition $A_j = a$, is equivalent to $Y(a, M_{ij}(\tilde{a}))$, so that

$$\begin{aligned} iTME^{a}(m_0, m_1, \mathbf{c}) =& \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1)\right) - Y_{ij}\left(a, M_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ =& \mathrm{E}\left[Y_{ij}\left(a, M_{ij}(1), N_{ij}(a)\right) - Y_{ij}\left(a, M_{ij}(0), N_{ij}(a)\right) | S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ =& \int_{n_0} \int_{n_1} iTME^{a}(m_0, m_1, n_0, n_1, \mathbf{c}) \cdot \pi_{n_0 n_1}(m_0, m_1, \mathbf{c}) \, d n_0 \, d n_1 \end{aligned} \tag{2.4.8}$$

As already mentioned, the net encouragement effect in CEDs can be viewed as a combination of two different mechanisms, namely interference by the treatment received by other subjects and other mechanisms that are not related to changes in the treatment receipt. The effects defined in this chapter allow us to formalize this

concept. Indeed, we can prove that the net encouragement effect decomposes into the spillover mediated effect and the pure encouragement effect:

$$NEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) = sME^{1-\tilde{a}}(m_0, m_1, \mathbf{c}) + PEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) \tag{2.4.9}$$

Proof.

$$\begin{aligned}
NEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) &= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a})\right) - Y_{ij}\left(0, M_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a}), N_{ij}(1)\right) - Y_{ij}\left(0, M_{ij}(\tilde{a}), N_{ij}(0)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a}), N_{ij}(1)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&\quad + \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(\tilde{a}), N_{ij}(0)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a}), N_{ij}(1)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&\quad + \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(\tilde{a}), N_{ij}(0)\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&= sME^{1-\tilde{a}}(m_0, m_1, \mathbf{c}) + PEE^{\tilde{a}}(m_0, m_1, \mathbf{c})
\end{aligned} \tag{2.4.10}$$

□

Critical Principal Strata: M-invariant, N-invariant and MN-invariant

Under homogeneity assumptions, identification of net encouragement effects in the entire population relies on those individual principal strata where $M_{ij}(0) = M_{ij}(1) = m$, namely never-takers and always-takers. As a matter of fact, units belonging to these strata are the only ones where potential outcomes of the type $Y_{ij}(a, M_{ij}(\tilde{a}))$ are observable. Here, we are going to focus on two particular types of units, those with $M_{ij}(0) = M_{ij}(1) = m$, i.e. with individual principal stratum of the form $S_{ij} = S^{mm}$, and those with $N_{ij}(0) = N_{ij}(1) = n$, i.e., with neighborhood principal stratum of the form $nS_{ij} = nS^{nn}$. We will refer to the former units as *M-invariant* and to the latter ones as as *N-invariant*.

As already mentioned, potential outcomes of the form $Y_{ij}\left(a, M_{ij}(\tilde{a}), N_{ij}(a')\right)$, with

$a \neq \tilde{a} \neq a'$, are in general not observable. As the analogous ones in the previous chapter, these can be called *a-priori counterfactuals*. In the three-way decomposition in (2.4.6), with any value of $\tilde{a} \in \{0,1\}$, we have two a-priori counterfactuals: $Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(1-\tilde{a})\big)$ and $Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big)$.

M-invariant superstratum: an M-invariant superstratum is a superstratum of the type $[S^{mm}, nS^{n_0 n_1}]$, which refers to the set of units with the same potential values of the individual treatment uptake, i.e. $M_{ij}(0) = M_{ij}(1) = m$. In such a principal superstratum, the potential outcome $Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(1-\tilde{a})\big)$ is actually in the observed data since $M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a})$. Therefore, this potential outcome is observed for those units of this type under the encouragement assignment $A_j = 1-\tilde{a}$. As a consequence, the individual treatment mediated effect for these units is zero:

$$iTME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) = 0 \qquad \forall \tilde{a} \in \{0,1\} \tag{2.4.11}$$

Proof.

$$\begin{aligned} iTME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) := \quad & \mathrm{E}\left[Y_{ij}\big(1-\tilde{a}, M_{ij}(1), N_{ij}(1-\tilde{a})\big) | S_{ij} = S^{mm}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ & - \mathrm{E}\left[Y_{ij}\big(1-\tilde{a}, M_{ij}(0), N_{ij}(1-\tilde{a})\big) | S_{ij} = S^{mm}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ & = \mathrm{E}\left[Y_{ij}\big(1-\tilde{a}, M_{ij}(0), N_{ij}(1-\tilde{a})\big) | S_{ij} = S^{mm}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ & - \mathrm{E}\left[Y_{ij}\big(1-\tilde{a}, M_{ij}(0), N_{ij}(1-\tilde{a})\big) | S_{ij} = S^{mm}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] = 0 \end{aligned}$$

□

N-invariant superstratum: an N-invariant superstratum is a superstratum of the typer $[S^{m_0 m_1}, nS^{nn}]$, which refers to the set of units with the same potential values of the neighbors' treatment receipt, i.e. $N_{ij}(0) = N_{ij}(1) = n$. In such a principal superstratum the spillover mediated effect for these units is zero:

$$sME^{1-\tilde{a}}(m_0, m_1, n, n, \mathbf{c}) = 0 \qquad \forall \tilde{a} \in \{0,1\} \tag{2.4.12}$$

Proof.

$$
\begin{aligned}
sME^{1-\tilde{a}}(m_0, m_1, n, n, \mathbf{c}) := \quad & \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(1)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& - \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(0)\right) | S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& = \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(0)\right) | S_{ij} = S^{mm}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& - \mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(0)\right) | S_{ij} = S^{mm}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right] = 0
\end{aligned}
$$

□

Finally, we can define a third superstratum which is both M-invariant and N-invariant.

MN-invariant superstratum: an MN-invariant superstratum is a superstratum of the type $[S^{mm}, nS^{nn}]$, which refers to the set of units with both the same potential values of the individual treatment uptake, i.e. $M_{ij}(0) = M_{ij}(1) = m$, and the same potential values of the neighbors' treatment receipt, i.e. $N_{ij}(0) = N_{ij}(1) = n$. In such a principal superstratum, the potential outcome $Y_{ij}\left(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right)$ is actually in the observed data since $M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a})$ and $N_{ij}(\tilde{a}) = N_{ij}(1-\tilde{a})$. Therefore, this potential outcome is observed for those units of this type under the encouragement assignment $A_j = 1-\tilde{a}$. As a consequence, the neighborhood principal causal effect $nPCE(m, m, n, n, \mathbf{c})$ for these units equals the pure encouragement effect:

$$
nPCE(m, m, n, n, \mathbf{c}) = PEE^{\tilde{a}}(m, m, n, n, \mathbf{c}) \qquad \forall \tilde{a} \in \{0, 1\} \tag{2.4.13}
$$

Proof.

$$
\begin{aligned}
nPCE(m, m, n, n, \mathbf{c}) = & \mathrm{E}\left[Y_{ij}(1) - Y_{ij}(0) \mid S_{ij} = S^{mm}, nS_{ij} = S^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& = \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(1), N_{ij}(1)\right) \mid S_{ij} = S^{mm}, nS_{ij} = nS^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0), N_{ij}(0)\right) \mid S_{ij} = S^{mm}, nS_{ij} = nS^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& = \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(1), N_{ij}(1)\right) \mid S_{ij} = S^{mm}, nS_{ij} = nS^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(1), N_{ij}(1)\right) \mid S_{ij} = S^{mm}, nS_{ij} = nS^{nn}, \mathbf{C}_{ij} = \mathbf{c}\right] \\
& = PEE^{1}(m, m, n, n, \mathbf{c})
\end{aligned}
$$

The proof is accomplished just by replacing the potential outcomes indexed only with one argument with the corresponding potential outcomes indexed with three arguments, and then by applying the properties of the superstratum, that is $M_{ij}(0) = M_{ij}(1)$ and $N_{ij}(0) = N_{ij}(1)$. Similarly, an opposite substitution of these potential values would yield $PEE^{0}(m, m, n, n, \mathbf{c})$. □

We refer to $nPCE^{1-\tilde{a}}(m, m, n, n, \mathbf{c})$ for these strata as *neighborhood dissociative causal effect*, denoted by $nDCE(m, n, \mathbf{c})$, which is dissociative of the effect on both post-assignment variables, M_{ij} and N_{ij}. These particular superstrata is the only one where the pure encouragement effect can be estimated from the observed data, similarly to never-takers and always-takers who were the only units who could provide information on net encouragement effects. Therefore, MN-invariant superstrata are crucial for the investigation of pure encouragement effects in the population, as we will see in later sections.

2.5 Identifying assumptions for Spillover Mediated Effects

As we did in section 1.5.1 in chapter 1, where we provided a set of assumptions allowing extrapolation of the a-priori counterfactual or of the overall net encouragement effect from never-takers to compliers, here we formulate similar homogeneity assumptions that would endorse both an extrapolation of pure encouragement effects from MN-invariant superstrata to all the other types of units, and an extrapolation of spillover mediated effects from M-invariant superstrata to compliers. While in chapter 1 we presented the assumptions in the particular case of interest for the application, postponing the generalization to the appendix, here - given the acquired familiarity of the reader with the concepts of this thesis - we will directly provide rather general assumptions.

Assumption 9. *Partial Homogeneity of the Pure Encouragement Effect across Principal Strata*

Partial homogeneity of the pure encouragement effect is assumed for specific values of $\tilde{a} \in \{0,1\}$ if the following identity holds $\forall \mathbf{c} \in \mathscr{C}, m_{\tilde{a}} = \{0,1\}, n_{\tilde{a}} \in [0,1]$:

$$\begin{aligned}
&\mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) | M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(\tilde{a}) = n_{\tilde{a}}, N_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\right] \\
&\equiv \\
&\mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) | M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(\tilde{a}) = n_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\right]
\end{aligned} \tag{2.5.1}$$

In words, it states that units sharing the same level of covariates, the same potential value of the individual and neighbors' treatment receipt under the encouragement status $\tilde{a}$, i.e. $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and $N_{ij}(\tilde{a}) = n_{\tilde{a}}$, have equivalent mean difference between potential outcomes under the two encouragement conditions, intervening to set the individual treatment receipt of unit ij to $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and his neighbors' treatment receipt to $N_{ij}(\tilde{a}) = n_{\tilde{a}}$, regardless of the potential value of the individual and neighbors' treatment receipt under the opposite encouragement status $1-\tilde{a}$, i.e. $M_{ij}(1-\tilde{a})$ and $N_{ij}(1-\tilde{a})$.

By virtue of the relations in (2.3.5) and (2.3.6), assumption 9 can also be expressed in terms of the presence of never-takers, always-takers and compliers in the unit's neighborhood.

$$\begin{aligned}
&\mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) | S_{ij} = S^{m_0 m_1}, nS_{ij}^{00} = 1 - n_1, nS_{ij}^{01} = n_1 - n_0, nS_{ij}^{11} = n_0, \mathbf{C}_{ij} = \mathbf{c}\right] \\
&\equiv \\
&\mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) | S_{ij} = S^{m'_0 m'_1}, nS_{ij}^{1-\tilde{a}\,1-\tilde{a}} = \tilde{a} - (2\tilde{a}-1)n_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\right] \quad \text{with } m'_{\tilde{a}} = m_{\tilde{a}}
\end{aligned} \tag{2.5.2}$$

This second expression of the partial homogeneity assumption 9 conveys the idea that the mean difference between potential outcomes under the two encouragement conditions, intervening to set the individual treatment receipt of unit ij to $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and his neighbors' treatment receipt to $N_{ij}(\tilde{a}) = n_{\tilde{a}}$, only depends on

the baseline covariates, the potential value of the individual treatment receipt under the encouragement status $\tilde{a}$, i.e., $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and on the presence of the principal stratum $S^{1-\tilde{a}1-\tilde{a}}$ in the unit's neighborhood, namely on the value of $nS_{ij}^{1-\tilde{a}1-\tilde{a}}$. If $\tilde{a} = 0$ it only depends on nS_{ij}^{11}, that accounts for presence of always-takers in the unit's neighborhood, whereas if $\tilde{a} = 1$ it only depends on nS_{ij}^{00}, that accounts for presence of never-takers in the surroundings.

In our vaccine example, the validity of assumption 9 for $\tilde{a} = 0$ requires that compliers and never-takers, both with $M_{ij}(0) = 0$, with the same level of covariates and the same presence of always-takers in the neighborhood, have an equivalent effect of the mobile immunization camp when the unit does not get vaccinated and his neighbors' are left under the vaccine status they would have without immunization camp, regardless of presence of never-takers and compliers in the unit's neighborhood. Similarly, the validity of assumption 9 for $\tilde{a} = 1$ requires that compliers and always-takers, both with $M_{ij}(1) = 1$, with the same level of covariates and the same presence of never-takers in the neighborhood, have an equivalent effect of the mobile immunization camp when the unit gets vaccinated and we intervene to set his neighbors' vaccine receipt to the one that they would naturally have with the immunization camp, regardless of presence of always-takers and compliers in the unit's neighborhood.

Assumption 9 with $\tilde{a} = 0$ is more feasible than with $\tilde{a} = 1$ if we assume that compliers and never-takers are more similar than compliers and always-takers are, in terms of their underlying characteristics affecting their compliance behavior. The hypothesized differences between compliers and never-takers outlined earlier (e.g. fear of side effects, misconception of vaccines, busy or ill relatives...) is likely not to affect their behavioral changes induced by the presence of immunization camps, while these children are kept without vaccine and their neighbors' vaccine status is kept unchanged. On the contrary, always-takers, assumed to be more educated or in general more sensible to the problem of tuberculosis, are more prone to respond

to any additional information received. Furthermore, if this is true, it makes sense to argue that the impact of the presence of immunization camps on never-takers' and compliers' outcome, net of any effect on the vaccine receipt, will not depend on the individual presence of never-takers and compliers in their neighborhood, but only on the presence of always-takers.

Theorem 3.

Part 1. If assumption 9 holds for a certain value of $\tilde{a} \in \{0,1\}$, then the pure encouragement effect

$PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c})$ for a superstratum $[S^{m_0 m_1}, nS^{n_0 n_1}]$, within levels of covariates, is equivalent to the neighborhood dissociative effect of an MN-invariant superstratum with individual principal stratum $S^{m_{\tilde{a}} m_{\tilde{a}}}$ with $M_{ij}(0) = M_{ij}(1) = m_{\tilde{a}}$ and neighborhood principal stratum $nS^{n_{\tilde{a}} n_{\tilde{a}}}$ with $N_{ij}(0) = N_{ij}(1) = n_{\tilde{a}}$:

$$PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) \equiv nDCE(m_0, n_0, \mathbf{c})(1-\tilde{a}) + nDCE(m_1, n_1, \mathbf{c})(\tilde{a}) = nDCE(m_{\tilde{a}}, n_{\tilde{a}}, \mathbf{c})$$

That is, if $\tilde{a} = 0$ the pure encouragement effect PEE^0 for compliers with any neighborhood principal stratum or for never-takers with $N_{ij}(0) \neq N_{ij}(1)$ is equivalent to the neighborhood dissociative causal effect of N-invariant never-takers with the same level of covariates and the same potential value of the neighbors' treatment receipt under the control encouragement condition, i.e. $N_{ij}(0) = N_{ij}(1) = n_0$. Analogously, if $\tilde{a} = 1$ the pure encouragement effect PEE^1 for compliers with any neighborhood principal stratum or for always-takers with $N_{ij}(0) \neq N_{ij}(1)$ is equivalent to the neighborhood dissociative causal effect of N-invariant always-takers with the same level of covariates and the same potential value of the neighbors' treatment receipt under the active encouragement condition, i.e. $N_{ij}(0) = N_{ij}(1) = n_1$.

Proof.

$$PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) = \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) - Y_{ij}\left(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, nS_{ij} = nS^{n_0 n_1}, \mathbf{C}_{ij} = \mathbf{c}\right]$$

$$= \mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}) = m_{1-\tilde{a}}, N_{ij}(\tilde{a}) = n_{\tilde{a}}, N_{ij}(1-\tilde{a}) = n_{1-\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\right]$$

$$= \mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(\tilde{a}) = n_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\right]$$

$$= \mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) \mid M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a}) = m_{\tilde{a}}, N_{ij}(\tilde{a}) = N_{ij}(1-\tilde{a}) = n_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\right]$$

$$= \mathrm{E}\left[Y_{ij}\left(1, m_{\tilde{a}}, n_{\tilde{a}}\right) - Y_{ij}\left(0, m_{\tilde{a}}, n_{\tilde{a}}\right) \mid S_{ij} = S^{m_{\tilde{a}} m_{\tilde{a}}}, nS_{ij} = nS^{n_{\tilde{a}} n_{\tilde{a}}}, \mathbf{C}_{ij} = \mathbf{c}\right] = nDCE(m_{\tilde{a}}, n_{\tilde{a}}, \mathbf{c})$$

The proof is accomplished by using the definition of $PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c})$ in the first equality, the consistency assumption in the second equality and finally by applying (2.5.1) in assumption 9 twice in the last equalities.

□

Part 2. If assumption 9 holds for a certain value of $\tilde{a} \in \{0, 1\}$, then the pure encouragement effect $PEE^{\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c})$, for a superstratum $[S^{m_0 m_1}, nS^{n_0 n_1}]$, within levels of covariates, is equivalent to the neighborhood dissociative effect of an MN-invariant superstratum with individual principal stratum $S^{m_{\tilde{a}} m_{\tilde{a}}}$ with $M_{ij}(0) = M_{ij}(1) = m_{\tilde{a}}$ and the same presence of the principal stratum $S^{1-\tilde{a}1-\tilde{a}}$ in the unit's neighborhood, namely with $nS_{ij}^{1-\tilde{a}1-\tilde{a}} = \tilde{a} - (2\tilde{a}-1)n_{\tilde{a}}$, but with $nS_{ij}^{01} = 0$:

$$\begin{aligned} PEE^{\tilde{a}}(m_0, m_1, nS_{ij}^{11} = n_0, 1 - nS_{ij}^{00} = n_1, \mathbf{c}) \equiv & nDCE(m_0, nS_{ij}^{11} = n_0, \mathbf{c})(1-\tilde{a}) \\ & + nDCE(m_1, 1 - nS_{ij}^{00} = n_1, \mathbf{c})(\tilde{a}) \\ = & nDCE(m_{\tilde{a}}, \tilde{a} - (2\tilde{a}-1)nS_{ij}^{1-\tilde{a}1-\tilde{a}} = n_{\tilde{a}}, \mathbf{c}) \end{aligned}$$

That is, if $\tilde{a} = 0$ the pure encouragement effect PEE^0 for compliers with any neighborhood principal stratum or for never-takers with $N_{ij}(0) \neq N_{ij}(1)$ is equivalent to the neighborhood dissociative causal effect of N-invariant never-takers with the same level of covariates and the same presence of always-takers in the neighborhood, $nS_{ij}^{11} = n_0$, but without any complier among the neighbors, i.e. $nS_{ij}^{01} = 0$. Analo-

gously, if $\tilde{a} = 1$ the pure encouragement effect PEE^1 for compliers with any neighborhood principal stratum or for always-takers with $N_{ij}(0) \neq N_{ij}(1)$ is equivalent to the neighborhood dissociative causal effect of N-invariant always-takers with the same level of covariates and the same presence of never-takers in the neighborhood, $n_{ij}S^{00} = 1 - n_1$, but without any complier among the neighbors, i.e. $nS_{ij}^{01} = 0$.

Proof. The proof follows from the expression of assumption 9 in (2.5.2). □

Theorem 3 is compelling for the investigation of causal mechanisms arising in clustered encouragement designs. Indeed, it allows the estimation of pure encouragement effect for non-MN-invariant principal strata by using the information provided by the observed data of MN-invariant principal strata.

Under assumption 9, deemed valid for a certain value of $\tilde{a}$, by virtue of theorem 3, $PEE^{\tilde{a}}$ for any superstratum of the type $\left[S^{m_0 m_1}, S^{n_0 n_1}\right]$ can be estimated from the observed data of units belonging to the superstratum of the type $\left[S^{m_{\tilde{a}} m_{\tilde{a}}}, S^{n_{\tilde{a}} n_{\tilde{a}}}\right]$. GIven this result, the observed data will also provide information on the spillover mediated effect in the particular superstratum $\left[S^{mm}, nS^{n_0 n_1}\right]$ with $M_{ij}(0) = M_{ij}(1) = m$ but $N_{ij}(0) = n_0 \neq N_{ij}(1) = n_1$.

Corollary 2. *If assumption 9 holds for a certain value of $\tilde{a}$, the spillover mediated effect in the superstratum $\left[S^{mm}, nS^{n_0 n_1}\right]$ is expressed as follows:*

$$sME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) = nPCE(m, m, n_0, n_1, \mathbf{c}) - nDCE(m, n_{\tilde{a}}, \mathbf{c})$$

Proof. By simply rearranging the terms in the three-decomposition in equation (2.4.6), we can write

$$sME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) = nPCE(m, m, n_0, n_1, \mathbf{c}) - iTME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) - PEE^{\tilde{a}}(m, m, n_0, n_1, \mathbf{c})$$

As shown in (2.4.11), in principal strata with $M_{ij}(0) = M_{ij}(1)$ the individual treatment

mediated effect is zero, so $iTME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) = 0$. The last step consists in applying the result expressed by part 1 of theorem 3 simply by substituting the term $PEE^{\tilde{a}}(m, m, n_0, n_1, \mathbf{c})$ with $nDCE(m, n_{\tilde{a}}, \mathbf{c})$, and the proof is complete. □

Information on the spillover mediated effect in the particular superstratum $\left[S^{mm}, nS^{n_0 n_1}\right]$ is then given by the observed data of units belonging to this superstratum as well as from units of the same M-invariant individual principal stratum S^{mm} and the same value of $N_{ij}(\tilde{a}) = n_{\tilde{a}}$ but belonging to an N-invariant neighborhood principal stratum $nS^{n_{\tilde{a}} n_{\tilde{a}}}$.

Assumption 10.A. *Partial Homogeneity of the Spillover Effect across Principal Strata*

Partial homogeneity of the spillover effect is said to be assumed for specific values of $\tilde{a} \in \{0, 1\}$ if the following identity holds $\forall \mathbf{c} \in \mathscr{C}, m_{\tilde{a}} = \{0, 1\}, n_{\tilde{a}} \in [0, 1]$:

$$\mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_1\right) - Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_0\right) | M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c}\right]$$

$$\equiv$$

$$\mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_1\right) - Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_0\right) | M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c}\right] \tag{2.5.3}$$

In words, the assumption holds for $\tilde{a} \in \{0, 1\}$ when the mean difference between potential outcomes under the two scenarios where the neighbors' treatment receipt is set to the values it would take under the two encouragement conditions, intervening to keep the individual treatment receipt of unit ij to $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and the clustered encouragement assignment to $A_j = 1 - \tilde{a}$, does not depend on the potential value of the individual treatment receipt under the opposite encouragement status $1 - \tilde{a}$, i.e. $M_{ij}(1 - \tilde{a})$, conditioning on baseline covariates, the individual treatment receipt under encouragement status $\tilde{a}$, i.e. $M_{ij}(\tilde{a}) = m_{\tilde{a}}$, and potential values of the neighbors' treatment receipt under the two encouragement conditions, i.e. $N_{ij}(0) = n_0$ and $N_{ij}(\tilde{a}) = n_1$.

By virtue of the relations in (2.3.5) and (2.3.6), assumption 10.A can also be expressed in terms of the presence of never-takers, always-takers and compliers in the unit's neighborhood:

$$\mathrm{E}\Big[Y_{ij}\big(1-\tilde{a},m_{\tilde{a}},n_1\big)-Y_{ij}\big(1-\tilde{a},m_{\tilde{a}},n_0\big)|S_{ij}=S^{m_0m_1},nS_{ij}^{00}=1-n_1,nS_{ij}^{01}=n_1-n_0,nS_{ij}^{11}=n_0,\mathbf{C}_{ij}=\mathbf{c}\Big]$$

$$\equiv$$

$$\mathrm{E}\Big[Y_{ij}\big(1,m_{\tilde{a}},n_{\tilde{a}}\big)-Y_{ij}\big(0,m_{\tilde{a}},n_{\tilde{a}}\big)|S_{ij}=S^{m'_0m'_1},nS_{ij}^{01}=n_1-n_0,nS_{ij}^{11}=n_0,\mathbf{C}_{ij}=\mathbf{c}\Big]\quad\text{with } m'_{\tilde{a}}=m_{\tilde{a}} \tag{2.5.4}$$

Again only two out of the three neighborhood variables are needed in the conditioning set.

This second expression of the partial homogeneity assumption 10.A conveys that the mean difference between potential outcomes under the two scenarios where the neighbors' treatment receipt is set to the values it would take under the two encouragement conditions, intervening to keep the individual treatment receipt of unit ij to $M_{ij}(\tilde{a})=m_{\tilde{a}}$ and the clustered encouragement assignment to $A_j=1-\tilde{a}$, only depends on the baseline covariates, the potential value of the individual treatment receipt under the encouragement status $\tilde{a}$, i.e. $M_{ij}(\tilde{a})=m_{\tilde{a}}$ and on the presence of the three principal strata in the unit's neighborhood, namely on the value of nS_{ij}^{00}, nS_{ij}^{01} and nS_{ij}^{11}, bearing in mind the constraint in (2.3.4) .

In the illustrative example, assumption 10.A with $\tilde{a}=0$ holds if compliers and never-takers, with the same baseline covariates and overall the same type of neighborhood, have an equivalent effect of a change in the neighbors' vaccine status, while an immunization camp is set up in their villages but they are kept without vaccine. Similarly, assumption 10.A with $\tilde{a}=1$ holds if compliers and always-takers, with the same baseline covariates and overall the same type of neighborhood, have an equivalent effect of a change in the neighbors' vaccine status, while an immunization camp is set up in their villages and they are vaccinated. Again, if we assume that always-takers are different types of children, it is reasonable to support assumption

10.A with $\tilde{a} = 0$, since compliers and always-takers might use different levels of preventive strategies affecting the impact of the neighbors' vaccine uptake. Moreover, the dependence on the neighborhood principal stratum as a whole is legitimated by the fact that the impact of a change in the neighbors' vaccine status is likely to depend on the vaccination coverage under the control condition, that is $N_{ij}(0) = nS_{ij}^{11}$, as well as on other cluster or neighborhood characteristics that are encoded by the neighborhood principal stratum and that may affect the individual outcome. Indeed, the presence of each individual principal stratum in the neighborhood might depend on neighbors' individual features and cluster factors, such as the level of education in the village, the population density, the distance from the closest public health facility, the amount of infection in the village and the vaccination coverage of the entire village population. For instance, the level of education of the neighbors has an influence on the extent to which they use preventive measures and hence on the probability of them getting infected and hence on transmitting the infection. The risk of transmitting the bacteria from infected people also depends on their readiness on getting treated, which in turn can depend on the distance from the closest dispensary.

Assumption 10.B. *Partial Homogeneity of the Spillover Effect across Principal Strata*

Partial homogeneity of the spillover effect is said to be assumed for specific values of $\tilde{a} \in \{0,1\}$ if the following identity holds $\forall \mathbf{c} \in \mathscr{C}, m_{\tilde{a}} = \{0,1\}, n_{\tilde{a}} \in [0,1]$:

$$\mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_1\right) - Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_0\right) | M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c}\right]$$

$$\equiv$$

$$\mathrm{E}\left[Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_1\right) - Y_{ij}\left(1-\tilde{a}, m_{\tilde{a}}, n_0\right) | M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(1) - N_{ij}(0) = n_1 - n_0, \mathbf{C}_{ij} = \mathbf{c}\right] \tag{2.5.5}$$

In words, it states that units sharing the same level of covariates, the same potential value of the individual treatment receipt under the encouragement status $\tilde{a}$, i.e.

$M_{ij}(\tilde{a}) = m_{\tilde{a}}$, and the same difference between the two potential values of the neighbors' treatment receipt, i.e. $N_{ij}(1) - N_{ij}(0) = n_1 - n_0$, have equivalent mean difference between potential outcomes under the two scenarios where the neighbors' treatment receipt is set to the values it would take under the two encouragement conditions, intervening to keep the individual treatment receipt of unit ij to $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and the clustered encouragement assignment to $A_j = 1 - \tilde{a}$, regardless of the potential value of the individual treatment receipt under the opposite encouragement status $1 - \tilde{a}$, i.e. $M_{ij}(1 - \tilde{a})$ and regardless of each individual value of $N_{ij}(0)$ and $N_{ij}(1)$.

By virtue of the relations in (2.3.5) and (2.3.6), assumption 10.A can also be expressed in terms of the presence of never-takers, always-takers and compliers in the unit's neighborhood:

$$\mathrm{E}\Big[Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_1\big) - Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_0\big) | S_{ij} = S^{m_0 m_1}, nS_{ij}^{00} = 1 - n_1, nS_{ij}^{01} = n_1 - n_0, nS_{ij}^{11} = n_0, \mathbf{C}_{ij} = \mathbf{c}\Big]$$

$$\equiv$$

$$\mathrm{E}\Big[Y_{ij}\big(1, m_{\tilde{a}}, n_{\tilde{a}}\big) - Y_{ij}\big(0, m_{\tilde{a}}, n_{\tilde{a}}\big) | S_{ij} = S^{m'_0 m'_1}, nS_{ij}^{01} = n_1 - n_0, \mathbf{C}_{ij} = \mathbf{c}\Big] \quad \text{with } m'_{\tilde{a}} = m_{\tilde{a}} \tag{2.5.6}$$

This second expression of the partial homogeneity assumption 10.B conveys that the mean difference between potential outcomes under the two scenarios where the neighbors' treatment receipt is set to the values it would take under the two encouragement conditions, intervening to keep the individual treatment receipt of unit ij to $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and the clustered encouragement assignment to $A_j = 1 - \tilde{a}$, only depends on the baseline covariates, the potential value of the individual treatment receipt under the encouragement status $\tilde{a}$, i.e. $M_{ij}(\tilde{a}) = m_{\tilde{a}}$ and on the presence of the principal stratum S^{01}, i.e. the compliers, in the unit's neighborhood, that is on the value of nS_{ij}^{01}.

In the illustrative example, assumption 10.B with $\tilde{a} = 0$ holds if compliers and never-takers, with the same baseline covariates and the same presence of compliers

in the neighborhood, have an equivalent effect of a change in the neighbors' vaccine status, while an immunization camp is set up in their villages but they are kept without vaccine. Similarly, assumption 10.B with $\tilde{a} = 1$ holds if compliers and always-takers, with the same baseline covariates and the same presence of compliers in the neighborhood, have an equivalent effect of a change in the neighbors' vaccine status, while an immunization camp is set up in their villages and they are vaccinated. Assumption 10.A is preferable to assumption 10.B when we can assume that the presence of compliers in the unit's neighborhood encodes the cluster characteristics that have the most influence.

Under one of the two homogeneity assumptions 10.A or 10.B the spillover mediated effect of a general superstratum with $M_{ij}(0) \neq M_{ij}(1)$ and $N_{ij}(0) \neq N_{ij}(1)$, can be recovered from the corresponding effect of units belonging to an M-invariant superstratum, which is identified under assumption 9 (see corollary 2). This result is formalized by the following theorems.

Theorem 4.A. *If assumption 10.A holds for a certain value of* $\tilde{a} \in \{0, 1\}$*, then the spillover mediated effect* $sME^{1-\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c})$ *for a superstratum* $[S^{m_0 m_1}, nS^{n_0 n_1}]$*, within levels of covariates, is equivalent to the corresponding spillover mediated effect of an M-invariant superstratum with individual principal stratum* $S^{m_{\tilde{a}} m_{\tilde{a}}}$*, where* $M_{ij}(0) = M_{ij}(1) = m_{\tilde{a}}$*, and the same neighborhood principal stratum* $nS^{n_0 n_1}$*, with* $N_{ij}(0) = n_0$ *and* $N_{ij}(1) = n_1$*:*

$$sME^{1-\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) \equiv sME^{1-\tilde{a}}(m_{\tilde{a}}, m_{\tilde{a}}, n_0, n_1, \mathbf{c})$$

That is, if $\tilde{a} = 0$ *the spillover mediated effect* sME^0 *for compliers (*$m_0 = 0$*) is equivalent to the spillover mediated effect of never-takers with the same level of covariates and the same potential values of the neighbors' treatment receipt under both encouragement condition. Analogously, if* $\tilde{a} = 1$ *the spillover mediated effect* sME^1 *for compliers (*$m_1 = 1$*) is equivalent to the spillover mediated effect of always-takers with*

the same level of covariates and the same potential values of the neighbors' treatment receipt under both encouragement condition.

The theorem can also be expressed as follows:

$$sME^{1-\tilde{a}}(m_0, m_1, nS_{ij}^{11} = n_0, 1 - nS_{ij}^{00} = n_1, \mathbf{c}) \equiv sME^{1-\tilde{a}}(m_{\tilde{a}}, m_{\tilde{a}}, nS_{ij}^{11} = n_0, 1 - nS_{ij}^{00} = n_1, \mathbf{c})$$

That is, if $\tilde{a} = 0$ *the spillover mediated effect* sME^0 *for compliers* ($m_0 = 0$) *is equivalent to the spillover mediated effect of never-takers with the same level of covariates and the same presence of the three principal strata in the unit's neighborhood, i.e.* $nS_{ij}^{00}, nS_{ij}^{01}$ *and* nS_{ij}^{1}. *Analogously, if* $\tilde{a} = 1$ *the spillover mediated effect* sME^1 *for compliers* ($m_1 = 1$) *is equivalent to the spillover mediated effect of always-takers with the same level of covariates and the same presence of the three principal strata in the unit's neighborhood, i.e.* $nS_{ij}^{00}, nS_{ij}^{01}$ *and* nS_{ij}^{1}.

Theorem 4.B. *If assumption 10.A holds for a certain value of* $\tilde{a} \in \{0, 1\}$, *then the spillover mediated effect*
$sME^{1-\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c})$ *for a superstratum* $[S^{m_0 m_1}, nS^{n_0 n_1}$, *within levels of covariates, is equivalent to the corresponding spillover mediated effect of an M-invariant superstratum* $[S^{m_{\tilde{a}} m_{\tilde{a}}}, nS^{n_0' n_1'}$ *with individual principal stratum* $S^{m_{\tilde{a}} m_{\tilde{a}}}$, *where* $M_{ij}(0) = M_{ij}(1) = m_{\tilde{a}}$ *and the same presence of principal stratum* S^{01} *in the unit's neighborhood, namely with* $nS_{ij}^{01} = n_1' - n_0' = n_1 - n_0$*:*

$$sME^{1-\tilde{a}}(m_0, m_1, n_0, n_1, \mathbf{c}) \equiv sME^{1-\tilde{a}}(m_{\tilde{a}}, m_{\tilde{a}}, n_0', n_1', \mathbf{c}) \quad \text{with } n_1' - n_0' = n_1 - n_0$$

That is, if $\tilde{a} = 0$ *the spillover mediated effect* sME^0 *for compliers* ($m_0 = 0$) *is equivalent to the corresponding spillover mediated effect of never-takers with the same level of covariates and the same difference between the two potential values of the neighbors' treatment receipt under both encouragement conditions. Analogously, if* $\tilde{a} = 1$ *the spillover mediated effect* sME^1 *for compliers* ($m_1 = 1$) *is equivalent to the*

corresponding spillover mediated effect of always-takers with the same level of covariates and the same difference between the two potential values of the neighbors' treatment receipt under both encouragement conditions.

The theorem can also be expressed as follows:'

$$sME^{1-\tilde{a}}(m_0, m_1, nS_{ij}^{11} = n_0, 1 - nS_{ij}^{00} = n_1, \mathbf{c}) \equiv sME^{1-\tilde{a}}(m_{\tilde{a}}, m_{\tilde{a}}, nS_{ij}^{11} = n'_0, 1 - nS_{ij}^{00} = n'_1, \mathbf{c})$$
$$with\ nS_{ij}^{01} = 1 - (nS_{ij}^{00} + nS_{ij}^{11}) \equiv n_1 - n_0 = n'_1 - n'_0$$

That is, if $\tilde{a} = 0$ *the spillover mediated effect* sME^0 *for compliers* ($m_0 = 0$) *is equivalent to the corresponding spillover mediated effect of never-takers with the same level of covariates and the same presence of the principal stratum* S^{01}*, i.e. the compliers, in the unit's neighborhood, that is on the value of* nS_{ij}^{01}*. Analogously, if* $\tilde{a} = 1$ *the spillover mediated effect* sME^1 *for compliers* ($m_1 = 1$) *is equivalent to the corresponding spillover mediated effect of always-takers with the same level of covariates and the same presence of the principal stratum* S^{01}*, i.e. the compliers, in the unit's neighborhood, that is on the value of* nS_{ij}^{01}*.*

Relation between Identifying assumptions

The results outlined in theorems 3 and 4 show the central role of assumptions 9 and 10.A/10.B for the investigation of spillover mechanisms. Analogously, in the previous chapter the same role was played by assumption 7 with respect to net encouragement effects, that is effects of the encouragement that are net of any effect on the individual treatment uptake. All the previous assumptions claim a partial homogeneity of the mean difference between two types of potential outcomes across principal strata. Assumption 7b concerns the mean difference between the two potential outcomes that are involved in the net encouragement effect $NEE^{\tilde{a}}$, that is $Y_{ij}\big(1, M_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a})\big)$, and its homogeneity across specific individual principal strata. Instead, assumptions 9 and 10.A/10.B concern, respectively, the two potential outcomes involved in the pure encouragement effect $PEE^{\tilde{a}}$, that is

$Y_{ij}\big(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big)$ and those involved in the spillover mediated effect $sME^{1-\tilde{a}}$, that is $Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(1)\big) - Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(0)\big)$, providing results on the homogeneity of their mean difference across both neighborhood and individual principal strata. In general, assumptions 9 and 10.A/10.B do not imply the identifying assumption 7, or rather its general version in 7b. Here we provide a set of sufficient conditions for assumption 7b of partial homogeneity of the mean difference between $Y_{ij}\big(1, M_{ij}(\tilde{a})\big)$ and $Y_{ij}\big(0, M_{ij}(\tilde{a})\big)$ across individual principal strata sharing the value of $M_{ij}(\tilde{a})$ and consequently, by theorem 2b, of partial homogeneity $NEE^{\tilde{a}}$.

Theorem 5 (Sufficient conditions for partial homogeneity of the net encouragement effect)**.**

If for a certain values of $\tilde{a} \in \{0,1\}$

a) assumption 9 holds

b) either assumption 10.A or 10.B holds

c) the probability of belonging to a specific neighborhood principal stratum does not depend on the potential value of the individual treatment uptake under encouragement condition $1-\tilde{a}$*,*

i.e. $\forall \mathbf{c} \in \mathscr{C}, \forall m_0, m_1 \in \{0,1\}, \forall n_0, n_1 \in [0,1]$

$$P(nS_{ij} = S^{n_0 n_1} \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}) = P(nS_{ij} = S^{n_0 n_1} \mid S_{ij} = S^{m_{\tilde{a}} m_{\tilde{a}}}, \mathbf{C}_{ij} = \mathbf{c})$$

then assumption 7b is satisfied.

Proof. Let us consider the following mean difference within individual principal strata: $\mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}\big) - Y_{ij}\big(0, m_{\tilde{a}}\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big]$. This represents the net encouragement effect $NEE^{\tilde{a}}$ in the particular individual principal stratum with potential values of the the individual treatment uptake $M_{ij}(\tilde{a}) = m$ and $M_{ij}(1-\tilde{a})$. In equation 2.4.9 we have shown that the net encouragement effect can be written as

the sum of pure encouragement and spillover mediated effects. Therefore, using the consistensy assumptions and similar manipulations used to prove equation 2.4.9,we can write:

$$
\begin{aligned}
&\mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}\big) - Y_{ij}\big(0, m_{\tilde{a}}\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad = \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad + \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a}), N_{ij}(1)\big) - Y_{ij}\big(0, M_{ij}(\tilde{a}), N_{ij}(0)\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad = \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad + \mathrm{E}\big[Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(1)\big) - Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(0)\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big]
\end{aligned}
$$

which can be averaged over all the possible values of $N_{ij}(0)$ and $N_{ij}(1)$

$$
\begin{aligned}
&\sum_{n_0 n_1} \Big[\mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(\tilde{a}) = n_{\tilde{a}}, N_{ij}(1-\tilde{a}) = n_{1-\tilde{a}} \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad + \mathrm{E}\big[Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(1)\big) - Y_{ij}\big(1-\tilde{a}, M_{ij}(\tilde{a}), N_{ij}(0)\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(0) = n_1, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c}\big]\Big] \\
&\times P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c})
\end{aligned}
$$

and by consistency we have

$$
\begin{aligned}
&\sum_{n_0 n_1} \Big[\mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}, n_{\tilde{a}}\big) - Y_{ij}\big(0, m_{\tilde{a}}, n_{\tilde{a}}\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(\tilde{a}) = n_{\tilde{a}}, N_{ij}(1-\tilde{a}) = n_{1-\tilde{a}} \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad + \mathrm{E}\big[Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_1\big) - Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_0\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}), N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c}\big]\Big] \\
&\times P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c})
\end{aligned}
$$

We now apply assumption 9 to the first term and assumption 10.A to the second one and we get

$$
\begin{aligned}
&\sum_{n_0 n_1} \Big[\mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}, n_{\tilde{a}}\big) - Y_{ij}\big(0, m_{\tilde{a}}, n_{\tilde{a}}\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(\tilde{a}) = n_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&\quad + \mathrm{E}\big[Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_1\big) - Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_0\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c}\big]\Big] \\
&\times P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c})
\end{aligned}
$$

that, using again assumption 9 on the first term, can also be expressed as

$$\sum_{n_0 n_1} \Big[\mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}, n_{\tilde{a}}\big) - Y_{ij}\big(0, m_{\tilde{a}}, n_{\tilde{a}}\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c} \big]$$
$$+ \mathrm{E}\big[Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_1\big) - Y_{ij}\big(1-\tilde{a}, m_{\tilde{a}}, n_0\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c} \big] \Big]$$
$$\times P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c})$$

As the second equivalence has been accomplished by adding and subtracting the same term, again the two differences between potential outcomes can be contracted in just the following

$$\sum_{n_0 n_1} \mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}, n_1\big) - Y_{ij}\big(0, m_{\tilde{a}}, n_0\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, N_{ij}(0) = n_0, N_{ij}(1) = n_1, \mathbf{C}_{ij} = \mathbf{c} \big]$$
$$\times P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c})$$

Finally, by condition c), if nS_{ij} does not depend on $M_{ij}(1-\tilde{a})$, then $P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}) = P(N_{ij}(0) = n_0, N_{ij}(1) = n_1 \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c})$, which leads to the conditional average

$$\mathrm{E}\big[Y_{ij}\big(1, m_{\tilde{a}}, n_1\big) - Y_{ij}\big(0, m_{\tilde{a}}, n_0\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c} \big]$$

that satisfies assumption 7b, concluding the proof. It is easy to show that assumption 10.A can be replaces by the stronger assumption 10.B, yielding the same result.

□

2.6 Hierarchical Models for Cluster Interventions

As in the previous chapter, we introduce here hierarchical models that can be used for the analysis. As we will see in the next section, the complete-data likelihood only depends on two models: a model for the potential outcome $Y_{ij}(a)$ and a model for the individual principal strata membership S_{ij}.

Individual Principal Strata Model

As for the model for the individual principal strata membership we will maintain the *ordinal mixed probit model* as it was defined in section 1.6. We report here the latent variable formulation with two linked probit models:

$$S_{ij} = \begin{cases} S^{00} & \text{if} \quad S_{ij}^{n} \equiv \boldsymbol{\alpha}_n^T Z_{ij}^{Sf} + \mathbf{a}_{nj}^T Z_{ij}^{Sr} + V_{ij} \leq 0 \\ S^{01} & \text{if} \quad S_{ij}^{n} \geq 0 \text{ and } S_{ij}^{c} \equiv \boldsymbol{\alpha}_c^T Z_{ij}^{Sf} + \mathbf{a}_{cj}^T Z_{ij}^{Sr} + U_{ij} \leq 0 \\ S^{11} & \text{if} \quad S_{ij}^{n} \geq 0 \text{ and } S_{ij}^{c} \geq 0 \end{cases} \tag{2.6.1}$$

where $\mathbf{Z}_{ij}^{Sf} = [1, \mathbf{C}_{ij}]$ and $\mathbf{Z}_{ij}^{Sr} = [1, \mathbf{X}_{ij}]$ are the covariate matrices of the fixed and random part, respectively, whereas U_{ij} and V_{ij} are two random terms independently distributed as $N(0,1)$.

Potential Outcome Model

Conversely, in order to disentangle the spillover mediated effect from the pure encouragement effect under assumptions 9 and 10, the potential outcome model has to be reformulated so as to include neighborhood principal strata indicators. Under the monotonicity assumption, while a binary encouragement assignment and a binary treatment lead to only three individual principal strata, the number of neighborhood principal strata will be $\frac{N(N+J)}{2J^2}$, which is already 55 with say $J = 10$ clusters and $N_j (= N/J) = 10$ observations per cluster. For this reason, modeling assumptions are necessary. For notational convenience, instead of the indicator for the neighborhood principal strata, we will use the three variables representing the presence of never-takers, compliers and always takers. As already said, because of the constraint in (2.3.4), only two of the three variables are independent and will be incorporated in the model. For practical convenience and coherence with the identifying assumptions, we will use nS_{ij}^{01} and $nS_{ij}^{1-\tilde{a}1-\tilde{a}}$, where $\tilde{a}$ depends on the value for which assumptions 9 and 10.A hold. In the previous chapter, the outcome used in

the bed nets application follows a relative binomial distribution. In the illustrative example of this chapter, the outcome is continuous. Therefore, letting $\mathbf{C}'_{ij} = [1, \mathbf{C}_{ij}]$, we will use the following hierarchical linear model for the potential outcome:

$$\begin{aligned} Y_{ij}(a) \mid S_{ij}, nS_{ij}^{01}, nS_{ij}^{1-\tilde{a}1-\tilde{a}}, \mathbf{C}_{ij} =& \boldsymbol{\beta}^{S_{ij}T}\mathbf{Z}_{ij}^{Yf} + \boldsymbol{b}_j^T\mathbf{Z}_{ij}^{Yr} + \epsilon_{ij} \\ =& \boldsymbol{\beta}_0^{S_{ij}T}\mathbf{C}'_{ij} + \boldsymbol{\beta}_1^{S_{ij}T}\mathbf{C}'_{ij}a + \boldsymbol{\beta}_2^{S_{ij}T}\mathbf{C}'_{ij}f_1\left(nS_{ij}^{1-\tilde{a}1-\tilde{a}}\right)a + \boldsymbol{\beta}_3^{S_{ij}T}\mathbf{C}'_{ij}f_2\left(nS_{ij}^{10}\right)a \\ &+ \boldsymbol{\beta}_4^{S_{ij}T}\mathbf{C}'_{ij}f_3\left(nS_{ij}^{10}nS_{ij}^{1-\tilde{a}1-\tilde{a}}\right)a + b_{0j} + \mathbf{b}_{1j}^T\mathbf{X}_{ij} + \epsilon_{ij} \\ & \epsilon_{ij} \sim N\left(0, \sigma_\epsilon^{2S_{ij}}\right) \end{aligned} \tag{2.6.2}$$

where for each principal stratum $\boldsymbol{\beta}^{S_{ij}}$ are the fixed effects and $\mathbf{b}_j$ are the random effects, with variable vectors $\mathbf{Z}_{ij}^{Yf} = \left[\mathbf{C}'_{ij}, \mathbf{C}'_{ij}a, \mathbf{C}'_{ij}f_1\left(nS_{ij}^{1-\tilde{a}1-\tilde{a}}\right)a, \mathbf{C}'_{ij}f_2\left(nS_{ij}^{10}\right)a, \mathbf{C}'_{ij}f_3\left(nS_{ij}^{10}nS_{ij}^{1-\tilde{a}1-\tilde{a}}\right)a\right]$ and $\mathbf{Z}_{ij}^{Yr} = \left[1, \mathbf{X}_{ij}\right]$ respectively, allowing for random intercepts and random individual covariates slopes. We let the variance of the individual random term $\sigma^{2S_{ij}}$ depend on the individual strata. It is worth noting that under assumption 10.B the interaction term cancels out, i.e. $\boldsymbol{\beta}_4 = \mathbf{0}$.

Under this parametrization, neighborhood principal causal effect for MN-invariant principal strata, referred to as neighborhood dissociative causal effect, is given by:

$$\begin{aligned} nDCE(m, n, \mathbf{c}) =& PEE^0(m, m, n, n, \mathbf{c}) = PEE^1(m, m, n, n, \mathbf{c}) \\ =& \boldsymbol{\beta}_1^{S^{mm}T}\mathbf{C}'_{ij} + \boldsymbol{\beta}_2^{S^{mm}T}\mathbf{C}'_{ij}f_1\left(n\tilde{a} + (1-n)(1-\tilde{a})\right) \end{aligned} \tag{2.6.3}$$

Now if assumption 9 holds with $\tilde{a} \in \{0, 1\}$, by theorem 3, pure encouragement effects for M-invariant principal $[S^{mm}, nS^{n_0n_1}]$, with $nS_{ij}^{01} \neq 0$, can be expressed as follows:

$$PEE^{\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) = \boldsymbol{\beta}_1^{S^{mm}T}\mathbf{C}'_{ij} + \boldsymbol{\beta}_2^{S^{mm}T}\mathbf{C}'_{ij}f_1\left(n_{\tilde{a}}\tilde{a} + (1-n_{\tilde{a}})(1-\tilde{a})\right) \tag{2.6.4}$$

The latter expression (2.6.4) also applies to pure encouragement effects $PEE^{\tilde{a}}(m_0, m_1, n'_0, n'_1, \mathbf{c})$ for non-M-invariant principal strata $[S^{m_0m_1}, nS^{n'_0n'_1}]$, i.e. compliers, with $M_{ij}(\tilde{a}) = m_{\tilde{a}} = m$ and $N_{ij}(\tilde{a}) = n'_{\tilde{a}} = n_{\tilde{a}}$. Under the same assumption, we are also provided

with an expression for spillover mediated effects for M-invariant principal strata $[S^{mm}, nS^{n_0 n_1}]$, with $nS_{ij}^{01} = n_1 - n_0$ and $nS_{ij}^{1-\tilde{a}1-\tilde{a}} = n_{\tilde{a}}\tilde{a} + (1 - n_{\tilde{a}})(1 - \tilde{a})$:

$$sME^{1-\tilde{a}}(m, m, n_0, n_1, \mathbf{c}) = \boldsymbol{\beta}_3^{S^{mm}T}\mathbf{C}'_{ij} f_2\Big((n_1 - n_0)\Big) + \boldsymbol{\beta}_4^{S^{mm}T}\mathbf{C}'_{ij} f_3\Big((n_1 - n_0)\big(n_{\tilde{a}}\tilde{a} + (1 - n_{\tilde{a}})(1 - \tilde{a})\big)\Big) \tag{2.6.5}$$

Finally if assumption 10.A holds with $\tilde{a} \in \{0, 1\}$, by theorem 4.A, equation (2.6.5) also estimates spillover mediated effects $sME^{1-\tilde{a}}(m_0, m_1, n'_0, n'_1, \mathbf{c})$ for non-M-invariant principal strata $[S^{m_0 m_1}, nS^{n_0 n_1}]$, i.e. compliers, with $M_{ij}(\tilde{a}) = m_{\tilde{a}} = m$, $n'_0 = n_0$ and $n'_1 = n_1$.

To reduce the amount of uncertainty associated to the outcome model, given by $\boldsymbol{\beta}$ coefficients, the random effects $\boldsymbol{b}_j$ with $J = 1, \ldots, J$, the individual and the neighborhood principal strata membership, we link the random coefficients $\boldsymbol{b}_j$ of the outcome model to the random coefficients $\boldsymbol{a}_{nj}$ and $\boldsymbol{a}_{cj}$ of the individual principal strata model in the following way:

$$\begin{aligned} \boldsymbol{b}_j &= \beta^r \mathbf{u}_j \\ \boldsymbol{a}_{nj} &= \alpha_n^r \mathbf{u}_j \qquad\qquad \mathbf{u}_j \sim N(\mathbf{0}, I) \\ \boldsymbol{a}_{cj} &= \alpha_c^r \mathbf{u}_j \end{aligned} \tag{2.6.6}$$

The random effects consist now of a cluster-specific random part $\mathbf{u}_j$, that is normally distributed and common to the outcome and the two principal strata probit models, and a different fixed coefficient for each model. It worth noting that this modeling assumption is plausible in most applications, because it stipulates that the cluster unmeasured factors that affect the outcome are also affecting the compliance behavior, allowing for different intensity of these effects. We can now simplify the expression of the outcome model as

$$Y_{ij}(a) \mid S_{ij}, nS_{ij}^{01}, nS_{ij}^{1-\tilde{a}1-\tilde{a}}, \mathbf{C}_{ij} = \boldsymbol{\beta}^{frS_{ij}T}\mathbf{Z}_{ij}^{Yfr} + \epsilon_{ij} \tag{2.6.7}$$

where $\boldsymbol{\beta}^{frS_{ij}} = [\boldsymbol{\beta}^{S_{ij}}, \beta^r]$ and $\mathbf{Z}_{ij}^{Yfr} = [\mathbf{Z}_{ij}^{Yf}, \mathbf{u}_j \cdot \mathbf{Z}_{ij}^{Yr}]$. Likewise, the individual principal strata model can be expressed as

$$S_{ij} = \begin{cases} S^{00} & \text{if} \quad S_{ij}^n \equiv \boldsymbol{\alpha}_n^{frT} Z_{ij}^{Sfr} + V_{ij} \leq 0 \\ S^{01} & \text{if} \quad S_{ij}^n \geq 0 \text{ and } S_{ij}^c \equiv \boldsymbol{\alpha}_c^{frT} Z_{ij}^{Sfr} + U_{ij} \leq 0 \\ S^{11} & \text{if} \quad S_{ij}^n \geq 0 \text{ and } S_{ij}^c \geq 0 \end{cases} \tag{2.6.8}$$

where $\boldsymbol{\alpha}_n^{fr} = [\boldsymbol{\alpha}_n, \alpha_n^r]$, $\boldsymbol{\alpha}_c^{fr} = [\boldsymbol{\alpha}_c, \alpha_c^r]$ and $\mathbf{Z}_{ij}^{Sfr} = [\mathbf{Z}_{ij}^{Sf}, \mathbf{u}_j \cdot \mathbf{Z}_{ij}^{Sr}]$. This new modeling scheme has an implication in the computation procedure as explained in the appendix in A 4.

2.7 Bayesian Inference

Due to the way we have defined the models in this chapter, the parameter vector $\boldsymbol{\theta}$ is now given by:

$$\boldsymbol{\theta} = (\boldsymbol{\beta}^{fr}, \boldsymbol{\alpha}^{fr}, \mathbf{u}, \boldsymbol{\sigma}_\epsilon^2)$$

where we have collected each set of parameters such that $\boldsymbol{\beta}^{fr} = (\boldsymbol{\beta}^{S^{00}}, \boldsymbol{\beta}^{S^{11}}, \boldsymbol{\beta}^{S^{01}}, \beta^r)$, $\boldsymbol{\alpha}^{fr} = [\boldsymbol{\alpha}_n^{fr}, \boldsymbol{\alpha}_c^{fr}]$, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_J)$, and $\boldsymbol{\sigma}_\epsilon^2 = [\sigma_\epsilon^{2S^{00}}, \sigma_\epsilon^{2S^{01}}, \sigma_\epsilon^{2S^{11}}]$. In the previous chapter, the complete-data likelihood function $\mathscr{L}(\boldsymbol{\theta}; \mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C})$, has been factorized into

$p(\mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C} \mid \boldsymbol{\theta}) = p(\mathbf{Y}^{obs} \mid \mathbf{S}, \mathbf{C}\ \boldsymbol{\theta}) p(\mathbf{S} \mid \mathbf{C},\ \boldsymbol{\theta}) p(\mathbf{C} \mid \boldsymbol{\theta})$. As before, we will assume that the random vector $\mathbf{u}_j$ accounts for all the unmeasured common factors affecting the outcome of all the units in cluster j, as well as unmeasured individual post-intermediate variables of every unit in the cluster affecting not only the unit's final outcome but also his neighbors', including the unit's outcome measured at previous time points or other behavior characteristics. As a consequence, given that two outcomes measured at the same time cannot causally affect one another, we make the assumption of independence between units' potential outcomes, conditioning on $\mathbf{u}_j$. According to cluster-level SUTVA (1), a unit's outcome does not depend

on other clusters' principal strata but only on those within the same cluster, i.e. $\mathbf{S}_j$. Moreover, the dependence of the subject's outcome from principal strata of other units within the same cluster is assumed to be through the neighborhood principal stratum, which is defined by a function of the two values $M_{ij}(0)$ and $M_{ij}(1)$ representing individual principal strata in the neighborhood. Finally, we assume independence between individual principal strata of different units, conditioning on the vector of random effects $\mathbf{a}_j$. For the foregoing reasons, letting $\delta_{ij}(S^{m_0m_1}, nS^{n_0n_1}) = \delta(S^{m_0m_1}, nS^{n_0n_1}, S_{ij}, nS_{ij})$ be 1 if $S_{ij} = S^{m_0m_1}$ and $nS_{ij} = nS^{n_0n_1}$ and 0 otherwise, we can write:

$$\mathscr{L}(\boldsymbol{\theta}; \mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C}, \mathbf{A}) = \prod_{j=1}^{\mathrm{J}} \prod_{i=1}^{N_j} \sum_{m_0m_1} \sum_{n_0n_1} \delta_{ij}(S^{m_0m_1}, nS^{n_0n_1}) p\big(Y_{ij} \mid A_j, S_{ij} = S^{m_0m_1}, nS_{ij} = nS^{n_0n_1}, \mathbf{C}_{ij}, \boldsymbol{\theta}\big) \times P\big(S_{ij} = S^{m_0m_1} \mid \mathbf{C}_{ij}, \boldsymbol{\theta}\big) p(\mathbf{C}_{ij} \mid \boldsymbol{\theta}) \tag{2.7.1}$$

The two models involved in the likelihood, for Y_{ij} and S_{ij}, have already been defined in (2.6.2) and (1.6.4) respectively.

2.7.1 Prior Specification

The prior distribution can be specified in a similar way to section 1.7.1. We assume an independence structure expressed in the following factorization of the prior distribution:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}^{fr}) p(\sigma_\epsilon^{2S^{00}}) p(\sigma_\epsilon^{2S^{01}}) p(\sigma_\epsilon^{2S^{11}}) p(\boldsymbol{\alpha}_n^{fr}) p(\boldsymbol{\alpha}_c^{fr}) \prod_j p(\mathbf{u}_j) \tag{2.7.2}$$

Since here the outcome is assumed to follow a normal distribution, normal prior distributions for the parameters of the outcome model reflect the conjugacy property. This will ease the computation providing a closed form for full conditional distributions and hence avoiding Metropolis-Hastings steps (see A 4 for details). Therefore, we posit a normal prior distribution for the coefficients of the outcome model as well as of the two principals strata probit probit models. The fixed effects of the

outcome model can be jointly modeled as

$$\boldsymbol{\beta}^{fr} \sim N\left(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0}\right) \tag{2.7.3}$$

whereas the fixed effects of the two principal strata models are separately modeled as

$$\boldsymbol{\alpha}_n^{fr} \sim N\left(\boldsymbol{\mu}_{\alpha 0}^n, \Lambda_{\alpha 0}^n\right) \quad \boldsymbol{\alpha}_c^{fr} \sim N\left(\boldsymbol{\mu}_{\alpha 0}^c, \Lambda_{\alpha 0}^c\right) \tag{2.7.4}$$

For the variances ${\sigma_\epsilon^2}^{S^{m_0 m_1}}$, due to its conjugacy property, we use an inverse-gamma distribution:

$${\sigma_\epsilon^2}^{S^{m_0 m_1}} \sim IG\left(\eta_0^\epsilon, s_0^\epsilon\right) \tag{2.7.5}$$

2.7.2 Imputation Approach for Finite Population Effects

So far we have defined superpopulation effects and, as seen in section 2.6, a model-based estimation can be accomplished by expressing them as a function of the parameters of the model for potential outcomes in (2.6.2). We extend here the bayesian procedure discussed in section 1.7.2 for the estimation of the finite population effects.

Relying on both homogeneity assumptions 9 and 10, we show how estimation of the finite population effects can be accomplished.

It is worth reminding that we denote with $(m_0, m_1, \mathbf{c})$ all effects within the individual principal stratum $S^{m_0 m_1}$ with level of covariates being $\mathbf{C}_{ij} = \mathbf{c}$, whereas $(m_0, m_1, n_0, n_1, \mathbf{c})$ implies the additional conditioning on the neighborhood principal stratum $nS^{n_0 n_1}$. Furthermore, we will denote all individual effects with subscript ij. Let $f_{m_0 m_1}^{n_0 n_1}(a \mid \mathbf{c})$ denote the predictive posterior distribution of the potential outcome $Y_{ij}(a)$:

$$f_{m_0 m_1}^{n_0 n_1}(a \mid \mathbf{c}) = p\left(Y_{ij}(a) \mid S_{ij} = S^{m_0 m_1}, nS_{ij}^{01} = n_1 - n_0, nS_{ij}^{1-\tilde{a}1-\tilde{a}} = n_{\tilde{a}}\tilde{a} + (1-n_{\tilde{a}})(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}, \mathscr{O}\right) \tag{2.7.6}$$

and $f_{m_0 m_1}^{n_0 n_1}(a \mid \mathbf{c}, \boldsymbol{\theta}^k)$ its conditional distribution evaluated at parameter values $\boldsymbol{\theta}$:

$$f^{n_0 n_1}_{m_0 m_1}(a \mid \mathbf{c}, \boldsymbol{\theta}^k) = p\big(Y_{ij}(a) \mid S_{ij} = S^{m_0 m_1}, nS^{01}_{ij} = n_1 - n_0, nS^{1-\tilde{a}1-\tilde{a}}_{ij} = n_{\tilde{a}}\tilde{a} + (1-n_{\tilde{a}})(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}, \boldsymbol{\theta}^k\big) \tag{2.7.7}$$

At each iteration $k = 1, \ldots, K$ of the MCMC, samples from the posterior distribution of individual and finite population effects are drawn as follows:

1. For each unit, belonging to any superstratum $[S^{m_0 m_1}, nS^{n_0 n_1}]$, i.e. $\forall i, j : S^k_{ij} = S^{m_0 m_1}, nS^k_{ij} = S^{n_0 n_1}$:

   a) Missing potential outcomes, $Y_{ij}^{mis} = Y_{ij}(1 - A_j^{obs})$, are imputed from their conditional distribution:

   $$Y_{ij}^{k,mis} \sim f^{n_0 n_1}_{m_0 m_1}\big(1 - A_j^{obs} \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k\big)$$

   b) Individual neighborhood principal causal effect is computed as:

   $$\widehat{nPCE}^k_{ij} = \big(2A_j^{obs} - 1\big)\big(Y_{ij}^{obs} - Y_{ij}^{k,mis}\big)$$

Let us now turn to the analysis of mechanisms. As already discussed, for MN-invariant principal strata of the type $[S^{mm}, nS^{nn}]$, i.e. never-takers and always-takers with no compliers in the neighborhood, pure encouragement effects coincide with the overall effect of the clustered encouragement, i.e. neighborhood principal causal effects, also called neighborhood dissociative causal effects in these strata. Therefore, for units belonging to these type of superstrata, i.e. $\forall i, j : S^k_{ij} = S^{mm}, nS^k_{ij} = S^{nn}$ with $m \in \{0, 1\}$ and $n \in [0, 1]$, we can write $\widehat{PEE}_{ij} = \widehat{nDCE}_{ij} = \widehat{nPCE}_{ij}$. On the contrary, for superstrata with $M_{ij}(0) \neq M_{ij}(1)$ or $N_{ij}(0) \neq N_{ij}(1)$, the overall effect of the encouragement does not coincide with pure encouragement effects, including in general individual treatment effect and spillover mediate encouragement effect. Indeed, in these principal strata potential outcomes of the type $Y_{ij}(a, M_{ij}(\tilde{a}), N_{ij}(\tilde{a}))$ with $a \neq \tilde{a}$ are not observable, hence pure encouragement effects are not identified from the observed data. Nevertheless, if assumption 9 holds with $\tilde{a} \in \{0, 1\}$, by virtue of theorem 3 estimation of pure encouragement effect can be accomplished using

the observed data of MN-invariant units. Note that individual pure encouragement effects cannot be identified, however their computation underpins the estimation of the corresponding finite population effects within principal strata. The estimation of $PEE_{ij}^{\tilde{a}}$ requires for each unit two potential outcomes: $Y_{ij}(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a}))$ and $Y_{ij}(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a}))$. Although information on one of them could be given by the observed data of units belonging to the same superstratum, theorem 3 involves an equivalence with MN-invariant superstrata of the mean difference between the two potential outcomes. Therefore, intuitively, both quantities should be drawn from posterior distributions of the corresponding MN-invariant superstratum.

3. For each unit belonging to superstratum $[S^{m_0 m_1}, nS^{n_0 n_1}]$, i.e.
   $\forall i, j : S_{ij}^{k} = S^{m_0 m_1}, nS_{ij}^{k} = S^{n_0 n_1}$:

   a) Both potential outcomes ${Y_{ij}}^{k}\big(a, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big)$ with $a = 0, 1$ are imputed from the likelihood distribution of $Y_{ij}(a)$ for the MN-invariant superstratum $[S^{mm}, nS^{nn}]$ with $m = m_{\tilde{a}}$ and $n = n_{\tilde{a}}$, given his values of covariates $\mathbf{C}_{ij}$:

$$ {Y_{ij}}^{k}\big(a, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) : \begin{cases} {Y_{ij}}^{k}\big(a, M_{ij}(0), N_{ij}(0)\big) \sim f_{m_0 m_0}^{n_0 n_0}(a \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k}) & \text{if } \tilde{a} = 0 \\ {Y_{ij}}^{k}\big(a, M_{ij}(1), N_{ij}(1)\big) \sim f_{m_1 m_1}^{n_1 n_1}(a \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k}) & \text{if } \tilde{a} = 1 \end{cases} $$

   b) Individual pure encouragement effects are computed:

$$ \widehat{PEE}_{ij}^{k,\tilde{a}} = {Y_{ij}}^{k}\big(1, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) - {Y_{ij}}^{k}\big(0, M_{ij}(\tilde{a}), N_{ij}(\tilde{a})\big) $$

With regard to spillover mediated effects for units with $N_{ij}(0) \neq N_{ij}(1)$ and hence $nS_{ij}^{01} \neq 0$, we should distinguish between those where the individual treatment uptake is not affected by the encouragement, i.e. $M_{ij}(0) = M_{ij}(1)$, from those where it is affected. For the former results depicted in corollary 2, whereas for the latter theorem 4.

4. For each unit belonging to an M-invariant superstratum of the type $[S^{mm}, nS^{n_0 n_1}]$,

that is never-takers and always takers with $n_0 \neq n_1$, individual spillover mediated effects are computed as follows:

$$\widehat{sME}_{ij}^{k,1-\tilde{a}} = \widehat{nPCE}_{ij}^{k} - \widehat{PEE}_{ij}^{k,\tilde{a}} \qquad \forall i,j : S_{ij}^{k} = S^{mm}, nS_{ij}^{k} = S^{n_0 n_1}$$

5. For each unit belonging to a non-M-invariant superstratum of the type $[S^{m_0 m_1}, nS^{n_0 n_1}]$, namely compliers with $m_0 \neq m_1$ and $n_0 \neq n_1$, i.e. $\forall i,j : S_{ij}^{k} = S^{m_0 m_1}, nS_{ij}^{k} = S^{n_0 n_1}$:

   a) ${Y_{ij}}^{k}\big(a, M_{ij}(\tilde{a}), N_{ij}(a)\big)$ with $a = 0,1$ are imputed from the likelihood distribution of $Y_{ij}(a)$ for the M-invariant superstratum $[S^{mm}, nS^{n_0' n_1'}]$ with $m = m_{\tilde{a}}$ and, depending on which of the two assumption 10.A or 10.B holds, with $n_0' = n_0$ and $n_1' = n_1$ or $n_1' - n_0' = n_1 - n_0$, given his values of covariates $\mathbf{C}_{ij}$:

$${Y_{ij}}^{k}\big(a, M_{ij}(\tilde{a}), N_{ij}(a)\big) : \begin{cases} {Y_{ij}}^{k}\big(a, M_{ij}(0), N_{ij}(a)\big) \sim f_{m_0 m_0}^{n_0 n_1}(a \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k}) & \text{if 10.A holds with } \tilde{a} = 0 \\ {Y_{ij}}^{k}\big(a, M_{ij}(0), N_{ij}(a)\big) \sim f_{m_0 m_0}^{n_0' n_1'}(a \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k}) \text{ with } n_1' - n_0' = n_1 - n_0 & \text{if 10.B holds with } \tilde{a} = 0 \\ {Y_{ij}}^{k}\big(a, M_{ij}(1), N_{ij}(a)\big) \sim f_{m_1 m_1}^{n_0 n_1}(a \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k}) & \text{if 10.A holds with } \tilde{a} = 1 \\ {Y_{ij}}^{k}\big(a, M_{ij}(1), N_{ij}(a)\big) \sim f_{m_1 m_1}^{n_0' n_1'}(a \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k}) \text{ with } n_1' - n_0' = n_1 - n_0 & \text{if 10.B holds with } \tilde{a} = 1 \end{cases}$$

   b) Individual net encouragement effect is then given by:

$$\widehat{NEE}_{ij}^{k,\tilde{a}} = {Y_{ij}}^{k}\big(1, M_{ij}(\tilde{a}), N_{ij}(1)\big) - {Y_{ij}}^{k}\big(0, M_{ij}(\tilde{a}), N_{ij}(0)\big)$$

   c) Since net encouragement effect consists of the spillover mediated effect and pure encouragement effect, as proven in (2.4.9), spillover mediated effects for these units can be obtained as follows:

$$\widehat{sME}_{ij}^{k,1-\tilde{a}} = \widehat{NEE}_{ij}^{k,\tilde{a}} - \widehat{PEE}_{ij}^{k,\tilde{a}}$$

6. Individual effects obtained in the previous steps are averaged out for each

individual principal stratum $S^{m_0 m_1}$ and within levels of covariates:

$$\widehat{PCE}^{k}(m_0, m_1, \mathbf{c}) = \frac{1}{|\mathscr{S}_c^{m_0 m_1}|} \sum_{i,j: S_{ij}^k = \mathscr{S}_c^{m_0 m_1}} \widehat{nPCE}_{ij}^{k}$$

$$\widehat{PEE}^{k,\tilde{a}}(m_0, m_1, \mathbf{c}) = \frac{1}{|\mathscr{S}_c^{m_0 m_1}|} \sum_{i,j: S_{ij}^k = \mathscr{S}_c^{m_0 m_1}} \widehat{PEE}_{ij}^{k,\tilde{a}}$$

$$\widehat{sME}^{k,1-\tilde{a}}(m_0, m_1, \mathbf{c}) = \frac{1}{|\mathscr{S}_c^{m_0 m_1}|} \sum_{i,j: S_{ij}^k = \mathscr{S}_c^{m_0 m_1}} \widehat{sME}_{ij}^{k,1-\tilde{a}}$$

where $\mathscr{S}_c^{m_0 m_1} = \{i, j : S_{ij}^k = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\}$. Again if the number of covariates is large and/or they are continuous some merging or marginalization across certain covariates is needed.

7. A last step is required for the estimation of individual treatment mediated effects for units with $M_{ij}(0) \neq M_{ij}(1)$, i.e. compliers $S_{ij} = S^{0,1}$:

$$\widehat{iTME}^{k,1}(01, \mathbf{c}) = \widehat{PCE}^{k}(01, \mathbf{c}) - \widehat{sME}^{k,1-\tilde{a}}(0, 1, \mathbf{c}) - \widehat{PEE}^{k,\tilde{a}}(0, 1, \mathbf{c})$$

These steps, for either assumption, are carried out repeatedly to account for the uncertainty in the imputation, resulting in the posterior distribution of the causal estimands. Finally, a summary statistics of these distributions, such as the mean or the median, can provide us with point estimates.

2.8 Application to the Illustrative Example: Simulation Study

In this section we apply our results to the aforementioned hypothetical study example. To investigate the performance of the proposed methodology and the bayesian estimation procedure, we conducted a simulation study based on a data generating model that partially mimics the real clustered encouragement experiment undertaken by Banarjee et al. (2010). As already said, we are going to focus on the effect of a mobile immunization camp on tuberculosis through a change in neighbors' BCG

vaccine uptake, referred here as *spillover mediated effect*. We have already explained in section 2.2 the variables used in the analysis. Suppose now that $J = 200$ villages are randomly assigned to either the active or control encouragement intervention, with $N_j = 50 \forall j$ observations in each village. We should remember that observations are healthy and unvaccinated children aged 0-18 months at baseline. We will make here some assumptions about the characterization of principal strata, that will in turn underpin further modeling and identifying assumptions. We posit a simplified setting where the compliance behavior of the study population hinges on few individual and cluster characteristics: sex, the level of parents' education, the distance of the village from the closest public health center, presence of family members in the child's daily life, knowledge and cultural beliefs about vaccination. In this simulation scenario, always-takers are those children belonging to educated families and/or living in villages that are not far from public health facilities . On the contrary, never-takers and compliers belong to less educated families and live in villages that are farther form urban center . The difference between these two principal strata relies on attitude towards immunization and sex. In fact, we can assume that never-takers belong to families characterized by resistance to vaccination due to lack of knowledge on the efficacy of immunization, fear of side effects or even to religious objections to vaccines. Moreover, gender differences in immunization status have been widely reported in India. Due to this gender discrimination, we can fairly think that females are more likely to never get vaccinated by their parents, that is to be never-takers. Suppose now that we only observe five individual and cluster characteristics for each study child: level of family education, sex, religion, level of family presence in the child's life, and distance of the village from the closest public health facility. Conversely, we do not have information neither on the number of infected subjects in the study villages nor on the level of knowledge and attitude of the families towards vaccination not related to religion. Therefore, the analysis will be based only on the five observed variables.

We will further assume that the unobserved factors that are responsible of the different immunization behavior of never-takers and compliers, i.e. unobserved factors related to the family's attitude towards vaccination, do not affect neither directly nor indirectly the disease outcome, i.e. bacterial load. For instance, we can assume that, despite of difference in knowledge and cultural beliefs about vaccination, never-takers and compliers have a similar behavior in terms of preventive measures and treatment uptake if infected with and without immunization camp. Conversely, we can suppose that always-takers are more prone to take preventive measures and also will get treated whenever infected.

In view of all the foregoing considerations, we can make the identifying assumption 9 and 10.A, both with $\tilde{a} = 0$, that is we can assume that, conditioning on the five observed covariates, pure encouragement effect PEE^0 and spillover effect sME^1 only depend on $M_{ij}(0)$ together with the presence in the neighborhood of always-takers and of the overall type of neighborhood, respectively. Furthermore, we can hypothesize that the protective effect of immunization camps through any type of mechanism are likely to decrease with the presence of alway-takers in the surroundings, given that this represents the immunization coverage without the immunization camp. In the present simulation study, we postulate the following realistic scenario where the expression of the effects within each super-stratum reflects the posited assumptions:

$$\begin{aligned}
PEE^0(1,1,nS^{11},nS^{11}+nS^{01},\mathbf{c}) &= -1+0.5nS^{11} \\
PEE^0(0,m,nS^{11},nS^{11}+nS^{01},\mathbf{c}) &= -2+nS^{11} && m=0,1 \\
sME^1(1,1,nS^{11},nS^{11}+nS^{01},\mathbf{c}) &= -2nS^{01}+nS^{01}nS^{11} \\
sME^1(0,m,nS^{11},nS^{11}+nS^{01},\mathbf{c}) &= -5nS^{01}+2nS^{01}nS^{11} && m=0,1 \\
iTME^1(0,1,nS^{11},nS^{11}+nS^{01},\mathbf{c}) &= -6+nS^{01}+1nS^{11}
\end{aligned} \tag{2.8.1}$$

Pure encouragement effect for always-takers is assumed to be given by an increased preventive and therapeutic behavior in the neighborhood, due to the presence of

an immunization camp. For never-takers and compliers pure encouragement effect is assumed higher because these subjects experience behavioral changes also for themselves. With regard to spillover mediated effect, always-takers are more likely to prevent contact with infected people, given the same type of neighborhood, and thus the protective effect of neighbors' being vaccinated is smaller than for compliers and never-takers. Finally, the greatest effect of immunization camps is assumed to the one through the individual vaccine receipt, that is the individual treatment mediated effect for compliers.

2.8.1 Data Generating Model

We consider a simplified setting with five observed variables, X_{1ij} for education, X_{2ij} for sex (1 if female, 0 if male), X_{3ij} for religion (1 if Islamic, 0 if others), X_{4ij} for the level of family presence (1 if low, 0 if high), and W_j for the distance of the village from the closest public health facility, with the following distributions:

$$X_{1ij} \sim discr.U(10) \quad X_{2ij} \sim ber(0.6) \quad X_{3ij} \sim ber(0.\bar{3})$$

$$X_{4ij} \sim ber(0.4) \quad W_j \sim discr.U(10)$$

The five variables are collected in the covariates vector $\mathbf{C}_{ij} = \left(X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij}, W_j\right)$. The clustered encouragement, i.e. the mobile immunization camp, is assumed to be assigned to clusters with probability equal to 0.5, hence A_j is distributed as:

$$A_j \sim ber(0.5)$$

Moreover, we have considered a generating model for individual principal strata membership that reflects the assumed characterization of individual principal strata:

$$S_{ij} = \begin{cases} S^{00} & \text{if} \quad S^n_{ij} \equiv 2.3 + 3.5X_{1ij} - 3X_{2ij} - 3.5X_{3ij} - 2X_{4ij} + 0.5W_j + a_{0nj} + V_{ij} \leq 0 \\ S^{01} & \text{if} \quad S^n_{ij} \geq 0 \text{ and } S^c_{ij} \equiv -1.5 + 5.1X_{1ij} - 0.3X_{2ij} - 1X_{3ij} + 1X_{4ij} - 2.5W_j + a_{0cj} + U_{ij} \leq 0 \\ S^{11} & \text{if} \quad S^n_{ij} \geq 0 \text{ and } S^c_{ij} \geq 0 \end{cases} \tag{2.8.2}$$

where U_{ij} and V_{ij} are independently distributed as $N(0,1)$. With respect to equation (2.6.1) there are only random intercepts a_{0nj} and a_{0cj}, with $\mathbf{a}_{1nj} = 0 \; \forall \, j = 1, \ldots, J$ and $\mathbf{a}_{1cj} = 0 \; \forall \, j = 1, \ldots, J$. The outcome follows a normal distribution. Below is the outcome generation model, mirroring the assumed effects in (2.8.1):

$$\begin{aligned} Y_{ij}(a) \mid S_{ij} = S^{00}, nS^{01}_{ij}, nS^{11}_{ij}, \mathbf{C}_{ij} &= 2.1 - X_{1ij} + 2W_j - 2a + 1nS^{11}_{ij}\, a - 5nS^{01}_{ij}\, a + 2nS^{10}_{ij}\, nS^{11}_{ij}\, a + b_{0j} + \epsilon_{ij} \\ Y_{ij}(a) \mid S_{ij} = S^{11}, nS^{01}_{ij}, nS^{11}_{ij}, \mathbf{C}_{ij} &= 0.6 - X_{1ij} + 2W_j - 1a + 0.5nS^{11}_{ij}\, a - 2nS^{01}_{ij}\, a + 1nS^{10}_{ij}\, nS^{11}_{ij}\, a + b_{0j} + \epsilon_{ij} \\ Y_{ij}(a) \mid S_{ij} = S^{01}, nS^{01}_{ij}, nS^{11}_{ij}, \mathbf{C}_{ij} &= 2.1 - X_{1ij} + 2W_j - 6a + 2nS^{11}_{ij}\, a - 4nS^{01}_{ij}\, a + 2nS^{10}_{ij}\, nS^{11}_{ij}\, a + b_{0j} + \epsilon_{ij} \end{aligned}$$

$$\epsilon_{ij} \sim N(0,1) \tag{2.8.3}$$

With respect to equation (2.6.2) we can see that we have made the following choices: $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$ and $f_4(\cdot)$ are all identity functions; there are no interactions between A_j and covariates, leading to effects that are independent of covariates levels; there is only a random intercept b_{0j} with $\mathbf{b}_{1j} = 0 \; \forall \, j = 1, \ldots, J$, and the variance of the individual random noise does not depend on the individual principal stratum. We considered two scenarios, a simplified one without random effects and the second one where random effects are present in all models.

Scenario 1: Absence of Random Effects in the Outcome Model

The first scenario simplifies the general setting in that there are no cluster unmeasured factors affecting the individual outcome, that is random effects are absent from the outcome model, whereas they are present in the compliance behavior as follows:

$$
\begin{aligned}
b_{0j} &= 0 \\
a_{0nj} &\sim N(0, 0.25) \qquad \forall\, j = 1, \ldots, J \\
a_{0cj} &\sim N(0, 0.25)
\end{aligned}
\tag{2.8.4}
$$

In the same way as in chapter 1, the two random effects, a_{0nj} and a_{0cj}, are independent. This will ease the computation of the estimation procedure.

Analysis Model

In the analysis we then use the following individual principal strata model

$$
S_{ij} = \begin{cases} S^{00} & \text{if} \quad S^n_{ij} \equiv \alpha_{0n} + \alpha_{1n} X_{1ij} + \alpha_{2n} X_{2ij} + \alpha_{3n} X_{3ij} + \alpha_{4n} X_{4ij} + \alpha_{5n} W_j + a_{0n} + V_{ij} \leq 0 \\ S^{01} & \text{if} \quad S^n_{ij} \geq 0 \text{ and } S^c_{ij} \equiv \alpha_{0c} + \alpha_{1c} X_{1ij} + \alpha_{2c} X_{2ij} + \alpha_{3c} X_{3ij} + \alpha_{4c} X_{4ij} + \alpha_{5c} W_j + a_{0c} + U_{ij} \leq 0 \\ S^{11} & \text{if} \quad S^n_{ij} \geq 0 \text{ and } S^c_{ij} \geq 0 \end{cases}
$$

with $a_{0nj} \sim N(0, \sigma^2_{a_n})$ and $a_{0cj} \sim N(0, \sigma^2_{a_c})$, and the following outcome model

$$
\begin{aligned}
Y_{ij}(a) \mid S_{ij}, nS^{01}_{ij}, nS^{11}_{ij}, \mathbf{X}_{ij}, W_j = {} & \beta_{00}^{S_{ij}} + \beta_{01} X_{1ij} + \beta_{02} W_j \\
& + \beta_{10}^{S_{ij}} a + \beta_{20}^{S_{ij}} nS^{11}_{ij} a + \beta_{30}^{S_{ij}} nS^{01}_{ij} a + \beta_{40}^{S_{ij}} nS^{11}_{ij} nS^{01}_{ij} a + \epsilon_{ij}
\end{aligned}
$$

$$
\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)
$$

Both are well specified according to the data generating models. The specification of the prior distribution has a similar structure to the one of the previous chapter in section 1.7.1, with normal distributed coefficients

$$
\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0}) \quad \boldsymbol{\alpha}_n \sim N(\boldsymbol{\mu}^n_{\alpha 0}, \Lambda^n_{\alpha 0}) \quad \boldsymbol{\alpha}_c \sim N(\boldsymbol{\mu}^c_{\alpha 0}, \Lambda^c_{\alpha 0})
$$

setting $\boldsymbol{\mu}_{\beta 0} = \boldsymbol{\mu}^n_{\alpha 0} = \boldsymbol{\mu}^c_{\alpha 0} = \mathbf{0}$ and $\Lambda_{\beta 0} = \Lambda^n_{\alpha 0} = \Lambda^c_{\alpha 0} = 10$ I, and the variance of the random effects following an inverse -gamma distribution, $\sigma^2_{a_n} \sim IG(\eta^n_0, s^n_0)$ and $\sigma^2_{a_c} \sim IG(\eta^c_0, s^c_0)$, where we set $\eta^n_0 = \eta^c_0 = 0.01$ and $s^n_0 = s^c_0 = 0.01$. In contrast to the prior specification of chapter 1, here the outcome model lacks any random effect but

requires the specification of the variance of the random noise: $\sigma_\epsilon^2 \sim IG\left(\eta_0^\epsilon, s_0^\epsilon\right)$, with $\eta_0^\epsilon = s_0^\epsilon = 0.01$.

Scenario 2: Presence of Random Effects in the Outcome Model

The second scenario generalizes the first one by including random effects in the outcome models, as specified in section 2.6:

$$\begin{aligned} b_{0j} &= 0.5u_j \\ a_{0nj} &= 0.5u_j \qquad u_j \sim N(0,1) \qquad \forall j = 1,\ldots,J \\ a_{0cj} &= 0.5u_j \end{aligned} \tag{2.8.5}$$

Here the three random effects are linked together by the random variable u_j. Cluster unmeasured factors, represented by u_j, affecting both the individual outcome and the compliance behavior, can be, for instance, the level of infection or the level of immunization in the cluster, the crowdedness, the absenteeism of health workers or the lack of sufficient of vaccine supplies in the nearest public health facility.

Analysis Model

In the analysis phase, models are changed with respect to the previous scenario by adding the random effects in the outcome model

$$\begin{aligned} Y_{ij}(a) \mid S_{ij}, nS_{ij}^{01}, nS_{ij}^{11}, \mathbf{X}_{ij}, W_j = {} & \beta_{00}^{S_{ij}} + \beta_{01} X_{1ij} + \beta_{02} W_j \\ & + \beta_{10}^{S_{ij}} a + \beta_{20}^{S_{ij}} nS_{ij}^{11} a + \beta_{30}^{S_{ij}} nS_{ij}^{01} a + \beta_{40}^{S_{ij}} nS_{ij}^{11} nS_{ij}^{01} a \\ & + b_{0j} + \epsilon_{ij} \end{aligned}$$

$$\epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right)$$

and specifying the three random effects as in (2.6.6). Prior specification follows section 2.7.1, with normal distributed coefficients

$$\boldsymbol{\beta}^{fr} \sim N(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0}) \quad \boldsymbol{\alpha}_n^{fr} \sim N(\boldsymbol{\mu}_{\alpha 0}^n, \Lambda_{\alpha 0}^n) \quad \boldsymbol{\alpha}_c^{fr} \sim N(\boldsymbol{\mu}_{\alpha 0}^c, \Lambda_{\alpha 0}^c)$$

setting $\boldsymbol{\mu}_{\beta 0} = \boldsymbol{\mu}_{\alpha 0}^n = \boldsymbol{\mu}_{\alpha 0}^c = \mathbf{0}$ and $\Lambda_{\beta 0} = \Lambda_{\alpha 0}^n = \Lambda_{\alpha 0}^c = 10$ I, and the variance of the

Table 1: Individual Principal Strata Rates and Neighborhood Characterization: mean and standard deviation of 500 Simulations

	Principal Strata Rates $P(S_{ij} = S^{m_0 m_1})$	Compliers in the Neighborhood $E\left[nS_{ij}^{01} \mid S^{m_0 m_1}\right]$	Always-takers in the Neighborhood $E\left[nS_{ij}^{11} \mid S^{m_0 m_1}\right]$
Principal Strata	Mean (SD)	Mean (SD)	Mean (SD)
NEVER-TAKERS	0.4219 (0.0069)	0.2527 (0.0088)	0.3144 (0.0081)
ALWAYS-TAKERS	0.3151 (0.0082)	0.2368 (0.0085)	0.3422 (0.0086)
COMPLIERS	0.2630 (0.0089)	0.3110 (0.0103)	0.2836 (0.0086)

random noise $\sigma_\epsilon^2 \sim IG\left(\eta_0^\epsilon, s_0^\epsilon\right)$, with $\eta_0^\epsilon = s_0^\epsilon = 0.01$.

2.8.2 Results

The generating process in both scenarios gives the following probabilities:

$$P(S_{ij} = S^{00}) = 0.26 \quad P(S_{ij} = S^{01}) = 0.42 \quad P(S_{ij} = S^{11}) = 0.32$$

These probabilities were empirically computed with 500 simulations using the data generating model (2.8.2) with random effects as in (2.8.4) of scenario 1. Scenario 2 would yield the same results, given that the two random effects a_{0nj} and a_{0cj}, even if not independent, have the same variance (=0.25) as those in scenario 1. Table 1 shows these individual principal strata rates as well as the average proportion of compliers and always-takers in the neighborhood, i.e. nS_{ij}^{01} and nS_{ij}^{11}.

500 data sets were generated from the model of each scenario. The bayesian estimation procedure outlined in section 2.7 was used to estimate the effects of interest. For a convergence check three chains were run for the first simulation of each model and only one chain was run for the subsequent simulations. Each chain consisted of 6000 iterations and 1000 of these were discarded as burn-in. Each model passed the convergence diagnostics of Gelman & Rubin (1996).

Scenario 1: Absence of Random Effects in the Outcome Model

Table 2: *Estimated Effects within Individual Principal Strata: Simulation Results of Scenario 1*

	PEE^0			sME^1		
Principal Strata	Mean	Median (SD)	95% Interval	Mean	Median (SD)	95% Interval
NEVER-TAKERS	-1.7278	-1.7276 (0.0813)	[-1.8877,-1.5693]	-0.7182	-0.7182 (0.0818)	[-0.8794,-0.5589]
ALWAYS-TAKERS	-0.8336	-0.8339 (0.0909)	[-1.0130,-0.6570]	-0.1544	-0.1543 (0.0883)	[-0.3380, 0.0078]
COMPLIERS	-1.7330	-1.7328 (0.0924)	[-1.9147,-1.5531]	-0.7831	-0.7830 (0.1044)	[-0.9951,-0.5862]
ALL	-1.4767	-1.4766 (0.0649)	[-1.6056,-1.3514]	-0.5803	-0.5802 (0.0677)	[-0.7317,-0.4665]

	$iTME^1$			PCE		
Principal Strata	Mean	Median (SD)	95% Interval	Mean	Median (SD)	95% Interval
NEVER-TAKERS		–		-2.4466	-2.4465 (0.0349)	[-2.5157,-2.3790]
ALWAYS-TAKERS		–		-0.9984	-0.9983 (0.0308)	[-1.0602,-0.9396]
COMPLIERS	-3.4204	-3.4205 (0.0587)	[-3.5347,-3.3049]	-5.9422	-5.9419 (0.0451)	[-6.0313,-5.8549]
ALL	-1.0046	-1.0045 (0.0189)	[-1.0427,-0.9686]	-3.0784	-3.0783 (0.0373)	[-3.1521,-3.0061]

Means, medians, standard deviations and 95% intervals of the posterior distribution of pure encouragement effects PEE^0, spillover mediated effect sME^1, individual treatment mediated effect $iTME^1$ and principal causal effects, are presented by individual principal strata. The last block of rows concerns the estimated effect in the whole population. All summary statistics are the average of the corresponding summary statistics obtained in 500 simulations.

Summary statistics of the posterior distributions of the estimated effects by individual principal strata are reported in Table 2. Pure encouragement effects and spillover mediated effects are greater for never-takers and compliers and the difference between these two principal strata is due to the different neighborhood. In fact, these effects are generated to be the same for never-takers and compliers with the same type of neighborhood (see (2.8.1)). As we can see in Table 1, never-takers have on average a slightly higher proportion of always-takers and a slightly lower proportion of compliers in the neighborhood, resulting mainly in a barely lower spillover mediated effect (sME(0,0): mean: -0.72, 95% interval: [-0.88,-0.56];

Table 3: Frequentist Performance of Bayesian Estimation Procedure for Causal Estimands in Sceanario 1 (500 simulations)

	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
NEVER-TAKERS								
PEE^0	93.5354	93.3333	0.0012	0.0014	-0.0693	-0.0791	0.0066	0.0066
sME^1	94.9495	94.9495	-0.0012	-0.0012	0.1706	0.1699	0.0067	0.0067
PCE	94.1414	94.3434	-0.0006	-0.0005	0.0237	0.0216	0.0012	0.0012
ALWAYS-TAKERS								
PEE^0	95.3535	95.5556	0.0104	0.0101	-1.2290	-1.2025	0.0084	0.0084
sME^1	94.7475	94.3434	-0.0026	-0.0025	1.6824	1.6404	0.0078	0.0078
PCE	94.5455	94.5455	-0.0024	-0.0023	0.2370	0.2356	0.0010	0.0010
COMPLIERS								
PEE^0	93.7374	93.5354	0.0050	0.0052	-0.2865	-0.2979	0.0086	0.0086
sME^1	94.1414	94.1414	-0.0051	-0.0050	0.6523	0.6457	0.0109	0.0109
$iTME^1$	91.9192	92.1212	0.0038	0.0037	-0.1096	-0.1086	0.0035	0.0035
PCE	94.9495	95.1515	-0.0022	-0.0019	0.0364	0.0320	0.0020	0.0020
ALL								
PEE^0	95.3535	95.3535	0.0053	0.0054	-0.3608	-0.3655	0.0042	0.0042
sME^1	95.5556	95.9596	-0.0033	-0.0032	0.5708	0.5606	0.0046	0.0046
$iTME^1$	93.3333	93.1313	0.0014	0.0015	-0.1382	-0.1462	0.0004	0.0004
PCE	95.7576	95.5556	0.0014	0.0013	0.0458	0.0411	0.0014	0.0014

sME(0,1): mean: -0.78, 95% interval: [-0.99,-0.59])). Although always-takers are more likely to be protected against infection thanks to preventive measure taken by their families, they still seem to benefit from the behavioral changes of the overall community responding to the presence of the immunization camp (PEE(1,1): mean: -0.83, 95% interval: [-1.01,-0.66], while there is little evidence that the camp affects always-takers through a change in the vaccination coverage of the village (sME(1,1): mean: -0.15, 95% interval: [-0.33,0.00]). This can occur, for example, if always-taker children are more exposed to adults than other children. Nevertheless, overall there is evidence of a spillover mediated effect in the population (sME: mean: -0.58, 95% interval: [-0.73,-0.47]).The greater effect is the individual treatment mediated effect for compliers (mean: -3.42, 95% interval: [-3.53,-3.30]). The individual treatment

mediated effect in the population falls by more than a third due to a proportion of 0.26 of compliers. The sum of the three effects gives a moderate average total effect of the mobile immunization camp on TB bacterial load (ITT: mean: -3.08, 95% interval: [-3.15,-3.00]).

Table 3 shows frequentist performance of the bayesian estimation procedure when the data generating process follows scenario 1. The performance is measured by the coverage (proportion of the time that the interval contains the true value) of the 95% credible interval using a normal approximation of the posterior distribution, the coverage of the quantile-based 95% credible interval, and three measures of the accuracy of both the mean and the median as parameter estimates, the bias, the percentage relative bias and the mean square error (MSE). Coverage rates are quite close to the nominal value of 95% and both point estimates, the mean and the median, show very little bias with a maximum MSE of 0.01. Therefore, even with noninformative priors, the Bayesian estimates have good frequentist properties.

Scenario 2: Presence of Random Effects in the Outcome Model

Summary statistics of the posterior distributions of the estimated effects are reported in Table 4. The simulation results for the Bayesian estimates of scenario 2 are similar to those of scenario 1. Both the mean and the median of the posterior distributions of all causal estimands are quite close to those in in Table 2. In contrast, the presence of a higher level of uncertainty, given by the presence of the random effect in the outcome model, results in bigger standard deviations and larger confidence intervals.

Table 5 shows frequentist performance of the bayesian estimation procedure when the data generating process follows scenario 2. The actual coverage probability is in general noticeably less than the nominal level, with an average across all effects of 82.8%, a maximum of 94.05% and a minimum of 78.17%. The bias of point estimates is still quite small but considerably larger than in scenario 1. In particular, the bias that has increased the most is the one for spillover mediated effects estimates.

Table 4: *Estimated Effects within Individual Principal Strata: Simulation Results of Scenario 2*

	PEE^0			sME^1		
Principal Strata	Mean	Median (SD)	95% Interval	Mean	Median (SD)	95% Interval
NEVER-TAKERS	-1.7906	-1.7911 (0.1351)	[-2.0540,-1.5248]	-0.6287	-0.6282 (0.1272)	[-0.8790,-0.3809]
ALWAYS-TAKERS	-0.9580	-0.9587 (0.1454)	[-1.2412,-0.6710]	-0.0889	-0.0883 (0.1383)	[-0.3619, 0.1801]
COMPLIERS	-1.7896	-1.7902 (0.1404)	[-2.0636,-1.5136]	-0.6936	-0.6933 (0.1408)	[-0.9703,-0.4188]
ALL	-1.5571	-1.5577 (0.1217)	[-1.7940,-1.3173]	-0.4966	-0.4961 (0.1159)	[-0.7251,-0.2709]

	$iTME^1$			PCE		
Principal Strata	Mean	Median (SD)	95% Interval	Mean	Median (SD)	95% Interval
NEVER-TAKERS		–		-2.4193	-2.4194 (0.0395)	[-2.4962,-2.3417]
ALWAYS-TAKERS		–		-1.0469	-1.0468 (0.0464)	[-1.1379,-0.9563]
COMPLIERS	-3.4329	-3.4330 (0.0556)	[-3.5417,-3.3240]	-5.9162	-5.9163 (0.0492)	[-6.0121,-5.8196]
ALL	-1.0103	-1.0102 (0.0175)	[-1.0447,-0.9763]	-3.0640	-3.0640 (0.0329)	[-3.1282,-2.9995]

Means, medians, standard deviations and 95% intervals of the posterior distribution of pure encouragement effects PEE^0, spillover mediated effect sME^1, individual treatment mediated effect $iTME^1$ and principal causal effects, are presented by individual principal strata. The last block of rows concerns the estimated effect in the whole population. All summary statistics are the average of the corresponding summary statistics obtained in 500 simulations.

Nevertheless, MSE is still small for all causal estimands, with a maximum of 0.035. In summary, even if the increased level of uncertainty of scenario 2 has affected the accuracy of Bayesian estimates, the estimation procedure has still good frequentist properties.

Table 5: Frequentist Performance of Bayesian Estimation Procedure for Causal Estimands of Scenario 2 (500 simulations)

	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
NEVER-TAKERS								
PEE^0	89.6825	89.4841	-0.0617	-0.0622	3.5923	3.6187	0.0261	0.0262
sME^1	86.9048	86.3095	0.0885	0.0889	-12.1749	-12.2408	0.0263	0.0263
PCE	86.3095	86.3095	0.0267	0.0266	-1.0936	-1.0898	0.0029	0.0029
ALWAYS-TAKERS								
PEE^0	86.3095	86.9048	-0.1138	-0.1145	13.6182	13.6993	0.0350	0.0351
sME^1	93.4524	93.4524	0.0629	0.0635	-37.6908	-38.1678	0.0223	0.0224
PCE	78.1746	78.5714	-0.0509	-0.0508	5.0923	5.0815	0.0054	0.0054
COMPLIERS								
PEE^0	90.6746	90.8730	-0.0519	-0.0524	3.0054	3.0376	0.0260	0.0261
sME^1	89.8810	89.4841	0.0845	0.0848	-10.7936	-10.8372	0.0284	0.0284
$iTME^1$	94.0476	94.8413	-0.0087	-0.0088	0.2612	0.2624	0.0033	0.0033
PCE	90.2778	90.0794	0.0238	0.0237	-0.4020	-0.3997	0.0035	0.0035
ALL								
PEE^0	88.2937	88.0952	-0.0729	-0.0735	4.9230	4.9617	0.0226	0.0227
sME^1	88.2937	88.4921	0.0805	0.0810	-13.9626	-14.0437	0.0209	0.0210
$iTME^1$	92.4603	92.4603	-0.0039	-0.0038	0.3909	0.3840	0.0004	0.0004
PCE	91.0714	91.0714	0.0037	0.0037	-0.1220	-0.1213	0.0015	0.0015

2.9 Concluding Remarks

This chapter extends the work presented in the previous one in that it is focused on a deeper investigation of how a clustered encouragement exerts its effect. The aim is to disentangle three causal mechanisms that may arise: through the individual uptake of the treatment, through the uptake of the treatment by other units of the same cluster, and through other processes not related to the treatment. In the decision-making process of the design phase, incorporating evidence on the diverse mechanisms, ensuing from the implementation of an intervention, can be crucial for increasing its cost-effectiveness. The first mechanism, here called *individual treatment mediated effect*, consists of the product of the effect of the encouragement on the treatment receipt and the effect of the treatment on the outcome of interest. This is the primary scope of the intervention. If the effect of the treatment is estimated to be high, but the encouragement is not much effective in achieving its main goal of increasing the treatment uptake (i.e. few compliers), the essential component of the intervention has to be improved. For example, if health workers hired to vaccinate children in the immunization camps turn out to be unreliable, not all the families that were convinced to vaccinate their children by the program promoters could actually get the vaccine. In this case, the fundamental element of the intervention, i.e. the realization of vaccination, is inadequate. The mechanism of interference, here referred to as *spillover mediated effect*, is of particular interest in resource limited settings, because, if beneficial, it can reinforce the individual treatment effect (for compliers) and it can also be a contribution for those whose treatment behavior is not affected by the encouragement (for never-takers and always-takers). If this effect is deemed substantial, one may, for example, think of an immunization camp supplied with fewer vaccines or fewer workers than necessary, given that even those who will not get the chance to receive the vaccine will get benefit from those who will. Even if this spillover effect is detrimental for those who do not get treated, it is useful to estimate the severity of such mechanism, because it provides insight into

what is narrowing down the overall effect and it urges to take measures to prevent this nuisance mechanism. Elucidating the effect of the clustered encouragement through interference is also important for predicting the impact of the program in its scale-up phase. The third mechanism, called *pure encouragement effect*, may play an important role when the intervention intrinsically involves a component that raises a general awareness on the issue related to the outcome and eventually promotes as a solution, not only the uptake of the specific treatment that is the focus of the intervention, but also other behavioral changes that may effective. An understanding of the efficacy of such component, net of the effect on the treatment uptake, can shed light on the possibility of improving the intervention by enhancing other elements that were not emphasized in the first place.

It is then clear how evidence-based policy demands that much greater priority is given to research that more reliably and relevantly identifies the potential mechanisms arising from public interventions, including information on heterogeneities of such mechanisms across subpopulations. The limited evidence on such causal mechanisms points to important gaps in evidence-based research. This thesis, and in particular this chapter, is an attempt to fill this gap. The framework presented in the previous chapter has been extended with the neighborhood principal stratification approach for the purpose of further disentangling the spillover mediated effect from the pure encouragement effect. The proposed homogeneity assumptions, that allow to identify such effects, take account of both individual and neighborhood compliance behavior, encoded by individual and neighborhood principal strata, i.e. superstrata. The two assumptions enables an extrapolation of information on pure encouragement effects from MN-invariant superstrata and on spillover mediated effects from M-invariant superstrata, respectively. As in the first chapter, our formalization of identifying assumptions allows a flexible specification of the principal strata involved in the homogeneity requirements, resulting in the identification of a combination of the three effects $PEE^{\tilde{a}}$, $sME^{1-\tilde{a}}$ and $iTME^{1-\tilde{a}}$, with $\tilde{a}=0$ or $\tilde{a}=1$.

The plausibility of each homogeneity assumptions has to be determined on a case-by-case basis, according to the prior knowledge on the application, and with the help of information on the similarity of principal strata that can be retrieved from the observed data.

The Bayesian estimation procedure accommodates an imputation-based approach where causal estimands are imputed on the basis of the hypothesized homogeneity assumptions. Simulation results show good frequentist performances for the Bayesian estimates, with an accuracy loss when the outcome model includes cluster-specific random terms. However, it is important to understand that these results have been obtained in two specific scenarios, with a particular sample size and specific values of the parameters. Hence, future work must be done to explore the extent to which the performance of our Bayesian estimation procedure depends on the sample size, the size of the effects and the variance of random terms. Likewise, future research can potentially address the difficulty in the specification of the functional form of the outcome model using a semi-parametric estimation, where the functions $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$ and $f_4(\cdot)$ are unknown. Another direction is to include a sensitivity analysis to assess the robustness of conclusions to departure from the homogeneity assumptions.

In any case, the framework proposed in this work, with the inclusion of neighborhood principal strata, provides a guideline to future research on spillover effects. As in the first chapter we used the individual treatment mediated effect to estimate the average treatment effect for the subpopulation of compliers (CACE), a natural direction for future research includes using spillover mediated effects to estimate the spillover effect of the treatment taken by other units for the subpopulation whose neighbors are affected by the encouragement. Moreover, even though an application from the vaccination field has been used to motivate the methodology proposed in this chapter, its use transcends beyond that specific context and is applicable to any clustered encouragement design.

Furthermore, the framework proposed could be extended to encouragement designs at individual level or to non-compliance settings where interference between units of the actual treatment receipt could bias the estimation of the treatment effect.

Appendix

A 1 Identifying assumptions for Net Encouragement Effects and Individual Treatment Mediated Effects

Here we provide a generalization of homogeneity assumptions (6) and (7) for net encouragement effects and individual treatment mediated effects presented in chapter 1. Each specification yields identification of $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ and the corresponding $iTME^{1-\tilde{a}}(m_0, m_1, \mathbf{c})$, with a specific value $\tilde{a} = 0, 1$ and for one of the two principal strata with $m_0 \neq m_1$. For each assumption we outline a comparison with sequential ignorability presented in section 1.5.

Assumption 6b. *Partial Stochastic Homogeneity of the Counterfactuals across Principal Strata*

Partial stochastic homogeneity of the counterfactuals across principal strata is said to be assumed if for specific values of $a, \tilde{a}, m \in \{0, 1\}$ if the following conditional independence holds:

$$Y_{ij}(a, m) \perp\!\!\!\perp M_{ij}(1 - \tilde{a}) \mid M_{ij}(\tilde{a}) = m, \mathbf{C}_{ij} = \mathbf{c} \qquad \forall \mathbf{c} \in \mathscr{C} \text{ and } \forall i, j$$

If assumption (6b) holds for a certain value of m and a certain value of $\tilde{a}$, with $a = \tilde{a}$, then the potential outcome $Y_{ij}(\tilde{a}, M_{ij}(\tilde{a}))$ is independent of $M_{ij}(1 - \tilde{a})$, conditioning on levels of covariates $\mathbf{C}_{ij}$ and on strata where $M_{ij}(\tilde{a}) = m$. In this particular case the assumption can be supported from the data if the distribution of outcomes under encouragement status $A_j = \tilde{a}$, within levels of covariates, is the same for the two strata that share the same potential value of the treatment receipt $M_{ij}(\tilde{a}) = m$.

When $a \neq \tilde{a}$, (6b) is an assumption on the distribution of potential outcomes of the form $Y_{ij}(a, M_{ij}(\tilde{a}))$, hence it is neither testable nor can find support in the data. However if assumption 6b holds for a certain value of m and a certain value of $\tilde{a}$, with $a = \tilde{a}$, we can also assume that it is valid for $a \neq \tilde{a}$.

The main result that follows from assumption (6b) is that if it is deemed valid for for specific values of $\tilde{a}$, a and m, then the two principal strata that share the same potential value $M_{ij}(\tilde{a}) = m$ present equal conditional mean of the potential outcome $Y_{ij}\big(a, M_{ij}(\tilde{a})\big)$:

$$\mathrm{E}\big[Y_{ij}\big(a, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}) = m_{1-\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\big] = \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a}) = m, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{A 1.1}$$

Theorem 1b. If assumption (6b) holds for $\tilde{a} = 0$, $a = 1$ and a specific value of $m \in \{0, 1\}$, the net encouragement effect $NEE^0(m, m_1, \mathbf{c})$ for the stratum S^{mm_1}, with $M_{ij}(0) = m$ and $M_{ij}(1) = m_1 \neq m$, within levels of covariates, is given by:

$$NEE^0(m, m_1, \mathbf{c}) = \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{mm}\mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{mm_1}\mathbf{C}_{ij} = \mathbf{c}\big]$$

Consequently, the individual treatment mediated effect $iTME^1(m, m_1, \mathbf{c})$ for the stratum S^{mm_1}, with $M_{ij}(0) = m$ and $M_{ij}(1) = m_1 \neq m$, within levels of covariates, is given by the following difference:

$$iTME^1(m, m_1, \mathbf{c}) = PCE(m, m_1, \mathbf{c}) - NEE^0(m, m_1, \mathbf{c})$$

If assumption (6b) holds for $\tilde{a} = 1$, $a = 0$ and a specific value of $m = 0, 1$, the net encouragement effect $NEE^1(m_0, m, \mathbf{c})$ for the stratum $S^{m_0 m}$, with $M_{ij}(0) = m_0 \neq m$ and $M_{ij}(1) = m$, within levels of covariates, is given by:

$$NEE^1(m_0, m, \mathbf{c}) = \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m}\mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0) \mid S_{ij} = S^{mm}\mathbf{C}_{ij} = \mathbf{c}\big]$$

Consequently, the individual treatment mediated effect $iTME^0(m_0, m, \mathbf{c})$ for the stra-

tum S^{mm_1}, with $M_{ij}(0) = m_0 \neq m$ and $M_{ij}(1) = m$, within levels of covariates, is given by the following difference:

$$iTME^0(m_0, m, \mathbf{c}) = PCE(m_0, m, \mathbf{c}) - NEE^1(m_0, m, \mathbf{c})$$

Proof. We show here the proof for the first part of the theorem relative to NEE^0. The proof simply uses the implication of assumption (6b) shown in (A 1.1), concerning homogeneity in terms of conditional mean:

$$\begin{aligned} NEE^0(m, m_1, \mathbf{c}) &= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0)\right) \mid S_{ij} = S^{mm_1}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0)\right) \mid S_{ij} = S^{mm_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ &= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0)\right) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}\left(0, M_{ij}(0)\right) \mid S_{ij} = S^{mm_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \\ &= \mathrm{E}\left[Y_{ij}(1) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] - \mathrm{E}\left[Y_{ij}(0) \mid S_{ij} = S^{mm_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \end{aligned}$$

where precisely the first equality, after the reported definition of NEE^0, makes use of the homogeneity of counterfactual conditional mean across the two strata and the second equality follows from the property of strata whose treatment uptake is unaffected by the encouragement, that is $Y_{ij}\left(1, M_{ij}(0)\right) = Y_{ij}\left(1, M_{ij}(1)\right)$. Similar manipulations demonstrate the second part of theorem. □

Corollary 3. *If assumption* (6b) *holds for* $\tilde{a} = 0$, $a = 1$ *and* $\forall m \in \{0, 1\}$, *the population mean of the counterfactual* $Y_{ij}(1, M_{ij}(0))$, *within levels of covariates, can be estimated using the following result:*

$$\mathrm{E}\left[Y_{ij}\left(1, M_{ij}(0)\right) \mid \mathbf{C}_{ij} = \mathbf{c}\right] = \sum_{m=0}^{1} \mathrm{E}\left[Y_{ij}(1) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] \sum_{m_1=m}^{1-m} \pi_{mm_1}(\mathbf{c})$$

so that the population $NEE^0(\mathbf{c})$ *is given by:*

$$NEE^0(\mathbf{c}) = \sum_{m=0}^{1} \mathrm{E}\left[Y_{ij}(1) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\right] \sum_{m_1=m}^{1-m} \pi_{mm_1}(\mathbf{c}) - \sum_{m_0=0}^{1} \sum_{m_1=0}^{1} \mathrm{E}\left[Y_{ij}(0) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right] \pi_{m_0 m_1}(\mathbf{c})$$

If monotonicity of compliers holds, the probability of defiers is zero, $\pi_{10} = 0$.

Proof. The second term of $NEE^0(\mathbf{c})$ is simply a weighted average of $Y_{ij}(0) = Y_{ij}\big(0, M_{ij}(0)\big)$ over the four principal strata. In the first term, $\mathrm{E}\big[Y_{ij}\big(1, M_{ij}(0)\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big]$, the same weighted average is performed but the change in the notation in the sums is used to distinguish the two different types of principal strata, so that:

$$\begin{aligned}\mathrm{E}\big[Y_{ij}\big(1, M_{ij}(0)\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big] &= \sum_{m=0}^{1} \sum_{m_1=m}^{1-m} \mathrm{E}\big[\big(1, M_{ij}(0)\big) \mid S_{ij} = S^{mm_1}\mathbf{C}_{ij} = \mathbf{c}\big]\pi_{mm_1}(\mathbf{c}) \\ &= \sum_{m=0}^{1} \mathrm{E}\big[Y_{ij}(1, m) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] \sum_{m_1=m}^{1-m} \pi_{mm_1} = \sum_{m=0}^{1} \mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{mm}, \mathbf{C}_{ij} = \mathbf{c}\big] \sum_{m_1=m}^{1-m} \pi_{mm_1}(\mathbf{c})\end{aligned}$$

where second equality follows from assumption (6b) and the consequent homogeneity in (A 1.1) for the two strata sharing the same potential value $M_{ij}(0) = m$. The last equality uses the fact that $Y_{ij}(1, m) = Y_{ij}(1)$ for strata where $M_{ij}(1) = m$. □

A similar result can be drawn for the counterfactual $NEE^1(\mathbf{c})$.

Remark

Assumption (6b) differs from the assumption of conditional unconfoundedness of the treatment receipt in (5) in a substantial way. (6b) assumes that, conditioning on levels of covariates, a potential outcome of the form $Y_{ij}\big(a, M_{ij}(\tilde{a})\big)$ only depends on one of the two potential values of the treatment receipt, precisely the one that we are assuming to keep fixed with the hypothetical intervention on M_{ij}, namely $M_{ij}(\tilde{a})$, and is instead independent of the other potential treatment receipt. On the contrary, the second assumption of sequential ignorability (5) requires the independence of the potential outcome from both potential values of the treatment receipt, so that it makes possible to extrapolate information across strata relying on the observed, instead of the potential, values of the treatment received. This substantial difference can be better understood if we express the identification formula (1.5.1), following from the sequential ignorability, in terms of principal strata:

$$\mathrm{E}\big[Y_{ij}\big(1, M_{ij}(0)\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big] = \sum_{m=0}^{1}\left(\sum_{m_0=m}^{1-m}\Big(\mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m}, \mathbf{C}_{ij} = \mathbf{c}\big]\frac{\pi_{m_0 m}(\mathbf{c})}{\pi_{mm}(\mathbf{c}) + \pi_{1-mm}(\mathbf{c})}\Big)\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})\right) \tag{A 1.2}$$

Proof. The proof starts by developing the population mean as a weighted average of the potential outcome over the four principal strata:

$$\mathrm{E}\big[Y_{ij}\big(1, M_{ij}(0)\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big] =$$

$$= \sum_{m=0}^{1}\sum_{m_1=m}^{1-m}\mathrm{E}\big[Y_{ij}(1, m) \mid M_{ij}(0) = m, M_{ij}(1) = m_1 \mathbf{C}_{ij} = \mathbf{c}\big]\pi_{mm_1}(\mathbf{c})$$

by virtue of unconfoundedness of the encouragement assignement (4)

$$= \sum_{m=0}^{1}\sum_{m_1=m}^{1-m}\mathrm{E}\big[Y_{ij}(1, m) \mid A_j = 0, M_{ij}(0) = m, M_{ij}(1) = m_1 \mathbf{C}_{ij} = \mathbf{c}\big]\pi_{mm_1}(\mathbf{c})$$

by virtue of unconfoundedness of the treatment receipt (5)

$$= \sum_{m=0}^{1}\mathrm{E}\big[Y_{ij}(1, m) \mid A_j = 0, \mathbf{C}_{ij} = \mathbf{c}\big]\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})$$

again by virtue of unconfoundedness of the encouragement assignement (4)

$$= \sum_{m=0}^{1}\mathrm{E}\big[Y_{ij}(1, m) \mid A_j = 1, \mathbf{C}_{ij} = \mathbf{c}\big]\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})$$

again by virtue of unconfoundedness of the treatment receipt (5)

$$= \sum_{m=0}^{1}\mathrm{E}\big[Y_{ij}(1, m) \mid A_j = 1, M_{ij}(1) = m, \mathbf{C}_{ij} = \mathbf{c}\big]\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})$$

we conclude the proof by taking now an average over all possible values of $M_{ij}(0)$

$$= \sum_{m=0}^{1}\sum_{m_0=m}^{1-m}\mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m}, \mathbf{C}_{ij} = \mathbf{c}\big]P\big(M_{ij}(0) = m_0 \mid M_{ij}(1) = m, \mathbf{C}_{ij} = \mathbf{c}\big)\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})$$

$$= \sum_{m=0}^{1}\left(\sum_{m_0=m}^{1-m}\Big(\mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m}, \mathbf{C}_{ij} = \mathbf{c}\big]\frac{\pi_{m_0 m}(\mathbf{c})}{\pi_{mm}(\mathbf{c}) + \pi_{1-mm}(\mathbf{c})}\Big)\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})\right)$$

□

If we now compare the identification result in corollary (3), yield by the homogeneity assumption (6b), with the identification result in equation (A 1.2), yield by the sequential ignorability assumptions (4) and (5), we can see that, in the latter, for all the strata where $M_{ij}(0) = m$, information on the mean of the counterfactual $Y_{ij}\big(1, M_{ij}(0)\big)$ for is taken from the mean value of the potential outcome $Y_{ij}(1)$ for those units where the potential value of the treatment received under $A_j = 1$, instead of $A_j = 0$, $M_{ij}(1)$, equals m. On the contrary, in (3), for the principal strata where $M_{ij}(0) = m$ and $M_{ij}(1) = m_1 \neq m$ information on the a priori counterfactual is borrowed just from those strata where $M_{ij}(0) = M_{ij}(1) = m$, who are the only ones for whom the mean value can be estimated from the data thanks to of the equality $Y_{ij}(1, M_{ij}(0)) \equiv Y_{ij}(1, M_{ij}(1)) \equiv Y_{ij}(1)$. For instance, when there are no defiers, this means to say that sequential ignorability allows to estimate $Y_{ij}(1, M_{ij}(0))$ for always-takers, where $M_{ij}(0) = 1$, not only from the values of $Y_{ij}(1) = Y_{ij}(1,1)$ for that sub-population but also borrowing information from the values of $Y_{ij}(1) = Y_{ij}(1,1)$ for compliers, whereas assumption (6b) does not use this extrapolation across these two strata.

A similar comparison could be shown for $\mathrm{E}\big[Y_{ij}\big(0, M_{ij}(1)\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big]$.

Assumption 7b. *Partial Homogeneity of the Mean Difference between Counterfactuals across Principal Strata*

Partial homogeneity of the mean difference between counterfactuals is said to be assumed if, for specific values of $\tilde{a} \in \{0,1\}$ and $m \in \{0,1\}$, the following identity holds:

$$\mathrm{E}\big[Y_{ij}(1,m) - Y_{ij}(0,m) \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}), \mathbf{C}_{ij} = \mathbf{c}\big]$$

$$=$$

$$\mathrm{E}\big[Y_{ij}(1,m) - Y_{ij}(0,m) \mid M_{ij}(\tilde{a}) = m, \mathbf{C}_{ij} = \mathbf{c}\big] \qquad \forall \mathbf{c} \in \mathscr{C}$$

In words, it states that the mean difference between potential outcomes under the two encouragement conditions and intervening to set the treatment receipt of each unit to the value it would take if A_j were set to $\tilde{a}$, i.e. $M_{ij}(\tilde{a}) = m$, is independent of the potential value of the treatment receipt under the opposite encouragement

status, $M_{ij}(1-\tilde{a})$.

Theorem 2b. If assumption (7b) is satisfied for a certain value of $\tilde{a} \in \{0,1\}$ and a specific value of $m \in \{0,1\}$, the net encouragement effect $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$, within levels of covariates, for the principal stratum $S^{m_0 m_1}$ where $M_{ij}(\tilde{a}) = m_{\tilde{a}} = m$, is given by:

$$NEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) \equiv DCE(m_0, \mathbf{c})(1-\tilde{a}) + DCE(m_1, \mathbf{c})(\tilde{a}) = DCE(m_{\tilde{a}}, \mathbf{c}) \qquad \text{(A 1.3)}$$

That is, if $\tilde{a} = 0$ the corresponding net encouragement effect for compliers ($m_0 = 0$) or defiers ($m_0 = 1$), depending on the value of m, is equal to the dissociative causal effect of never-takers or always-takers, respectively. Analogously, if $\tilde{a} = 1$ the corresponding net encouragement effect for compliers ($m_1 = 1$) or defiers ($m_1 = 0$), depending on the value of m, is equal to the dissociative causal effect of always-takers or never-takers, respectively.

Proof. The proof is accomplished by using the definition of $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ in (1.4.5):

$$\begin{aligned} NEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) &= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a})\right) - Y_{ij}\left(0, M_{ij}(\tilde{a})\right) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c} \right] \\ &= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a})\right) - Y_{ij}\left(0, M_{ij}(\tilde{a})\right) \mid M_{ij}(0) = m_0, M_{ij}(1) = m_1, \mathbf{C}_{ij} = \mathbf{c} \right] \end{aligned}$$

Let us rewrite the potential values of the treatment receipt using $\tilde{a}$ and $1-\tilde{a}$ so that this proof can apply to any value of $\tilde{a}$

$$= \mathrm{E}\left[Y_{ij}\left(1, M_{ij}(\tilde{a})\right) - Y_{ij}\left(0, M_{ij}(\tilde{a})\right) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, M_{ij}(1-\tilde{a}) = m_{1-\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c} \right]$$

Now the proof simply proceeds by applying assumption (7b) twice

$$
\begin{aligned}
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = m_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a}) = m_{\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\big] \\
&= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) - Y_{ij}\big(0, M_{ij}(\tilde{a})\big) \mid S_{ij} = S^{m_{\tilde{a}} m_{\tilde{a}}}, \mathbf{C}_{ij} = \mathbf{c}\big] = DCE(m_{\tilde{a}}, \mathbf{c})
\end{aligned}
$$

□

Remark

Assumption (7b) differs from the assumption of conditonal ignorability of the treatment receipt in (5) on three main provisions. First, the latter states a stochastic independence whereas the former is an assumption about independence in terms of the expected value. Second, conditonal ignorability of the treatment receipt concerns separately each counterfactual, whereas (7b) concerns a difference between pairs of counterfactuals. Third, A way to interpret (5) is saying that the counterfactual $Y_{ij}(a, m)$ does not depend neither on $M_{ij}(\tilde{a})$ nor on $M_{ij}(1-\tilde{a})$, conditioning on levels of covariates and the observed encouragement, so that information on $Y_{ij}(a, m)$, for for all units, can be extrapolated from $Y_{i'j'}(a)$ for all those units with $M_{i'j'}(a) = m$, regardless of the values of $M_{ij}(a)$, $M_{ij}(1-a)$ and $M_{i'j'}(1-a)$. Conversely, partial homogeneity assumption (7b) is solely based on the independence of the mean difference between potential outcomes $Y_{ij}(1, m)$ and $Y_{ij}(0, m)$ from $M_{ij}(1-\tilde{a})$, conditioning on covariates but more important on $M_{ij}(\tilde{a}) = m$, with specific values of $a, \tilde{a}$ and m. This means that extrapolation across strata is only carried out for the a priori counterfactual $Y_{ij}(a, m)$ for those whose compliance behavior is given by $M_{ij}(\tilde{a}) = m$ and $M_{ij}(1-\tilde{a}) \neq m$ from $DCE(m, \mathbf{c})$ for the principal stratum with the same value m of treatment receipt under both encouragement conditions, i.e. $M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a}) = m$. For these three reason we can conclude that assumption (7b) of partial homogeneity is a much weaker assumption that the second of the sequential ignorability assumptions. Mixing information across strata with the same behavior under a specific encouragement assignment seems more reasonable that

mixing across all the principal strata, especially when these strata are most likely very different because of the presence of latent characteristics.

Furthermore, note that the first two differences between assumptions (7b) and (5) also apply to a comparison between assumptions (7b) and (6b). Intuitively in general it is more plausible to assume homogeneity in terms of a mean difference rather that a stochastic homogeneity of each specific counterfactual.

Theorems (1b) and (2b) give rise to an identification result for the net encouragement effect in the whole population:

Corollary 4. *If either assumption* (6b) *holds for a value of* $\tilde{a} = 0$ *and both* $a = 0$ *and* $a = 1$ *and* $\forall m \in 0,1$*, or assumption* (7b) *holds for a value of* $\tilde{a} = 0$ *and* $\forall m \in 0,1$*, the population net encouragement effect* $NEE^0(\mathbf{c})$*, within levels of covariates, is given by:*

$$NEE^0(\mathbf{c}) = \sum_{(m_0,m_1)} NEE^0(m_0,m_1,\mathbf{c})\pi_{m_0m_1}(\mathbf{c}) = \sum_{m=0}^{1}\left(DCE(m,\mathbf{c})\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})\right) \quad \text{(A 1.4)}$$

If either assumption 6b holds for a value of $\tilde{a} = 1$ *and both* $a = 0$ *and* $a = 1$ *and* $\forall m \in 0,1$*, or assumption 7b holds for a value of* $\tilde{a} = 1$ *and* $\forall m \in 0,1$*, the population net encouragement effect* $NEE^1(\mathbf{c})$*, within levels of covariates, is given by:*

$$NEE^1(\mathbf{c}) = \sum_{(m_0,m_1)} NEE^1(m_0,m_1,\mathbf{c})\pi_{m_0m_1}(\mathbf{c}) = \sum_{m=0}^{1}\left(DCE(m,\mathbf{c})\sum_{m_0=m}^{1-m}\pi_{m_0m}(\mathbf{c})\right) \quad \text{(A 1.5)}$$

Proof. The proof of the corollary simply follows from equation (A 1.1) applied for the specified values of $a, \tilde{a}$ and m and from theorem 2b, by performing a weighted average over all four principal strata.

□

Both assumptions (6b) and (7b) provide the possibility of a generalization of the information on one potential outcome or the net encouragement effect from a stratum S^{mm} to the stratum $S^{m_0m_1}$ with $M_{ij}(\tilde{a}) = m$ and $M_{ij}(1-\tilde{a}) \neq m$, as stated

by theorems (1b) and (2b). As a fair consequence of this generalization, the estimation of the individual treatment mediated effect for strata with $M_{ij}(0) \neq M_{ij}(1)$ in this stratum is straightforward and given by the difference between the estimated principal causal effect and net encouragement effect: $iTME^{1-\tilde{a}}(m_0, m_1, \mathbf{c}) = PCE(m_0, m_1, \mathbf{c}) - NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$.

Corollary 5. *If either assumption* (6b) *holds for a specific value of* $\tilde{a}$, $\forall m \in 0,1$ *and both* $a = 0$ *and* $a = 1$ *or assumption* (7b) *holds for a specific value of* $\tilde{a}$ *and* $\forall m \in 0,1$, *the individual treatment mediated effect in the whole population is given by the weighted sum over the compliers and the defiers, as reported in* (1.4.10).

$$iTME^{1-\tilde{a}}(\mathbf{c}) = \sum_{m_0 \neq m_1} \Big(PCE(m_0, m_1, \mathbf{c}) - DCE(m_{\tilde{a}}, \mathbf{c}) \Big) \pi_{m_0 m_1}(\mathbf{c}) \qquad \text{(A 1.6)}$$

Note that when the defiers are not present the $iTME^{1-\tilde{a}}(m_0, m_1, \mathbf{c})$ will just be scaled by the conditional probability of compliers.

A 2 Imputation Approach For Net Encouragement Effects and Individual Mediated Treatment Effects

Here we generalize the Bayesian imputation approach, described in section 1.7.2, for the estimation of net encouragement effects and individual mediated treatment effects. The first two steps for the estimation of principal causal effects remain unchanged, whereas at each iteration draws from the posterior distribution of causal mechanisms for compliers and defiers, under assumptions (6b) or (7b) with $\tilde{a} \in \{0,1\}$, are obtained with four different steps:

3. For each unit with $M_{ij}(0) = m_0 \neq M_{ij}(1) = m_1$, i.e compliers and defiers, at iteration k, the potential outcome $Y_{ij}^{k}\big(\tilde{a}, M_{ij}(\tilde{a})\big) = Y_{ij}^{k}(\tilde{a})$ is derived as follows: if assumption (6b) holds, $Y_{ij}^{k}(\tilde{a})$ is simply taken from Y_{ij}^{obs} or Y_{ij}^{mis}, depending on A_j^{obs}; if assumption (7b) holds, in order to follow the identification result in theorem 2b, $Y_{ij}^{k}(\tilde{a})$ is imputed from the likelihood distribution of $Y_{ij}(\tilde{a})$ for

principal strata $S^{m_{\tilde{a}}m_{\tilde{a}}}$ with $M_{ij}(0) = M_{ij}(1) = m_{\tilde{a}}$, i.e. never-takers or alway-takers, given his values of covariates $\mathbf{C}_{ij}$:

$$Y_{ij}^{\,k}(0): \begin{cases} \text{3a. } \textit{if assumption } 6\text{b: } Y_{ij}^{\,k}(\tilde{a}) = Y_{ij}^{\,obs} \cdot (1 - A_j^{obs}) + Y_{ij}^{\,k,mis} \cdot A_j^{obs} \\ \text{3b. } \textit{if assumption } 7\text{b: } Y_{ij}^{\,k}(\tilde{a}) \sim f_{m_{\tilde{a}}m_{\tilde{a}}}(\tilde{a} \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k) \end{cases}$$
$$\forall i, j : S_{ij}^k = S^{m_0 m_1}$$

4. For each unit with $M_{ij}(0) = m_0 \neq M_{ij}(1) = m_1$, i.e compliers and defiers, at iteration k , $Y_{ij}^{\,k}\big(1-\tilde{a}, M_{ij}(\tilde{a})\big)$ is imputed from the likelihood distribution of $Y_{ij}(1-\tilde{a})$ for principal strata $S^{m_{\tilde{a}}m_{\tilde{a}}}$, with $M_{ij}(0) = M_{ij}(1) = m_{\tilde{a}}$, i.e. never-takers or alway-takers, given his values of covariates $\mathbf{C}_{ij}$:

$$Y_{ij}^{\,k}\big(1-\tilde{a}, M_{ij}(\tilde{a})\big) \sim f_{m_{\tilde{a}}m_{\tilde{a}}}(1 \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k) \qquad \forall i, j : S_{ij}^k = S^{m_0 m_1}$$

5. $NEE^{k,\tilde{a}}$ for compliers and defiers is computed by taking the average, within levels of covariates, of the difference between the two imputed potential outcomes:

$$\widehat{NEE}^{k,\tilde{a}}(m_0, m_1, \mathbf{c}) = \frac{1}{|\mathscr{S}_c^{m_0 m_1}|} \sum_{i,j:S_{ij}^k = \mathscr{S}_c^{m_0 m_1}} \big(Y_{ij}^{\,k}\big(1, M_{ij}(\tilde{a})\big) - Y_{ij}^{\,k}\big(0,, M_{ij}(\tilde{a})\big)\big)$$

Again subgroup analysis based on covariates might require some restrictions.

Estimation of individual treatment effects requires a last step, that is subtracting the estimated net encouragement effects from the principal causal effects for compliers and defiers:

6. $$i\widehat{TME}^{k,1}(m_0, m_1, \mathbf{c}) = \widehat{PCE}^{k}(m_0 m_1, \mathbf{c}) - \widehat{NEE}^{k,0}(m_0, m_1, \mathbf{c})$$

These steps, for either assumption, are carried out repeatedly to account for the uncertainty in the imputation, resulting in the posterior distribution of the causal estimands. Finally, a summary statistics of these distributions, such as the mean or

the median, can provide us with point estimates.

A 3 Controlled net encouragement effects within principal strata

We define the *Controlled Net Encouragement Effect* (CNEE) within principal stratum $S^{m_0 m_1}$ and level of covariates $\mathbf{C}_{ij} = \mathbf{c}$, as follows:

$$CNEE^{m}(m_0, m_1 \mathbf{c}) := \mathrm{E}\big[Y_{ij}(1,m) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0,m) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{A 3.1}$$

From the definition of net encouragement effects within principal strata it follows that net encouragement effects $NEE^{a}(m_0, m_1)$ for the stratum where $M(0) = m_0$ is equal to the controlled net encouragement effects for that strata with treatment receipt fixed at m_0:

$$NEE^{0}(m_0, m_1, \mathbf{c}) \equiv CNEE^{m_0}(m_0, m_1, \mathbf{c})$$

and, analogously, the net encouragement effect $NEE^{1}(m_0, m_1)$ for the strata where $M(1) = m_1$ is equal to the controlled net encouragement effects with treatment receipt fixed at m_1:

$$NEE^{1}(m_0, m_1, \mathbf{c}) \equiv CNEE^{m_1}(m_0, m_1, \mathbf{c})$$

Proof. The proof is straightforward and follows from the definition of NEE by noticing that within strata potential intermediate variables are constant and their value can be replaced in potential outcomes:

$$\begin{aligned} NEE^{a}(m_0, m_1, \mathbf{c}) &= \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(a)\big) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}\big(0, M_{ij}(a)\big) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \\ &= \mathrm{E}\big[Y_{ij}(1, m_a) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0, m_a) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \\ &= CNEE^{m_a}(m_0, m_1, \mathbf{c}) \end{aligned}$$

□

By virtue of this equivalence, theorem (2b) can also be expressed in terms of *CNEE*.

Corollary 6. *If either assumption* (6b) *hold for a specific value of* $\tilde{a} \in \{0,1\}$*, both* $a = 0$ *and* $a = 1$ *and a specific value of* $m \in \{0,1\}$*, or assumption* (7b) *holds for specific values of* $\tilde{a} \in \{0,1\}$ *and* $m \in \{0,1\}$*, then the controlled net encouragement effect, within level of covariates, for the stratum* $S^{m_0 m_1}$ *where* $M_{ij}(\tilde{a}) = m_{\tilde{a}} = m$ *and* $M_{ij}(1-\tilde{a}) = m_{1-\tilde{a}} \neq m$*, setting the treatment receipt to* $m_{\tilde{a}}$*, is equal to the corresponding controlled net encouragement effect for the stratum* $S^{m_{\tilde{a}} m_{\tilde{a}}}$ *where both* $Mij(\tilde{a}) = M_{ij}(1-\tilde{a}) = m_{\tilde{a}} = m$.

$$CNEE^{m_{\tilde{a}}}(m_0, m_1, \mathbf{c}) \equiv CNEE^{m_{\tilde{a}}}(m_{\tilde{a}}, m_{\tilde{a}}, \mathbf{c})$$

As a final result we can claim that, if assumptions (6b) or (7b) are satisfied for both encouragement conditions, $\tilde{a} = 0$ and $\tilde{a} = 1$, the controlled net encouragement effect $CNEE^{m}(m_0, m_1 \mathbf{c})$ is the same for all the strata with at least one of the potential values $M_{ij}(0)$ or $M_{ij}(1)$ equal to m.

A 4 Computation of the Posterior Distribution: Gibbs-Sampling and Data Augmentation

As stated earlier, the Bayesian inference in a Principal Stratification framework is based on the joint posterior distribution of $(\boldsymbol{\theta}, \mathbf{S})$, since the vector of individual principal strata $\mathbf{S}$ is not observed. Moreover, according to the proposed multinomial probit model for the strata membership, the two latent variables S_{ij}^{n} and S_{ij}^{c} have to be included as unknown variables. An approximation of this joint posterior distribution can be performed with a Gibbs-sampling approach. At every iteration of the Markov chain each set of parameters, the strata indicators S_{ij} and the latent variables S_{ij}^{n} and S_{ij}^{c} are drawn in turns from their full conditional distributions. At the end of the chain, given the sequence of samples drawn at each iteration, we can

obtain the histogram of the marginal posterior distributions of each parameter.

In the following we will describe each step of the Gibbs sampler used in both chapters. Let $\boldsymbol{\theta}^{(0)}$, $\mathbf{S}^{(0)}$, $\mathbf{S}^{n(0)}$ and $\mathbf{S}^{c(0)}$ be the vectors of starting values of the parameters, the strata indicators and the strata latent variables. At each iteration of the Monte Carlo Markov chain the sampling procedure is as follows.

The first part of the algorithm concerns the imputation of potential outcomes and hence of causal estimands from their posterior predictive distributions. Imputation of missing potential outcomes and principal causal effects within each individual principal stratum is described in the first two steps of the algorithm in both sections 1.7.2 and 2.7.2. Estimation of net encouragement effects and individual treatment effects follows the procedure outlined in section A 2 under assumption (6b) (7b). Similarly, spillover mediated effects, pure encouragement effects and individual mediated effects of chapter 2, are imputed as described in section 2.7.2, under assumptions (9) and (10).

1. The missing outcome $Y_{ij}^{mis} = Y_{ij}1 - A_j$ for each unit is drawn from the conditional distribution $f_{m_0 m_1}(1 - A_j \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^k)$, as defined by the models (1.6.1) and (2.6.2). In addition, for the estimation of causal mechanisms of chapter 1, for each complier, i.e. with strata indicator $S_{ij} = S^{01}$, we draw two random samples, $Y_{ij}^{k}(\tilde{a})$ and $Y_{ij}^{k}(1-\tilde{a}, M_{ij}(\tilde{a}))$, as described in section 1.7.2. Finally, $PCE(m_0, m_1, \mathbf{c})$ and $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ for all three individual principal strata and $iTME^{1-\tilde{a}}(0, 1, \mathbf{c})$ for compliers are derived. Similarly, causal mechanisms of chapter 2, for MN-invariant and M-invariant superstata, are derived as described in section 2.7.2.

2. Chapter 1: The vector of parameters $\boldsymbol{\beta}$ of the outcome model is drawn from its full conditional distribution $p\left(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{S}, \mathbf{Z}^{Yf}, \mathbf{Z}^{Yr}, \boldsymbol{b}\right)$. In the application of bed nets used in chapter 1, since the outcomes follows a binomial distribution, this is accomplished by a random walk Metropolis-Hastings algorithm with a normal proposal distribution, whose covariance matrix is a scaled version of an initial estimate.

Chapter 2: Given the conjugacy property of the normal prior with the normal outcome model used in the simulation study of chapter 2, the full conditional distribution has a closed form. The vector of parameters here is $\boldsymbol{\beta}^{fr}$, where the coefficient of the random part is included, and it is drawn from a different full conditional distribution $p\left(\boldsymbol{\beta}^{fr} \mid \mathbf{Y}, \mathbf{S}, \mathbf{Z}^{Yfr}, \boldsymbol{\sigma}_{\epsilon}^{2}\right)$.

3. Chapter 1: Cluster-specific $\boldsymbol{b}_j$ are drawn independently for each cluster from their posterior distribution $p\left(\boldsymbol{b}_j \mid \boldsymbol{\beta}, \mathbf{Y}_j, \mathbf{Z}_j^{Yr}, \mathbf{Z}_j^{Yf}\right)$. Another step of random walk Metropolis-Hastings is used for the purpose, with a normal proposal distribution, a likelihood derived from the binomial regression model in (1.6.1) and (1.6.2) and a normal prior distribution given in (1.7.7), where the prior covariance matrix $\boldsymbol{\Sigma}_b$ is drawn at the previous iteration from its own posterior distribution.
   Chapter 2: This step is not included in the computation.

4. Chapter 1: This step is not included in the computation.
   Chapter 2: Each variance $\sigma_{\epsilon}^{2 S^{m_0 m_1}}$ is independently drawn from its full conditional distribution $p(\sigma_{\epsilon}^{2 S^{m_0 m_1}} \mid \boldsymbol{\beta}^{fr}, \{Y_{ij}, \mathbf{Z}_{ij}^{Yfr} : S_{ij} = S^{m_0 m_1}\})$.

5. Chapter 1: The drawing of the covariance matrix Σ_b of the random effects is from the Inverse-Wishart posterior distribution, derived as the posterior distribution of a covariance matrix of multivariate normal random variable, $\boldsymbol{b}_j$ in this case, with Inverse-Wishart prior as defined in (1.7.8).
   Chapter 2: This step is not included in the computation.

This second part of the algorithm concerns the principal strata model.

6. Chapter 1: The vectors of parameters $\boldsymbol{\alpha}_n$ and $\boldsymbol{\alpha}_c$ of the individual strata membership model are drawn independently from their normal full conditional distributions
   $p\left(\boldsymbol{\alpha}_n \mid \mathbf{S}^n, \mathbf{Z}^{Sf}, \mathbf{Z}^{Sr}, \mathbf{a}_n\right)$ and $p\left(\boldsymbol{\alpha}_c \mid \mathbf{S}^c, \mathbf{Z}^{Sf}, \mathbf{Z}^{Sr}, \mathbf{a}_c\right)$ computed from their likelihood resulting from the linear models of the latent variables S_{ij}^n and S_{ij}^c in (1.6.5), and their prior distributions in (1.7.9). This time bayesian regression are run with

offsets $\mathbf{a}_{nj}^T Z_{ij}^{Sr}$ and $\mathbf{a}_{cj}^T Z_{ij}^{Sr}$ respectively.

Chapter 2: The vectors of parameters here are $\boldsymbol{\alpha}_n^{fr}$ and $\boldsymbol{\alpha}_c^{fr}$, where the coefficients of the random part are included, and they are drawn from different full conditional distributions $p\left(\boldsymbol{\alpha}_n^{fr} \mid \mathbf{S}^n, \mathbf{Z}^{Sfr}\right)$ and $p\left(\boldsymbol{\alpha}_c^{fr} \mid \mathbf{S}^c, \mathbf{Z}^{Sfr}\right)$.

7. Chapter 1: According to distributional assumptions presented above, cluster-specific random effects $\mathbf{a}_{nj}$ and $\mathbf{a}_{cj}$ are drawn independently for each cluster from their normal posterior distributions $p\big(\mathbf{a}_{nj} \mid \mathbf{S}_j^n, \mathbf{Z}_j^{Sr}, \mathbf{Z}_j^{Sf}, \boldsymbol{\alpha}_n\big)$ and $p\big(\mathbf{a}_{cj} \mid \mathbf{S}_j^c, \mathbf{Z}_j^{Sr}, \mathbf{Z}_j^{Sf}, \boldsymbol{\alpha}_c\big)$ derived from the linear regression model in (1.6.5), this time with offsets $\boldsymbol{\alpha}_n^T \mathbf{Z}_{ij}^{Sf}$ and $\boldsymbol{\alpha}_c^T \mathbf{Z}_{ij}^{Sf}$, and normal prior distribution given in (1.7.10), where the prior covariance matrices Σ_{a_n} and Σ_{a_c} come from the previous iteration.

   Chapter 2: This step is not included in the computation.

8. Chapter 2: As with outcome random effects, the drawing of the covariance matrices of the strata model random effects, Σ_{a_n} and Σ_{a_c} , is from the Inverse-Wishart posterior distributions $p(\Sigma_{a_n} \mid \mathbf{a}_n)$ and $p(\Sigma_{a_c} \mid \mathbf{a}_c)$, derived as the posterior distribution of a covariance matrix of multivariate normal random variable, in this case $\mathbf{a}_{nj}$ and $\mathbf{a}_{cj}$, with Inverse-Wishart prior as defined in (1.7.11).

   Chapter 2: This step is not included in the computation.

9. Chapter 1: This step is not included in the computation.

   Chapter 2: Cluster-specific $\boldsymbol{u}_j$ are drawn independently for each cluster from their normal posterior distribution $p\big(\boldsymbol{u}_j \mid Y_{ij}, \boldsymbol{\beta}^{fr}, \mathbf{Z}_{ij}^{Yr}, \mathbf{Z}_{ij}^{Yf}, \boldsymbol{\sigma}_\epsilon^2, \mathbf{S}^n, \boldsymbol{\alpha}_n^{fr}, \mathbf{S}^c, \boldsymbol{\alpha}_c^{fr}, \mathbf{Z}^{Sf}, \mathbf{Z}^{Sr},\big)$, which depends on the observed data, the coefficients of all three models, the error variances, and the latent variables representing the principal strata membership.

10. Given the fixed effects, the random effects and the observed data, the vector of latent individual strata membership $\mathbf{S}$ has to be generated from its full conditional distribution $p(\mathbf{S} \mid \mathbf{Y}, \mathbf{M}, \mathbf{A}, \mathbf{C}, \boldsymbol{\theta})$, which this time depends as well on the vector of individual treatment receipt $\mathbf{M}$ being the principal stratum defined based on the potential mediators. This is the typical data augmentation step of the principal

strata framework. As far as the average effects for each individual principal stratum are concerned, within each cluster the strata memberships of the unit are independent and hence strata indicators can be drawn independently from the conditional distribution factorized as:

$$
\begin{aligned}
&p(S_{ij} = S^{m_0 m_1} \mid Y_{ij}, M_{ij}, A_j, \mathbf{C}_{ij}, \boldsymbol{\theta}) \\
&= \frac{p\left(Y_{ij} \mid S_{ij} = S^{m_0 m_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m_0 m_1}}, \mathbf{b}_j^{S^{m_0 m_1}}\right) p(S_{ij} = S^{m_0 m_1} \mid M_{ij}, A_j, \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a})}{\sum_{S^{m'_0 m'_1}} p(Y_{ij} \mid S_{ij} = S^{m'_0 m'_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m'_0 m'_1}}, \mathbf{b}_j^{S^{m'_0 m'_1}}) p(S_{ij} = S^{m'_0 m'_1} \mid M_{ij}, A_j, \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a})} \\
&= \frac{p\left(Y_{ij} \mid S_{ij} = S^{m_0 m_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m_0 m_1}}, \mathbf{b}_j^{S^{m_0 m_1}}\right) p(S_{ij} = S^{m_0 m_1} \mid \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a}) I\left(M_{ij}(A_j) = M_{ij}\right)}{\sum_{S^{m'_0 m'_1}} p(Y_{ij} \mid S_{ij} = S^{m'_0 m'_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m'_0 m'_1}}, \mathbf{b}_j^{S^{m'_0 m'_1}}) p(S_{ij} = S^{m'_0 m'_1} \mid \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a}) I\left(M_{ij}(A_j) = M_{ij}\right)}
\end{aligned}
\tag{A 4.1}
$$

When monotonicity assumption holds, individuals with $A_j = 0$ and $M_{ij} = 1$ or $A_j = 1$ and $M_{ij} = 0$ are necessarily always-takers and never-takers respectively. Instead in the other situations two strata are possible fit, never takers or compliers when $A_j = 0$ and $M_{ij} = 0$ and always-takers or compliers when $A_j = 1$ and $M_{ij} = 1$. The drawing of one or the other possibility is made according to a bernoulli distribution with probability resulting from the conditional probabilities reported above.

11. For the estimation of causal mechanisms of chapter 2, the neighborhood principal strata have to be considered. The tow variables nS_{ij}^{01} and $nS_{ij}^{1-\tilde{a}1-\tilde{a}}$ are derived as a function of the vector of individual strata membership, as in eqation (2.3.3) in section 2.3

12. Each iteration ends with another data augmentation step resulting from the specific choice of the two linked probit models for S_{ij}. Precisely the latent variable S_{ij}^c and S_{ij}^c are drawn from their posterior distribution conditional on the strata indicators S_{ij} as they are updated at the previous step. These are normal linear models but truncated to the left or to the right depending on S_{ij}. In particular

lower and upper limits of the truncated normal distribution are:

$$S_{ij}^{n} \sim \begin{cases} N_{-}(\boldsymbol{\alpha}_n Z_{ij}^{Sf} + \mathbf{a}_{nj}^{T} Z_{ij}^{Sr}, 1) I(S_{ij}^{n} \leq 0) & \text{if } S_{ij} = S_{ij}^{00} \\ N_{+}(\boldsymbol{\alpha}_n Z_{ij}^{Sf} + \mathbf{a}_{nj}^{T} Z_{ij}^{Sr}, 1) I(S_{ij}^{n} > 0) & \text{if } S_{ij} = S_{ij}^{01} \text{ or } S_{ij} = S_{ij}^{11} \end{cases}$$

$$S_{ij}^{c} \sim \begin{cases} N(\boldsymbol{\alpha}_c Z_{ij}^{Sf} + \mathbf{a}_{cj}^{T} Z_{ij}^{Sr}, 1) & \text{if } S_{ij} = S_{ij}^{00} \\ N_{-}(\boldsymbol{\alpha}_c Z_{ij}^{Sf} + \mathbf{a}_{cj}^{T} Z_{ij}^{Sr}, 1) I(S_{ij}^{c} \leq 0) & \text{if } S_{ij} = S_{ij}^{01} \\ N_{+}(\boldsymbol{\alpha}_c Z_{ij}^{Sf} + \mathbf{a}_{cj}^{T} Z_{ij}^{Sr}, 1) I(S_{ij}^{c} > 0) & \text{if } S_{ij} = S_{ij}^{11} \end{cases} \tag{A 4.2}$$

A 5 Probability of Neighborhood Principal Strata membership

Let $\mathbf{M}_{ij} = [M_{ij}(0), M_{ij}(1)]$ be the vector of potential values of the treatment uptake under the two encouragement conditions of unit ij, so that $\mathbf{M}_{-ij}$ is the corresponding $2(N_j-1)$-dimensional vector of the unit's neighbors, i.e. $\mathbf{M}_{-ij} = [\mathbf{M}_{1j}, \ldots, \mathbf{M}_{i-1j}, \ldots, \mathbf{M}_{i+1j}, \ldots, \mathbf{M}_{N_j j}]$. Let also $\mathbf{N}_{ij}$ be the vector representing the neighborhood principal stratum, i.e. $\mathbf{N}_{ij} = [N_{ij}(0), N_{ij}(1)]$ and $\mathbf{n} = [n_0, n_1]$ its realization. Let finally $\mathbf{N}_{ij}^{\star}$ be a $N_j - 1$-dimensional variable, with realization denoted by , $\mathbf{n}^{\star}$, obtained by appending to $\mathbf{N}_{ij}$ $N_j - 3$ fictitious random variables. To compute the probability of a neighborhood principal stratum, given an individual principal stratum and baseline covariates, we will use method of transformation of multidimensional random variables with transformation $G_{ij}(\cdot)$ which is not injective.

$$P\left(nS_{ij} = nS^{n_0 n_1} \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right) = \sum_{l=1}^{L(\mathbf{n})} P\left(\mathbf{M}_{-ij} = \mathbf{m}^{l} \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\right) \times \frac{1}{|J(\mathbf{m}^{l})|} \tag{A 5.1}$$

where $\mathbf{m}^{l} = G_{ij}^{\star -1}(\mathbf{n}^{\star})$ real root since, with $G_{ij}^{\star}$ being a vector function such that $\mathbf{N}_{ij}^{\star} = G_{ij}^{\star}(\mathbf{M}_{-ij})$, and $|J(\mathbf{m}^{l})|$ is the jacobian determinant of the transformation. We will now refer to the hierarchical model (1.6.4) where individual principals strata of different units in the same cluster are assumed to be independent given $\mathbf{a}_j$, resulting

in the following factorization:

$$
\begin{aligned}
&= \sum_{l=1}^{L(\mathbf{n})} \sum_{\mathbf{a}_j} P\big(\mathbf{M}_{-ij} = \mathbf{m}^l \mid S_{ij} = S^{m_0 m_1}, \mathbf{a}_j, \mathbf{C}_{ij} = \mathbf{c}\big) p(\mathbf{a}_j \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}) \\
&= \sum_{l=1}^{L(\mathbf{n})} \sum_{\mathbf{a}_j} \prod_{k \in \mathcal{N}_i} P\big(\mathbf{M}_{kj} = [m_0^{kl}, m_1^{kl}] \mid \mathbf{C}_{ij} = \mathbf{c}, \mathbf{a}_j\big) p(\mathbf{a}_j \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}) \qquad \text{(A 5.2)} \\
&= \sum_{l=1}^{L(\mathbf{n})} \sum_{\mathbf{a}_j} \prod_{k \in \mathcal{N}_i} P\big(S_{kj} = S^{m_0^{kl} m_1^{kl}} \mid \mathbf{a}_j, \mathbf{C}_{ij},\big) p(\mathbf{a}_j \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c})
\end{aligned}
$$

where $[m_0^{kl}, m_1^{kl}]$ elements of the vector $\mathbf{m}^l$ corresponding to unit kj.

A 6 Results of Simulation Study

Scenario 1: Absence of Random Effects in the Outcome Model

Table 1: *Posterior Means and Confidence Intervals of Model Coefficients (Average of 500 simulations)*

	mean	median	sd	95% Interval
β_{00}: constant				
Never-takers	2.0984	2.0983	0.0331	[2.0336, 2.1632]
Always-takers	0.5955	0.5955	0.0413	[0.5145, 0.6765]
Compliers	2.0954	2.0954	0.0427	[2.0117, 2.1790]
β_{01}: education (X_1)	-0.9989	-0.9989	0.0387	[-1.0747,-0.9231]
β_{01}: distance form health facility (W)	2.0054	2.0053	0.0398	[1.9276, 2.0832]
β_{10}: encouragement				
Never-takers	-1.9326	-1.9309	0.2072	[-2.3738, -1.5626]
Always-takers	-0.8908	-0.2672	0.2489	[-1.4541,-0.479]
Compliers	-5.8340	-3.8894	0.3000	[-6.4824,-5.3077]
β_{20}: encouragement*complier neighbors				
Never-takers	-5.4047	-5.4188	0.6689	[-6.4561,-3.8352]
Always-takers	-2.5523	-2.5407	0.8237	[-3.7601, -0.5341]
Compliers	-5.6521	-5.6911	0.8545	[-6.0588, -2.7117]
β_{30}: encouragement*always-taker neighbors				
Never-takers	1.0471	1.0480	0.4730	[-0.0151, 1.8369]
Always-takers	0.5354	0.5354	0.5451	[-0.6376, 1.4968
Compliers	2.0914	2.0928	0.6885	[0.4325, 3.1287]
β_{40}: encouragement*complier neighbors *always-taker neighbors				
Never-takers	1.9319	1.9300	0.3480	[1.4125, 2.7754]
Always-takers	0.9687	0.9693	0.4172	[0.2495, 1.8830]
Compliers	1.8597	1.8563	0.4217	[1.3707, 3.0220]
variance σ^2_ϵ	1.0032	1.0030	0.0145	[0.9751, 1.0321]

Table 2: *Frequentist Performance of Bayesian Estimation Procedure for Model Coefficients (500 simulations)*

	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias % Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
β_{00}: constant								
Never-takers	95.0980	94.9020	-0.0020	-0.0020	-0.0780	-0.0790	0.0010	0.0010
Always-takers	96.0780	96.0780	-0.0040	-0.0040	-0.7440	-0.7470	0.0010	0.0010
Compliers	96.2750	96.2750	-0.0050	-0.0050	-0.2200	-0.2200	0.0020	0.0020
β_{01}: education (X_1)	96.4710	96.4710	0.0010	0.0010	-0.1100	-0.1090	0.0010	0.0010
β_{01}: distance form health facility (W)	94.5100	94.5100	0.0050	0.0050	0.2680	0.2670	0.0020	0.0020
β_{10}: encouragement								
Never-takers	92.9410	93.1370	0.0674	0.0691	-3.3710	-3.4530	0.0475	0.0477
Always-takers	96.6670	97.0590	0.1100	0.1092	-10.9980	-10.9180	0.0741	0.0739
Compliers	94.9020	94.7060	0.1636	0.1660	-2.7270	-2.7660	0.1168	0.1176
β_{20}: encouragement*complier neighbors								
Never-takers	92.9410	92.7450	-0.4047	-0.4188	8.0940	8.3750	0.6112	0.6228
Always-takers	95.8820	95.6860	-0.5523	-0.5407	27.6160	27.0350	0.9835	0.9708
Compliers	94.9020	94.9020	-1.6521	-1.6911	41.3020	42.2780	3.4596	3.59
β_{30}: encouragement*always-taker neighbors								
Never-takers	92.9410	92.9410	0.0471	0.0480	4.7080	4.8000	0.2259	0.2260
Always-takers	96.2750	96.2750	0.0354	0.0354	7.0880	7.0770	0.2984	0.2984
Compliers	95.2940	95.2940	0.0914	0.0928	4.5710	4.6390	0.4824	0.4826
β_{40}: encouragement*complier neighbors *always-taker neighbors								
Never-takers	93.1370	93.1370	-0.0681	-0.0700	-3.4030	-3.4990	0.1257	0.1260
Always-takers	95.2940	95.2940	-0.0313	-0.0307	-3.1290	-3.0660	0.1750	0.1750
Compliers	94.1180	94.3140	-0.1403	-0.1437	-7.0140	-7.1840	0.1975	0.1985
σ^2_ϵ	95.2940	94.7060	0.0030	0.0030	0.3190	0.3050	0.0000	0.0000

Table 3: *Estimated Parameters for Individual Principal Strata Model*

	$S^n Model$		$S^c Model$	
	Mean	95% Interval	Mean	95% Interval
Constant	2.3119	[2.0660, 2.5608]	1.5085	[-1.7123,-1.3081]
Education X_1	3.5274	[3.2965, 3.7623]	5.1298	[4.8573, 5.4112]
Sex X_2	-3.0175	[-3.2025,-2.8372]	-0.3059	[-0.4205,-0.1913]
Religion X_3	-3.5237	[-3.7184,-3.3335]	-1.0077	[-1.1717,-0.8436]
Family Presence X_4	-2.0146	[-2.1568,-1.8750]	1.0122	[0.8787, 1.1487]
Distance from Health facility W	0.5021	[0.2016, 0.8039]	-2.5066	[-2.8152,-2.2044]
Random Intercept Variance, σ_a^2	0.2719	[0.1934, 0.3702]	0.2667	[0.1922, 0.3587]

Table 4: *Frequentist Performance of Bayesian Estimation Procedure for Model Coefficients (500 simulations)*

	$S^n Model$							
	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias % Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
Constant	93.7250	93.9220	0.0120	0.0120	0.5180	0.5000	0.0170	0.0170
Education X_1	92.7450	92.3530	0.0270	0.0270	0.7820	0.7630	0.0160	0.0160
Sex X_2	93.9220	93.7250	-0.0170	-0.0170	0.5830	0.5610	0.0090	0.0090
Religion X_3	93.5290	93.5290	-0.0240	-0.0230	0.6770	0.6590	0.0110	0.0110
Family Presence X_4	93.9220	94.1180	-0.0150	-0.0140	0.7280	0.6990	0.0060	0.0060
Distance from Health facility W	92.5490	93.1370	0.0020	0.0020	0.4110	0.3020	0.0260	0.0260
Random Intercept Variance, σ_a^2	97.4510	94.7060	0.0220	0.0180	8.7420	7.3410	0.0020	0.0020

	$S^c Model$							
	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias % Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
Constant	94.9020	94.9020	-0.0080	-0.0080	0.5630	0.5130	0.0110	0.0110
Education X_1	93.9220	93.7250	0.0300	0.0280	0.5840	0.5530	0.0210	0.0210
Sex X_2	92.9410	93.1370	-0.0060	-0.0060	1.9620	1.9810	0.0040	0.0040
Religion X_3	94.9020	94.7060	-0.0080	-0.0080	0.7660	0.7570	0.0070	0.0070
Family Presence X_4	94.1180	94.1180	0.0120	0.0120	1.2200	1.1690	0.0050	0.0050
Distance from Health facility W	95.2940	95.4900	-0.0070	-0.0060	0.2650	0.2200	0.0230	0.0230
Random Intercept Variance, σ_a^2	97.2550	96.4710	0.0170	0.0140	6.6670	5.4280	0.0020	0.0020

Scenario 2: Presence of Random Effects in the Outcome Model

Table 5: *Posterior Means and Confidence Intervals of Model Coefficients (Average of 500 simulations)*

	mean	median	sd	95% Interval
β_{00}: constant				
Never-takers	2.1508	2.1507	0.0967	[1.9624,2.3404]
Always-takers	0.6366	0.6365	0.0970	[0.4472,0.8265]
Compliers	2.1615	2.1614	0.1016	[1.9637,2.3608]
β_{01}: education (X_1)	-1.0057	-1.0057	0.0353	[-1.0748,-0.9367]
β_{01}: distance form health facility (W)	1.8910	1.8913	0.1572	[1.5840,2.1966]
β_{10}: encouragement				
Never-takers	-1.7675	-1.7682	0.2715	[-2.2967,-1.2336]
Always-takers	-0.7847	-0.7850	0.3217	[-1.4141,-0.1533]
Compliers	-5.7118	-5.7125	0.3434	[-6.3822,-5.0373]
β_{20}: encouragement*complier neighbors				
Never-takers	-4.3341	-4.3316	0.7761	[-5.8598,-2.8207]
Always-takers	-1.3123	-1.3096	0.9633	[-3.2086,0.5662]
Compliers	-3.4102	-3.4056	0.9604	[-5.3045,-1.5424]
β_{30}: encouragement*always-taker neighbors				
Never-takers	-0.0855	-0.0853	0.6255	[-1.3117,1.1385]
Always-takers	-0.5532	-0.5563	0.7133	[-1.9420,0.853]
Compliers	0.6919	0.6886	0.8137	[-0.8926,2.2949]
β_{40}: encouragement*complier neighbors *always-taker neighbors				
Never-takers	1.7163	1.7158	0.3921	[0.9489,2.4843]
Always-takers	0.6846	0.6860	0.4753	[-0.2495,1.6124]
Compliers	1.7480	1.7487	0.4912	[0.7866,2.7088]
β^r: random intercept u	0.7795	0.7773	0.0661	[0.6566,0.9160]
variance σ^2_ϵ	0.9913	0.9912	0.0147	[0.9630,1.0204]

Table 6: *Frequentist Performance of Bayesian Estimation Procedure for Model Coefficients (500 simulations)*

	Coverage % (Normal)	Coverage % (quantiles)	Bias (Mean)	Bias % (Median)	Bias % (Mean)	Bias % (Median)	MSE (Mean)	MSE (Median)
β_{00}: constant								
Never-takers	93.4524	94.0476	0.0508	0.0507	2.4203	2.4129	0.0100	0.0100
Always-takers	95.2381	95.4365	0.0366	0.0365	6.0971	6.0819	0.0088	0.0088
Compliers	92.2619	92.4603	0.0615	0.0614	2.9307	2.9230	0.0122	0.0121
β_{01}: education (X_1)	93.0556	92.6587	-0.0057	-0.0057	0.5702	0.5708	0.0014	0.0014
β_{01}: distance form health facility (W)	90.8730	89.8810	-0.1090	-0.1087	-5.4485	-5.4363	0.0311	0.0311
β_{10}: encouragement								
Never-takers	85.9127	85.3175	0.2325	0.2318	-11.6265	-11.5892	0.1294	0.1289
Always-takers	92.4603	92.0635	0.2153	0.2150	-21.5286	-21.5044	0.1281	0.1280
Compliers	87.3016	86.9048	0.2882	0.2875	-4.8029	-4.7919	0.1892	0.1892
β_{20}: encouragement*complier neighbors								
Never-takers	86.9048	86.7063	0.6659	0.6684	-13.3180	-13.3671	0.9827	0.9861
Always-takers	93.4524	93.2540	0.6877	0.6904	-34.3873	-34.5193	1.0949	1.0996
Compliers	93.0556	93.4524	0.5898	0.5944	-14.7446	-14.8605	1.1313	1.1377
β_{30}: encouragement*always-taker neighbors								
Never-takers	67.8571	68.0556	-1.3081	-1.3114	-65.4072	-65.5712	2.2533	2.2616
Always-takers	59.3254	59.3254	-1.0855	-1.0853	-108.5462	-108.5286	1.5430	1.5423
Compliers	70.8333	71.0317	-1.0532	-1.0563	-210.6360	-211.2652	1.5099	1.5158
β_{40}: encouragement*complier neighbors *always-taker neighbors								
Never-takers	89.2857	89.0873	-0.2837	-0.2842	-14.1868	-14.2089	0.2130	0.2131
Always-takers	92.8571	92.2619	-0.3154	-0.3140	-31.5373	-31.4026	0.2636	0.2625
Compliers	92.4603	92.4603	-0.2520	-0.2513	-12.5984	-12.5657	0.2817	0.2816
β^r: random intercept u	0.0000	0.0000	0.2795	0.2773	55.9084	55.4555	0.4444	0.4413
variance σ^2_ϵ	90.0794	90.2778	-0.0087	-0.0088	-0.8693	-0.8840	0.0003	0.0003

Table 7: *Estimated Parameters for Individual Principal Strata Model*

	S^n *Model*		S^c *Model*	
	Mean	95% Interval	Mean	95% Interval
Constant	2.3019	[2.0031, 2.6154]	-1.5119	[-1.8077,-1.2612]
Education X_1	3.5131	[3.3425, 3.6845]	5.1336	[4.9280,5.2430]
Sex X_2	-3.0211	[-3.2912,-2.7681]	-0.3002	[-0.1990,-0.4020]
Religion X_3	-3.5359	[-4.7123,-2.3714]	-1.0187	[-1.1102,0.9278]
Family Presence X_4	-2.0301	[-2.3327,-1.7419]	1.0189	[0.9334,1.1667]
Distance from Health facility W	0.5395	[0.2378, 0.8442]	-2.5066	[-2.8152,-2.2044]
Random Intercept u	0.6479	[0.5184, 0.7899]	0.7176	[0.5824, 0.8665]

Table 8: *Frequentist Performance of Bayesian Estimation Procedure for Model Coefficients (500 simulations)*

	S^n *Model*							
	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias % Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
Constant	93.8492	94.0476	0.0019	-0.0010	0.1063	-0.0531	0.0296	0.0297
Education X_1	95.2381	95.2381	0.0131	0.0130	1.8766	1.8607	0.0077	0.0077
Sex X_2	93.2540	92.0635	-0.0211	-0.0178	0.7033	0.5943	0.0243	0.0241
Religion X_3	94.0476	93.4524	-0.0359	-0.0337	1.0257	0.9746	0.0360	0.0361
Family Presence X_4	94.0476	93.8492	-0.0301	-0.0275	1.5050	1.3885	0.0293	0.0294
Distance from Health facility W	95.4365	94.8413	0.0395	0.0387	1.975	1.8692	0.0233	0.0232
Random Intercept u	0.1984	0.1984	0.1479	0.1457	29.5710	29.1385	0.2714	0.2691

	S^c *Model*							
	Coverage % (Normal)	Coverage % (quantiles)	Bias Mean	Bias % Median	Bias % Mean	Bias % Median	MSE Mean	MSE Median
Constant	96.6270	96.2302	-0.0119	-0.0111	0.7933	0.7229	0.0121	0.0120
Education X_1	94.6429	94.0476	0.0336	0.0329	0.6720	0.5911	0.0109	0.0109
Sex X_2	95.6349	95.6349	-0.0002	-0.0000	0.0666	0.0317	0.0027	0.0027
Religion X_3	94.2460	93.6508	-0.0187	-0.0186	1.8705	1.8555	0.0034	0.0034
Family Presence X_4	92.6587	92.2619	0.0189	1.896	1.8569	1.1428	0.0083	0.0083
Distance from Health facility W	97.6190	97.0238	-0.0376	-0.0367	2.5052	2.4460	0.0217	0.0216
Random Intercept u	0.0000	0.0000	0.2176	0.2153	43.5233	43.0615	0.3472	0.3446

Acknowledgements

One of the best things when completing a PhD is to look over the past journey and realizing that it has been a wonderful experience. This is not just because of the topics I have learnt, or the experience of having my own projects, of writing papers and giving talks, but also, and more importantly, because of the unexpected and fascinating atmosphere around all this and the people that helped create this atmosphere, supported me and changed with me along this long but fulfilling road. This thesis would not have been possible without the inspiration and support of a number of wonderful individuals – my thanks and appreciation to all of them for being part and make possible this unforgettable journey.

First of all, I would like to take this opportunity to sincerely thank my advisor, Fabrizia Mealli, a talented teacher, a passionate scientist and an exceptional person. Having her as advisor has been such a natural decision. Her research topics seemed to be interesting and applicable to my interests, evaluation of health and social programs, and, above all, she was a 'though' and somewhat 'uncommon' woman. It was her. She picked me at a critical stage, where and I had just made an about-turn from my previous engineering studies and I was still confused. She supported me throughout my thesis with trust, immense knowledge and a non-systematic guidance, while allowing me the room to work in my own way. With her funding she gave me the opportunity to attend conferences, gain a wide breadth of experience and meet so many interesting people. The passion and enthusiasm she has for her research was contagious and motivational for me. It it weren't for her encouragement and contagious passion I probably would not have considered a career in research and in April I would have been lost somewhere in Africa or India. I admire her ability to balance research interests and personal pursuits, and also her braveness in always being ready to jump in a new adventure, without constrains. With her example she has exceptionally inspired and enriched my growth as a researcher and as a woman.

I could say she has become more of a mentor and friend than a professor.

This thesis would also not have been possible without the encouragement and support of my co-advisor, Alessandra Mattei, whose mathematical insight is impressive. I thank her for providing an experienced ear for my doubts, for offering thorough suggestions and comments on works in progress and for kindly warning me "I might be wrong but?", say, "I think 2 times 2 is not 3 but 4?but again I have to check?". Coming from another field, at the beginning it was though for me to keep up with even basic concepts, but Fabrizia's and Alessandra's encouragement and passion pushed me to study hard.

I am indebted and thankful for the fresh opportunity given to me by Tyler VanderWeele, Harvard School of Public Health, who was the source of inspiration for me in my early days. His advise and bright thoughts were very fruitful for shaping up my ideas and research.

My second unofficial mentor has been another 'though' and 'uncommon' professor, Donal Rubin, Harvard University. I am extremely grateful to him for his unwavering support and encouragement, for our long scientific and life discussions, for sharing his wisdom with me, for making sure I was having a good time in Boston, for taking care of me and for treating me like a daughter and a friend. I learned a lot from him about statistics, research and life. His technical excellence and tremendous grasp of experimental issues, his involvement and his originality has triggered and nourished my motivation to continue down this path. I expand my thanks to Don's grad students. They welcomed me with a stimulating and pleasant environment, and made me feel like part of their group by showing an unexpected interest in my work and helping me valuable suggestions. Appreciation also goes out to Corwin Zigler, Harvard School of Public Health, who made sure I had a desk to work (he still did not know I could work anywhere) and gave me encouraging and constructive feedback.

The application results of the first part of the thesis would not have been possible without Gunther Fink, Harvard School of Public Health, who provided me with the

dataset and supported me in the early stage. Thanks to Gunther Fink, Jessica Cohen, Jennifer Leaning, Richard Cash and other professors of the Department of Global Health and Population, I have bumped into this enchanting world of global health and I have grasped how academic research can contribute to generate knowledge and solve problems concerning health equity and human rights.

My first experience abroad as a visiting student researcher has been at the University of California, Berkeley. I gratefully acknowledge Jennifer Ahern for welcoming me and Hannah Leslie for letting me collaborate with her and for looking at me as if I were this experienced statistician.

The simulation study of the second part of the thesis benefited from my benefactors Simone Bucci and Matteo Morelli, who allowed me to use the computer cluster and facilities at Agenzia Regionale per la Protezione Ambientale (ARPA).

I also would like to record my gratitude to the two NGOs where I had the opportunity to take on an internship during the time of my PhD studies: Techo, Haiti; and Fundación François-Xavier Bagnoud, Colombia. These two experiences gave me the chance to meet great people with an immense faith in their projects and whose trust in my work has been a tremendous incentive for me.

This list is definitely incomplete without acknowledging my family, my stronghold. In particular, words are short to express my deep sense of gratitude to my parents, who have always been there for me–although I keep them at least at arm's length– to give me unwavering support and to help me overcome the roadblocks that unavoidably crop up throughout my life and my studies. It has been bumpy at times, but their unconditional and persistent confidence in me has enhanced my ability to get through it all. I thank them for for letting me take the opportunities and experiences that have made me who I am, for taking the blows and giving me a chance to thrive, for helping me move my vast collections of "stuff" across places. Thank you for teaching me that my job in life was to learn, to be happy, and to know and understand myself. I am particularly appreciative to my dad, who that day by the beach helped me decide

to make that about-turn and selflessly encouraged me to explore new directions in life and seek my own destiny, and to my mom, who silently hides her worries when I am stumbling, does not ask for explanations of my peculiar and mysterious behavior and tries not to impose her will. This journey would not have been possible if not for them.

My sister Serena and my twin sister Anna also deserve my sincere expression of thanks: Serena, for her light heartedness and for helping soften the difficulties; Anna, for being my life companion and for making me feel understood. Anna, in my master thesis I promised you would make it and now, after these years I basically left you alone for pushing you through but also for fear, I finally see it coming, and that?s the best gift I could have ever received. I cannot imagine being the person I am today without you.

Collective and individual acknowledgments are also owed to my friends and roommates scattered around the world. My time in Boston was made enjoyable in large part due to my roommates Jeff, Brandon, Zena, Faye, James, Andrea and Sarah, who have been like a surrogate family and made me change my mind about the american culture. I will miss your support when you asked me "Is you code working?", without a clue of what I was doing, and when you tried to help me give sense to 'my story of malaria, mosquitos and farmers'. I am particularly thankful to Sarah, for all the fun we had, for having followed me when I was working in Colombia and, above all, for having the patience to get through my resistance to opening myself up and promising me you will look for me whenever I will disappear. In Boston I was also very lucky to have crossed paths with Linda Valeri and Gianluca Mazzarella, who have become precious friends and a source of great support.

Further special thanks go to my present roommates Marco, Francesca, Carmine and Anna, for being awake with me during my sleepless nights of work and for the faithful support through the ups and downs during the final stretch of the thesis write-up (even if somehow they thought until the very last moment I was studying

anthropology).

Although he probably won't read these notes, I would like to send my thanks out to Matthias because at that time he helped me make up my mind on my future and selflessly gave me the freedom to take my own way forward and sail.

Of course during this long journey I was extraordinarily fortunate in having as a cornerstone my best friends Ilaria, Maggie, Claudia and Francesca. Each one in a different city of the world, each one with her own freakishness and daftness. These individuals accept me for what I am, probably without fully understanding me, they help me going through challenges of life, without asking, they never fail to lift my spirits (when I let them). I will never leave you (even if I flee to Brazil jeje). A journey is easier when you travel together, even from miles away. Thank you for banding together over beers, chats and life. Without you I would be lost. Thank you for just being there.

A very special thanks goes also out to to the many friends in the various households that have sheltered me over the years and to all the other friends who have shared part of themselves with me. You have always been patiently waiting for me to come back and reappear in your lives–Tiziana one big hug is for you and your little Leila–.

Finally to the countless cafes, bars, libraries and weird spots that have hosted me for the writing-up phase (Italy lacks 24/7 libraries and wifi cafes).

Last, but certainly not least, I take this opportunity to sincerely acknowledge my grandfather Giuseppe Scarinci for his constant encouragement and support, for our special relationship, for being worried about me, for asking everyday to my sister "Notizie di Laura?", for his admiration and trust, for accepting everything I did he would not agree with, for his groaning of concern and helplessness, for sheltering me in his mansard and for being the one I would always come back to. Although he has never really liked the idea that I was not going to design artificial hearts and he would say "Do you need a PhD in statistics to count the unemployed in Italy?", it is to him that I dedicate this work.

Bibliography

Abadie, A. *&* Imbens,G. W. *(2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. Econometrica, 74(1), 235–267.*

Albert, J. *(2008). Mediation analysis via potential outcomes models. Statistics in Medicine, 27, 1282–1304.*

Alonso, P. L., Lindsay, S. W., Armstrong Schellenberg, J. R., Keita, K., Gomez, P., Shenton, F. C., Hill, A. G., David, P. H., Fegan, G. *&* Cham, K. *(1993). A malaria control trial using insecticide-treated bed nets and targeted chemoprophylaxis in a rural area of The Gambia, west Africa. 6. The impact of the interventions on mortality and morbidity from malaria. Transactions of the Royal Society of Tropical Medicine and Hygiene, 87, 37–44.*

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

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

Baird, S. J., Garfein, R. S., McIntosh, C. T., *&* Özler, B. *(2012). Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial. Lancet, , 379, 1320–1329.*

Banarjee, A., Duflo, E., Glennerster, R., *&* Kothari, D. *(2010). Improving Immunization Coverage in Rural India: A Clustered Randomized Controlled Evaluation of Immunization Campaigns with and without Incentives. British Med-*

ical Journal, 340:c2220.

Barnard, J., Frangakis, C., Hill, J., & Rubin, D. B. *(2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City (with discussion). Journal of the American Statistical Association, 98, 299–323.*

Baron, R. M. & Kenny, D. A. *(1986). The moderator?mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173–1182.*

Binka, F. N., Indome, F. & Smith, T. *(1998). Impact of spatial distribution of permethrin-impregnated bed nets on child mortality in rural northern Ghana. American Journal of Tropical Medicine and Hygiene, 59(1), 80–5.*

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

Bowers, J. & Hansen, B. *(2005). Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference. Working Paper.*

Baron, E. T. *(1998). Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design. Marketing Letters, 9(4), 383–391.*

Christakis, N. A. & Iwashyna, T. I. *(2003). The Health Impact of Health Care on Families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. Social Science & Medicine, 57 (3), 465–475.*

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

D'Alessandro, U., Olaleye, B. O., McGuire, W., Langerock, P., Bennett, S., Aikins, M. K., Thomson, M. C., Cham, M. K., Cham, B. A. & Greenwood, B. M. *(1995). Mortality and morbidity from malaria in Gambian children after introduction of an impregnated bednet programme. Lancet, 345, 479–83.*

de Finetti, B. *(1974). Theory of Probability: A Critical Introductory Treatment. Wiley; New York.*

Diaconis, P. *(1977). Finite forms of deFinetti's theorem on exchangeability. Synthese, 36, 271–181.*

Duflo, E. *&* Saez, E. *(2002). Participation and investment decisions in a retirement plan: the influence of colleagues' choices. Journal of Public Economics, 85(1), 121–148.*

Dunn, G. *&* Bentall, R. *(2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). Statistics in Medicine, 26(26), 4719–4745.*

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, 353–372.*

Fink, G., Masiye, F. *(2012). Assessing the impact of scaling-up bednet coverage through agricultural loan programmes: evidence from a cluster randomised controlled trial in Katete, Zambia. Transactions of the Royal Society of Tropical Medicine and Hygiene,106, 660–667.*

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

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

Flores, C. A. *&* Flores-Lagunes, A. *(2009a). Identication and estimation of causal mechanisms and net effects of a treatment. IZA Discussion Paper, No. 4237.*

Flores, C. A. *&* Flores-Lagunes, A. *(2009b). Nonparametric partial and point identification of net or direct causal effects. American Economic Association, Annual Meeting Paper 2009.*

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

Frangakis, C. E., Rubin, D. B. & Zhou, X. H. *(2002). Clustered encouragement design with individual noncompliance: Bayesian inference and application to Advance Directive Forms (with discussion). Biostatistics, 3, 147–164.*

Galiani, S. & McEwan, P. J. *(2013). The heterogeneous impact of conditional cash transfers. Journal of Public Economics,103, 85–96.*

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffee, M., & Ten Have, T. R. *(2009). Mediation analysis with principal stratification.. Statistics In Medicine, 28, 1108–1130.*

Gelman, A. *(1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), Markov Chain Monte Carlo in practice, Boca Raton, FL:Chapman and Hall, (pp. 131–143).*

Gelman, A. & Rubin, D. B. *(1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7, 457–472.*

Giné, X. & Mansuri, G. *(2011). Together we will : experimental evidence on female voting behavior in Pakistan. Policy Research Working Paper Series 5692, The World Bank.*

Hafeman, D. M. & VanderWeele, T. J. *(2011). Alternative assumptions for the identification of direct and indirect effects. Epidemiology, 22, 753–764.*

Hawley, W. A., Phillips-Howard, P. A., ter Kuile, F. O., Terlouw, D. J., Vulule, J. M., Ombok, M., Nahlen B. L., Gimnig, J. E., Kariuki, S. K., Kolczak, M. S. & Hightower, A. W. *(2003). Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. American Journal of Tropical Medicine Hygiene, 68, 121–27.*

Heckman, J. & Hotz, V. J. *(1989). Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Man-*

power Training (with discussion). Journal of the American Statistical Association, 84(408), 862–880.

Hill, J., Waldfogel, J. *&* Brooks-Gunn, J. *(2002). Assessing differential impacts: The effects of high-quality child care on children's cognitive development. Journal of Policy Analysis and Management, 21, 601–628.*

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, 69–88.*

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

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, 901–910.*

Howard, W. A., Omumbo, J., Nevill, C., Some, E. S., Donnelly C. A. *&*Snow, R. W. *(2000). Evidence for a mass community effect of insecticide-treated bednets on the incidence of malaria on the Kenyan coast. Transactions of the Royal Society of Tropical Medicine and Hygiene, 94, 357–60.*

Hudgens, M. G., *&* Halloran, M. E. *(2006). Causal vaccine effects on binary post-infection outcomes. Journal of the American Statistical Association, 101, 51–64.*

Hudgens, M. G., *&* Halloran, M. E. *(2008). Towards causal inference with interference. Journal of the American Statistical Association, 103, 832–842.*

Hudgens, M. G., *&* Halloran, M. E. *(2012). Comparing Bounds for Vaccine Effects on Infectiousness. Epidemiology, 23(6), 931–932.*

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

Imbens, G. W., *&* Rubin, D. B. *(1997). Bayesian inference for causal effects in*

randomized experiments with noncompliance. Annals of Statistics, 25, 305–327.

Imai, K. *(2005). Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments. American Political Science Review, 99 (2),283?300.*

Imai, K., Keele, L., & Tingley, D. *(2010a). A General Approach to Causal Mediation Analysis. Psychological Methods, 15, 309–334.*

Imai, K., Keele, L., & Yamamoto, T. *(2010b). Identification, inference, and sensitivity analysis for causal mediation effects. Statistical Science, 25, 51–71.*

Jo, B. *(2008). Causal inference in randomized experiments with mediational processes. Psychological Methods, 13, 314–336.*

Jo, B., Asparouhov, T., Muthén B. O., Ialongo N. S. & Brown C. H. *(2008a). Cluster randomized trials with treatment noncompliance. Psychological Methods, 13, 1–18.*

Jo, B., Asparouhov, T. & Muthén B. O. *(2008b). Intention-to-Treat Analysis in Cluster Randomized Trials with Noncompliance. Statistics in Medicine, 27, 5565–5577.*

Lynch, K. G., Cary, M., Gallop, R. & Ten Have, T. R. *(2008). Causal mediation analysis for randomized trial. Health. Serv. Outcome Res. Meth., 8, 57–76.*

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. & Sheets, V. *(2002). A comparison of methods to test mediation and other intervening variable effects. Psychological Methods, 7, 83–104.*

Maluccio, J. A., Murphy, A. & Regalia, F. *(2011). Does Supply Matter? Initial School Conditions and the Effectiveness of Conditional Cash Transfers for Grade Progression in Nicargaua. Journal of Development Effectiveness, 2(1), 87–116.*

Manski, C. F. *(1990). Non-parametric bounds on treatment effects. American*

Economic Review, Papers and Proceedings, 80, 319–323.

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

Manski, C. F. *(2013). Identification of treatment response with social interactions. The Econometrics Journal, 16(1):S1–S23.*

Mattei, A. & Mealli, F. *(2011). Augmented designs to assess principal strata direct effects. Journal of the Royal Statistical Society, 73(5), 729–752.*

McDonald, C., Hiu, S. & Tierney, W. *(1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. MD Computing, 9, 304–312.*

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

Mealli, F. & Mattei, A. *& (2012). A refreshing account of principal stratification. The International Journal of Biostatistics, 8(1), 246–254.*

Morris, 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. Lancet, 364, 2030–2037.*

Neyman, J. *(1923). On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in Statistical Science, 5, 465–80, 1990.*

Nevill, C. G., Some, E. S., Mung'ala, V. O., Mutemi, W., New, L., Marsh, K., Lengeler, C. & Snow, R. W. *(1996). Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast.*

Tropical Medicine in International Health, 1(2), 139–46.

Page, L. C. *(2012). Principal Stratification as a Framework for Investigating Mediational Processes in Experimental Settings. Journal of Research on Educational Effectiveness, 5, 215–244.*

Pearl, J. *(2001). Direct and indirect effects. In Proc. 17th Conf. Uncertainty in Artificial Intelligence (eds. J. S. Breese & D. Koller), San Francisco: Morgan Kaufman, (pp. 411–420).*

Pearl, J. *(2011). Principal stratification – A goal or a tool?. International Journal of Biostatistics, 7(1), Article 20.*

Petersen, M., Sinisi, S. E. & Van Der Laan, M. *(2006). Estimation of direct causal effects. Epidemiology, 17, 276–284.*

Robins, J. M., & Greenland, S. *(1992). Identifiability and Exchangeability for Direct and Indirect Effects. Epidemiology, 3, 143–155.*

Rosenbaum, P. R. *(2002). Observational Studies. New York: Springer-Verlag 2nd edition.*

Rosenbaum, P. R. *(2007). Interference Between Units in Randomized Experiments. Journal of the American Statistical Association, 102 (477), 191–200.*

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

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

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

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

Rubin, B. D. *(1980). Comment on "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by D. Basu. Journal of the American Statistical Association, 75, 591–593.*

Rubin, B. D. *(1986). Which Ifs have Causal Answers? Comment on "Statistics and Causal Inference" by P. Holland. Journal of the American Statistical Association, 81, 961–962.*

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

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

Rubin, B. D. *(1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. Annals of Internal Medicine, 127 (8S), 757–763.*

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

Rubin, B. D. *(2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Health Services & Outcomes Research Methodology, 2 (1), 169–188.*

Rubin, B. D. *(2004). Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics, 31, 161–170.*

Rubin, B. D. *(2006). Matched Sampling for Causal Effects. Cambridge, England: Cambridge University Press.*

Schultz T. P. *(2004). School subsidies for the poor: evaluating the Mexican Progresa poverty program. Journal of Development Economics, 74,199–250.*

Sinclair, B., McConnell, M. & Green, D.P. *(2012). Detecting Spillover Effects: Design and Analysis of Multilevel Experiments. American Journal of Political Science, 56(4) 1055–1069.*

Small, D. S. *(2012). Mediation Analysis without Sequential Ignorability: Using Baseline Covariates Interacted with Random Assignment as Instrumental Variables. Journal of Statistical Research, 42(2), 89–101.*

Smith, H. *(1997). Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological Methodology, 27, 305–353.*

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, 1398–1407.*

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

Tanner, M. A. *&* Wong, W. H.*(1987). The Calculation of Posterior Distributions by Data Augmentation (with discussions). Journal of the American Statistical Association, 82, 528–550.*

Tchetgen Tchetgen E. J. *&* VanderWeele T. J. *(2011). Bounding the infectiousness effect in vaccine trials. Epidemiology, 22, 686–693.*

Tchetgen Tchetgen E. J. *&* VanderWeele T. J. *(2012). On causal inference in the presence of interference.. Statistical Methods in Medical Research, 21, 55–75.*

TenHave, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A. *&* Beck, A. T. *(2007). Causal mediation analyses with rank preserving models. Biometrics, 63, 926–934.*

TenHave, T. R. *&* Joffe, M. M. *(2012). A review of causal estimation of effects in mediation analyses. Statistical Methods in Medical Research, 21, 77–107.*

VanderWeele T. J. *(2008). Simple relations between principal stratification and direct and indirect effects. Statistics and Probability Letters, 78, 2957–2962.*

VanderWeele T. J. *&* Vansteelandt, S. *(2009). Conceptual issues concerning mediation, interventions and composition. Statistics and its Interface, 2, 457–468.*

VanderWeele T. J. *(2010). Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology, 21, 540–551.*

VanderWeele T. J. *(2010). Direct and Indirect Effects for Neighborhood-Based*

Clustered and Longitudinal Data. Sociological Research and Methods, 38, 515–544.

VanderWeele T. J. *(2012). Comments: Should Principal Stratification Be Used to Study Mediational Processes?. Journal of Research on Educational Effectiveness, 5, 245–249.*

VanderWeele T. J., Tchetgen Tchetgen, E. J. & Halloran, M. E. *(2012). Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. Epidemiology, 23(5), 751–761.*

VanderWeele T. J., Hong G., Jones S. M. & Brown J. L. *(2013). Mediation and Spillover Effects in Group-Randomized Trials: A Case Study of the 4Rs Educational Intervention. Journal of the American Statistical Association, 108:502, 469–482.*

Winship, C. & Morgan, S. *(1999). The estimation of causal effects from observational data. Annual Review of Sociology, 25, 659–707.*