Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets Using Bayesian Principal Stratification

Laura Forastiere, Fabrizia Mealli & Tyler J. VanderWeele

To cite this article: Laura Forastiere, Fabrizia Mealli & Tyler J. VanderWeele (2016) Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets Using Bayesian Principal Stratification, Journal of the American Statistical Association, 111:514, 510-525, DOI: 10.1080/01621459.2015.1125788

To link to this article: http://dx.doi.org/10.1080/01621459.2015.1125788

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 514, 510–525, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2015.1125788

Taylor & Francis
Taylor & Francis Group

Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets Using Bayesian Principal Stratification

Laura Forastiere, Fabrizia Mealli, and Tyler J. VanderWeele

ABSTRACT

Exploration of causal mechanisms is often important for researchers and policymakers to understand how an intervention works and how it can be improved. This task can be crucial in clustered encouragement designs (CEDs). Encouragement design studies arise frequently when the treatment cannot be enforced because of ethical or practical constraints and an encouragement intervention (information campaigns, incentives, etc.) is conceived with the purpose of increasing the uptake of the treatment of interest. By design, encouragements always entail the complication of noncompliance. Encouragements can also give rise to a variety of mechanisms, particularly when encouragement is assigned at the cluster level. Social interactions among units within the same cluster can result in spillover effects. Disentangling the effect of encouragement through spillover effects from that through the enhancement of the treatment would give better insight into the intervention and it could be compelling for planning the scaling-up phase of the program. Building on previous works on CEDs and noncompliance, we use the principal stratification 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. We show how the latter stratum-specific causal effects are related to the decomposition commonly used in the literature and provide flexible homogeneity assumptions under which an extrapolation across principal strata allows one to disentangle the effects. Estimation of causal estimands can be performed with Bayesian inferential methods using hierarchical models to account for clustering. We illustrate the proposed methodology by 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. Farmer households assigned to the program could take advantage of a deferred payment and a discount in the purchase of new bed nets. Our analysis shows a lack of evidence of an effect of the offering of the program to a cluster of households through spillover effects, that is, through a greater bed net coverage in the neighborhood. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2015
Revised October 2015

KEYWORDS
Bayesian causal inference; Interference; Mediation analysis; Principal causal effects

1. Introduction

The main purpose of clinical trials and impact evaluations is to provide evidence to guide the development of policies or programs as well as preventive or control measures. Evidence-based practice has gained considerable interest over the last decades in the fields of economics, psychology, political, social, and health sciences. With mediation analysis, research has gone far beyond providing evidence of overall effects. A deep understanding of underlying mechanisms could be used by psychologists, social workers, health services managers, or policymakers to better design their interventions. Improvements could include tailoring and focusing on particular successful components of the interventions. 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. Indeed, since the first attempts to exploration of causal mechanisms (Baron and 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 and Greenland 1992; Pearl 2001). Because of their definitions, they involve quantities, sometimes named a priori counterfactuals (Rubin 2004) or *cross-world counterfactuals*, which cannot be estimated from the observed data without strong assumptions. Most of the approaches to mediation analysis, from parametric to semiparametric estimators, hinge on *sequential ignorability* assumption (Imai, Keele, and Tingley 2010; Imai, Keele, and Yamamoto 2010; Hafeman and VanderWeele 2011), which requires unconfoundedness of both the assignment and the intermediate variable, given the baseline covariates. The key assumption of unconfoundedness of the intermediate variable is strong and will often not hold; several authors have tried to address the problem through different techniques such as

CONTACT Laura Forastiere forastiere@disia.unifi.it Department of Statistics, Computer Science, Applications, University of Florence, Florence 50134, Italy.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

instrumental variables (Dunn and Bentall 2007; Ten Have et al. 2007; Albert 2008; Small 2012), sensitivity analysis (Imai, Keele, and Yamamoto 2010; VanderWeele 2010a), or *Principal Stratification* (PS; Frangakis and Rubin 2002).

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 and Rubin 2003; Rubin 2004; Mealli and Mattei 2012). Its use to address these limitations has been developed by Hill, Waldfogel, and Brooks-Gunn (2002) and subsequently applied with Bayesian estimation techniques by Gallop et al. (2009), Elliott, Raghunathan, and Li (2010), and Page (2012). A specific augmented design was proposed by Mattei and Mealli (2011); further comparisons can be found in the literature (e.g., Jo 2008; Lynch et al. 2008; Ten Have and Joffe 2012). The precise scope of the use of PS is still subject to an ongoing debate (Pearl 2011; Mealli and Mattei 2012; VanderWeele, Tchetgen Tchetgen, and Halloran 2012).

Oftentimes, the so-called principal strata direct effects (PSDE), that is, the effect of the assignment for those whose treatment uptake does not depend on the assignment, are "wrongly" interpreted as "direct" effects for the whole population, implicitly making some kind of homogeneity assumptions across all principal strata (Jo 2008; Flores and Flores-Lagunes 2009a,b; Elliott et al. 2010; Page 2012).

Interesting questions concerning mechanisms can be raised in a typical noncompliance setting: encouragement designs. Encouragements, such as incentives, different strategies for treatment supply, or public policies in general, are used when a treatment cannot be enforced for ethical or practical reasons. Treatments can be therapeutic drugs or programs, preventive measures (e.g., vaccines, condoms, bed nets), protective, or risky behaviors (e.g., drug or alcohol abuse). We may think of the treatment as 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 circumstances, encouragement interventions can be conceived to foster a behavioral change of the target population. 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, that is, PCE, with and without exclusion restriction assumptions (Imbens and Angrist 1994; Angrist et al. 1996; Imbens and Rubin 1997). Oftentimes, in fact, the encouragement is itself a cause of alternative behaviors that would affect the outcome even without inducing a change in the treatment received. Even when the major interest relies on the effect of the encouragement on the treatment uptake and in turn on the outcome, investigating the underlying mechanisms through which the encouragement program achieves its goal is important for both descriptive and prescriptive reasons.

Here we consider *cluster randomized encouragement designs* (CEDs), 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, public policies, etc.) or because of particular constraints, but compliance is at the individual level. CEDs with individual noncompliance can be found relatively frequently in many field experiments (Sommer and Zeger 1991; McDonald, Hiu, and Tierney 1992; Hirano et al. 2000; Morris et al. 2004; among others). Frangakis, Rubin, and 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 noncompliance in cluster encouragement settings, with the treatment being the intermediate variable. CEDs are intriguing because they can give rise to several mechanisms that are worthwhile to investigate. In fact, not only does their relationship with the outcome depend on a change in the treatment uptake but often encouragements 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 what in the literature is referred to as *interference* or *spillover effects* (Sobel 2006; Hudgens and Halloran 2008; Tchetgen Tchetgen and VanderWeele 2012).

VanderWeele et al. (2013) had already attempted to disentangle spillover effects in cluster randomized trials, using sequential ignorability assumptions that accommodate cluster-level assignment and spillovers, but again the assumption made in this setting are very strong.

This article 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 one to extrapolate information across principal strata and use the estimated PCE to estimate 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 one of them. Third, building on previous work (Frangakis, Rubin, and Zhou 2002; Jo et al. 2008a,b), we incorporate an imputation-based procedure for the estimation of these intervention-based causal effects under the required assumptions.

The article is organized as follows. Section 2 describes the motivating study that we will use to illustrate the methodology. Section 3 provides notation and setup. In Section 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 5 presents our innovative structural assumptions deriving the identification results. Section 6 concerns the models for Bayesian inference. In Section 7, we study the KAHS data and Section 8 concludes.

2. Motivational Study: Katete Agriculture and Health Study (KAHS)

The proposed methodology is motivated by the Katete Agriculture and Health Study (KAHS) implemented in Katete District, a rural area with highly endemic malaria in Zambia's Eastern Province (Fink and Masiye 2012). From a list of 256 clusters, corresponding to small rural settlements of about 250 households each, the study was restricted to 49 noncontiguous

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 three arms, were surveyed twice, once prior to the rainy season and a second time 5 months later. All the 11 households selected in the clusters assigned to the third arm, after the baseline interview, were allowed to obtain insecticide-treated bed nets (ITNs) at a subsidized price, with repayments due at the end of the harvesting season with a crop sale deduction system. However, not all the households that were offered the subsidies took advantage of them by ordering new bed nets, whereas in the control clusters families could also buy new ITNs from local markets. Fink and 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, past 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 concern of interference across clusters can be reasonably ruled out. In contrast interference within clusters is likely to take place. Bed net 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, Tchetgen Tchetgen, and Halloran 2012). In addition, bed nets commonly used in the last two decades are insecticide-treated bed nets (ITNs). 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 analogs to the *infectiousness effect* (VanderWeele, Tchetgen Tchetgen, and Halloran 2012).

Some of these components of the mass community effect of bed nets have been assessed by researchers in randomized trials (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 different mechanisms through which the offer of agricultural loans had an effect.

3. Notation and Definitions

In this section, we will give formal definitions of the aforementioned effects in the potential outcomes framework (Rubin 1974). The setting consists of $j = 1, \ldots, J$ clusters and $i = 1, \ldots, N_j$ 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 the outcome variables for unit i in cluster j. We also introduce a vector of covariates, $\mathbf{C}_{ij} = (\mathbf{X}_{ij}, \mathbf{V}_j, h_i(\mathbf{X}_{-ij})) \in \mathcal{C}$, 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 treatments received and outcomes, respectively.

In the KAHS study, farmer households are the units of analysis and 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 bought at least one more bed net and $M_{ij} = 0$ if no purchase was carried out. 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 the article, we will use the term "individual" to refer to 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 household i that 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 values of the intermediate variable and potential outcomes do not vary with encouragements assigned to clusters other than the individual's own cluster, that is, $M_{ij}(\mathbf{A}) \equiv M_{ij}(A_j)$ and $Y_{ij}(\mathbf{A}, \mathbf{M}) \equiv Y_{ij}(A_j, \mathbf{M}_j)$, where $\mathbf{M}_j$ is the vector of dimension $N_J \times 1$ of treatment received by individuals of cluster j.

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

$$\begin{aligned} &\text{if} \quad A_j = A_j' \quad \text{then} \quad M_{ij}(A_j) = M_{ij}(A_j') \quad \text{and} \\ &\text{if} \quad A_j = A_j' \quad \text{and} \quad \mathbf{M}_j = \mathbf{M}_j' \\ &\text{then} \quad Y_{ij}(A_j, \mathbf{M}_j) = Y_{ij}(A_j', \mathbf{M}_j'). \end{aligned}$$

Cluster-level SUTVA is an extension of the individual-level SUTVA introduced by Rubin (1978, 1980) to settings with cluster-level assignments and individual-level intermediate variable. Yet it is worth noting that part (i) 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, $Y_{ij}(a, \mathbf{M}_j(a))$. 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 level $\mathbf{c}$ of baseline covariates, is defined as the following contrast:

$$\text{ITT}(\mathbf{c}) := \text{E}\big[Y_{ij}(1) \mid C_{ij} = \mathbf{c}\big] - \text{E}\big[Y_{ij}(0) \mid C_{ij} = \mathbf{c}\big]. \quad (3.1)$$

To shed light on the heterogeneity of the effects, we will define all causal estimands as average effects within levels of the baseline covariates C_{ij}.

4. Principal Stratification Approach

Principal stratification has been first introduced by Frangakis and Rubin (2002), to address post-treatment complications. Its use in mediation analysis has been proposed as a way to relax the sequential ignorability assumption and focus on so-called principal strata direct effects (VanderWeele 2008; Gallop et al. 2009; Elliott et al. 2010; Mattei and Mealli 2011; Page 2012). 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:

$$S^{m_0 m_1} := \big\{i : M_{ij}(0) = m_0, M_{ij}(1) = m_1\big\}. \quad (4.1)$$

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 four 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 can argue that this last category is not plausible in this setting and thus we make the following 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 conveys the idea 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 conditional distribution of compliance status to be consistently estimated. In fact, if we let $\pi_{m_0 m_1}(\mathbf{c}) := P(S_{ij} = S^{m_0 m_1} \mid C_{ij} = \mathbf{c})$ denote the probability of belonging to stratum $S^{m_0 m_1}$ conditional on baseline covariates, the monotonicity assumption implies the following result $\forall \mathbf{c} \in \mathcal{C}$:

$$\begin{aligned} &\pi_{10}(\mathbf{c}) = 0; \quad \pi_{11}(\mathbf{c}) = P(M_{ij}(0) = 1 \mid C_{ij} = \mathbf{c}); \\ &\pi_{00}(\mathbf{c}) = P(M_{ij}(1) = 0 \mid C_{ij} = \mathbf{c}); \\ &\pi_{01}(\mathbf{c}) = 1 - \pi_{11}(\mathbf{c}) - \pi_{00}(\mathbf{c}). \end{aligned} \quad (4.2)$$

As mentioned previously, in the KAHS study this assumption is plausible because there should not be any reason to buy a bed net at a full price but not with subsidies.

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

$$\begin{aligned} \text{PCE}(\text{m}_0, \text{m}_1, \mathbf{c}) : = &\text{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m_1}, C_{ij} = \mathbf{c}\big] \\ &-\text{E}\big[Y_{ij}(0) \mid S_{ij} = S^{m_0 m_1}, C_{ij} = \mathbf{c}\big] \end{aligned} \quad (4.3)$$

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

$$\text{ITT}(\mathbf{c}) = \sum_{m_0 m_1} \text{PCE}(\text{m}_0, \text{m}_1, \mathbf{c}) \cdot \pi_{m_0 m_1}(\mathbf{c}). \quad (4.4)$$

In principal strata where the treatment receipt is unaffected by the encouragement, that is, never-takers and always-takers, principal causal effect, $\text{PCE}(m, m, \mathbf{c})$ with $m \in \{0, 1\}$, are called *dissociative causal effect* ($\text{DCE}(m, \mathbf{c})$).

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 in individual behaviors other than in the treatment receipt (Frangakis, Rubin, and Zhou 2002), and effects due to mechanisms of *interference*, by behavioral changes in other inhabitants of the same cluster, both in terms of treatment receipt or in terms of other actions. Several behavioral changes often occur when the encouragement intervention is provided by information campaigns which will increase the awareness of the problem and also encourage the use of other measures together with the treatment of interest. Interventions designed to boost the use of bed nets often comprise different components that are responsible for different mechanisms. 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 dissociative effects for never-takers and always-takers can be substantial. On the one hand, this can be due to the possible encouragement–treatment interaction, 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 C_{ij} would allow one to identify individual as well as cluster characteristics of the units who do not get any benefit from the cluster intervention when they do not 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, for example, free distribution of bed nets. 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 subpopulation. As far as compliers are concerned, $\text{PCE}(0, 1, \mathbf{c})$ is a combination of all the aforementioned mechanisms as well as the effect of the encouragement involving a change in the individual treatment uptake.

4.2. Individual Treatment Mediated Effect and Net Encouragement Effect

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 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 four 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 $\tilde{a}$, that is, $Y_{ij}(a, M_{ij}(\tilde{a}))$. 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 where 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 it creates the condition for which that household finds nets out of stock, assuming no secondary consequences.

Potential outcomes of the type $Y_{ij}(a, M_{ij}(\tilde{a}))$, 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. We now 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:

$$\begin{aligned}\text{NEE}^{\tilde{a}}(m_0, m_1, \mathbf{c}) :&= \text{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) \mid S_{ij} = S^{m_0 m_1}, C_{ij} = \mathbf{c}\big] \\ &\quad - \text{E}\big[Y_{ij}\big(0, M_{ij}(\tilde{a})\big) \mid S_{ij} = S^{m_0 m_1}, C_{ij} = \mathbf{c}\big]. \end{aligned} \tag{4.5}$$

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 $C_{ij} = \mathbf{c}$. 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, $\text{NEE}^{\tilde{a}}(m_0, m_1, \mathbf{c})$ indicates the average, over all units with $C_{ij} = \mathbf{c}$ 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:

$$\begin{aligned}\text{iTME}^a(m_0, m_1, \mathbf{c}) := \text{E}\big[Y_{ij}\big(a, M_{ij}(1)\big) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \\ -\text{E}\big[Y_{ij}\big(a, M_{ij}(0)\big) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big].\end{aligned} \tag{4.6}$$

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 characteristic, that is, 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 and Greenland 1992; Pearl 2001) within principal strata (VanderWeele 2008; Mealli and 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:

$$\text{DCE}(m, \mathbf{c}) \equiv \text{NEE}^0(m, m, \mathbf{c}) = \text{NEE}^1(m, m, \mathbf{c}). \tag{4.7}$$

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

$$\text{PCE}(0, 1, \mathbf{c}) = \text{NEE}^{1-a}(0, 1, \mathbf{c}) + \text{iTME}^a(0, 1, \mathbf{c}). \tag{4.8}$$

In our example, $\text{iTME}^a(0, 1, \mathbf{c})$ 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 $A_j = a$.

Equation (4.8) is the typical decomposition found in mediation analysis. The intuition behind this decomposition, with different indexes for the two complementary effects, arises from the possible presence of an interaction between the treatment variable and the encouragement intervention. In fact, translating the work by VanderWeele (2014) into this context, the total effect of the intervention for compliers is given by the sum of the effect of the intervention when they do not buy new bed nets, the effect of buying new bed nets when the program has not been offered to the whole cluster plus the differential effect of new bed nets in the presence or absence of the program in the cluster. This last interaction effect could be included in the individual treatment mediated effect, resulting in the decomposition given by $\text{NEE}^0(0, 1, \mathbf{c}) + \text{iTME}^1(0, 1, \mathbf{c})$, or in the net encouragement effect, resulting in $\text{NEE}^1(0, 1, \mathbf{c}) + \text{iTME}^0(0, 1, \mathbf{c})$. The choice of one or the other decomposition, indexed by a, depends on the specific application. As we can see, compliers are the only units to actually exhibit a nonzero iTME besides a possible difference between the two NEEs.

In any case, a conceptual point has to be made. In this application, the effects $\text{NEE}^1(0, 1, \mathbf{c})$ and $\text{iTME}^0(0, 1, \mathbf{c})$ 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 likely lead to other mechanisms. On the contrary, $\text{NEE}^0(0, 1, \mathbf{c})$ and $\text{iTME}^1(0, 1, \mathbf{c})$ 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. 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 $\text{NEE}^0(0, 1, \mathbf{c})$ and $\text{iTME}^1(0, 1, \mathbf{c})$ for compliers 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:

$$\begin{aligned}\text{NEE}^0(\mathbf{c}) &= \sum_{(m_0, m_1)} \text{NEE}^0(m_0, m_1, \mathbf{c})\pi_{m_0 m_1}(\mathbf{c}) \\ &= \sum_m \text{DCE}(m, \mathbf{c})\pi_{mm}(\mathbf{c}) + \text{NEE}^0(0, 1, \mathbf{c})\pi_{m_0 m_1}(\mathbf{c}).\end{aligned} \tag{4.9}$$

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

$$\text{iTME}^1(\mathbf{c}) = \text{iTME}^1(0, 1, \mathbf{c})\pi_{01}(\mathbf{c}). \tag{4.10}$$

By virtue of the particular behavior of compliers, that is, $M_{ij}(0) = 0$ and $M_{ij}(1) = 1$, we can interpret their individual treatment mediated effect as the average causal effect of the receipt of treatment within this subpopulation (see supplemental materials). This makes clear, then, how 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).

5. Assumptions for Causal Mechanisms

Throughout we will make the following assumption:

Assumption 3. Unconfoundedness of the cluster 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:

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

When the encouragement is randomized, unconfoundedness of the encouragement assignment holds without conditioning on covariates. This is actually the case in the KAHS study. It is worth remarking 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. Observed value of the intermediate variables are in general mixtures of different principal strata. The monotonicity Assumption 2 allows one to identify as always-takers the units in the control group who take the treatment and, similarly, as never-takers those who do not take the treatment under encouragement. However, in all other observed groups strata membership is not known. Without additional assumptions—such as exclusion restrictions, which substantially rule out the presence of net effects and hence cannot be invoked in these settings where these are effects of interest—this missing information results in models and causal parameters that are not fully identified, that is, for which a consistent estimator does not exist. 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. Once principal causal effects have been estimated, a full assessment of causal mechanisms, as defined in the previous section, solely requires a last step, which is the decomposition of $\text{PCE}(0, 1, \mathbf{c})$ for compliers into $\text{NEE}^0(0, 1, \mathbf{c})$ and $\text{iTME}^1(0, 1, \mathbf{c})$.

NEEs and iTMEs involve potential outcomes of the form $Y_{ij}(a, m)$ and, in particular, causal estimands of interest here are defined based on comparisons between $Y_{ij}(1, M_{ij}(0))$, $Y_{ij}(0, M_{ij}(0))$, and $Y_{ij}(1, M_{ij}(1))$. 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 is randomized, only one of all these possible potential outcomes is ultimately observed, namely, $Y_{ij}(A_j, M_{ij}(A_j))$, where A_j is the encouragement status assigned to cluster j. Potential outcomes of the type $Y_{ij}(a, m)$, with m set to a particular value for all units or to $M_{ij}(\tilde{a})$, are observable only if $m \equiv M_{ij}(a)$, in which case the potential outcome collapses in $Y_{ij}(a)$. This 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 $M_{ij}(0) \equiv M_{ij}(1)$, that is, the unit is a never-taker or an always-taker. On the contrary, potential outcomes can never be (not even potentially) observed for units with $M_{ij}(a) \neq m$, hence they are called a priori counterfactuals (Rubin 2004). Here the problematic counterfactuals are $Y_{ij}(1, M_{ij}(0))$ for compliers. Estimation of these missing potential outcomes would require an extrapolation from other individuals in the data, based on specific assumptions.

Identification of effects involving a priori counterfactuals typically relies on sequential ignorability assumptions (see Ten Have and Joffe (2012) for a review of the different specifications). 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, that is, 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. In fact, in our empirical study, households' 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 individual treatment mediated effects and net encouragement effects for all principal strata.

5.1. Homogeneity Assumptions

When the sequential ignorability assumption does not hold, information about $Y_{ij}(1, M_{ij}(0))$ for compliers cannot be extrapolated across strata and thus the principal causal effect for this subpopulation cannot be decomposed into the two causal estimands of interest, $\text{NEE}^0(0, 1, \mathbf{c})$ and $\text{iTME}^1(0, 1, \mathbf{c})$. Here, we provide two alternative homogeneity assumptions that enable 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 solely concern the missing information and only allow the extrapolation that is strictly needed across strata with a similar compliance behavior at least under one encouragement condition. Here, we focus on identification of $Y_{ij}(1, M_{ij}(0))$ for compliers and present two alternative assumptions that allow to identify the same causal estimands, namely, $\text{NEE}^0(0, 1, \mathbf{c})$ and hence $\text{iTME}^1(0, 1, \mathbf{c})$, in two different ways.

Assumption 4. Stochastic homogeneity of the counterfactual across never-takers and compliers.

Stochastic homogeneity of the counterfactual $Y_{ij}(1, M_{ij}(0))$ across never-takers and compliers holds if

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

Assumption 4 coveys the idea that the distribution of the counterfactual $Y_{ij}(1, M_{ij}(0))$, which corresponds to $Y_{ij}(1, 0)$ for never-takers and compliers, is the same for these two principal strata, conditioning on baseline covariates. This allows one to estimate the a priori counterfactual $Y_{ij}(1, M_{ij}(0))$ 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 $Y_{ij}(0, M_{ij}(0))$, we could assume that it is also true when encouragement is set to the opposite condition.

Theorem 1. If Assumption 4 holds, for compliers the net encouragement effect $\text{NEE}^0(0, 1, \mathbf{c})$, within levels of covariates, is given by

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

Proof. see supplemental materials. □

Often Assumption 4 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 5. Homogeneity of mean difference between counterfactuals for never-takers and compliers.

Homogeneity of the mean difference between counterfactuals across never-takers and compliers holds if

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

In words, it 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, that is, 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 give the following theorem.

Theorem 2. If Assumption 5 is satisfied, for compliers the net encouragement effect $\text{NEE}^0(0, 1, \mathbf{c})$, within levels of covariates, can be extrapolated from the dissociative causal effect for never-takers:

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

Proof. see supplemental materials. □

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 5 allows then to estimate $\text{NEE}^0(0, 1, \mathbf{c})$ for compliers and hence $\text{NEE}^0(\mathbf{c})$ in the entire population.

Assumptions 4 and 5 provide the possibility of a generalization of the potential outcome $Y_{ij}(1, M_{ij}(0))$ or the net encouragement effect NEE^0 from never-takers to compliers, as stated by Theorems 1 and 2. As a consequence, these assumptions also yield identification of the individual treatment mediated effect in the latter principal stratum, $\text{iTME}^1(0, 1, \mathbf{c})$.

Corollary 1. If Assumption 4 holds, the individual treatment mediated effect for compliers $\text{iTME}^1(0, 1, \mathbf{c})$, within levels of covariates, is given by

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

If Assumption 5 holds, the individual treatment mediated effect for compliers $\text{iTME}^1(0, 1, \mathbf{c})$, within levels of covariates, is given by

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

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 as correlation among individuals arising from common environmental factors and even reciprocal influence cannot be ignored. 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. Further, the level of resemblance in outcomes may vary across different individual strata. Correlation in cluster randomized trials with individual noncompliance has been intensively studied by Jo et al. (2008a,b), after Frangakis, Rubin, and Zhou (2002), who were the first authors to accommodate in their analysis correlation in both outcome and noncompliance status. Here, we extend the model framework used by Frangakis, Rubin, and Zhou (2002) for a different outcome distribution.

Potential outcome model. Although the general framework applies to any outcome models, here we present the model used to analyze KAHS study. 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 $Y_{ij}(a)$

$$Y_{ij}(a) | S_{ij}, \mathbf{C}_{ij} \sim \frac{\text{Bin}\big(n_{ij}, p_{ij}\big)}{n_{ij}} \tag{6.1}$$

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

$$g\Big(p_{ij}\big(a, S_{ij}, \mathbf{C}_{ij}\big)\Big) = \boldsymbol{\beta}^{S_{ij}T}\mathbf{Z}_{ij}^{Yf} + \boldsymbol{b}_j^T\mathbf{Z}_{ij}^{Yr} = \boldsymbol{\beta}_0^{S_{ij}T}\mathbf{C}_{ij}' + \boldsymbol{\beta}_1^{S_{ij}T}\mathbf{C}_{ij}'a + b_{0j} + \mathbf{b}_{1j}^T\mathbf{X}_{ij}\mathbf{b}_j \sim N(\mathbf{0}, \Sigma_b), \tag{6.2}$$

where $g(\cdot)$ 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 $\mathbf{Z}_{ij}^{Yf} = [1, \mathbf{C}_{i_j}, a, \mathbf{C}_{i_j}a]$ and $\mathbf{Z}_{ij}^{Yr} = [1, \mathbf{X}_{ij}]$, 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:

$$g\Big(P\big(S_{ij}=S^{m_0m_1}|\mathbf{C}_{i_j}\big)\Big) = \boldsymbol{\alpha}^T\mathbf{Z}_{ij}^{Sf}+\mathbf{a}_j^T\mathbf{Z}_{ij}^{Sr}=\boldsymbol{\alpha}^T\mathbf{C}'_{i_j} + \mathrm{a}_{0j} + \mathbf{a}_{1j}^T\mathbf{X}_{ij}$$
$$\mathbf{a}_j \sim N(\mathbf{0}, \Sigma_a), \tag{6.3}$$

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

Here, we follow the approach used in Frangakis, Rubin, and 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 $P(Y_i \le l)|\ \mathbf{C}_{ij}) = \Phi(\boldsymbol{\alpha}_l\ \mathbf{C}_{i_j}))$, with $l = 1, \ldots, L-1$, so that the probability of belonging to the category l ends up being $P(Y_i = l|\ \mathbf{C}_{i_j}) = (P(Y_i \le l+1)|\ \mathbf{C}_{i_j}))(1 - \Phi(\boldsymbol{\alpha}_l\ \mathbf{C}_{i_j})))$. The function $\Phi(\cdot)$ is the standard normal cumulative distribution function.

According to this parameterization 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 exclusion from the never-taker stratum. In our setting of cluster-based interventions, we extend the above model to an *ordinal mixed probit model*, parameterized as

$$\begin{aligned}
\Psi_n(\mathbf{C}_{i_j}, \boldsymbol{\alpha}, \mathbf{a}) &= P(S_{ij} = S^{00})|\mathbf{C}_{i_j}) = 1 - \Phi\big(\boldsymbol{\alpha}_n^T Z_{ij}^{Sf} + \mathbf{a}_{nj}^T Z_{ij}^{Sr}\big)\\
\Psi_c(\mathbf{C}_{i_j}, \boldsymbol{\alpha}, \mathbf{a}) &= P(S_{ij} = S^{01})|\mathbf{C}_{i_j}) = \big(1 - \Psi_n(\mathbf{C}_{i_j}, \boldsymbol{\alpha}, \mathbf{a})\big)\\
&\quad \times\big(1 - \Phi\big(\boldsymbol{\alpha}_c^T Z_{ij}^{Sf} + \mathbf{a}_{cj}^T Z_{ij}^{Sr}\big)\big)\\
\Psi_a(\mathbf{C}_{i_j}, \boldsymbol{\alpha}, \mathbf{a}) &= P(S_{ij} = S^{11})|\mathbf{C}_{i_j}) = 1 - \Psi_n(\mathbf{C}_{i_j}, \boldsymbol{\alpha}, \mathbf{a})\\
&\quad -\Psi_c(\mathbf{C}_{i_j}, \boldsymbol{\alpha}, \mathbf{a}),
\end{aligned} \tag{6.4}$$

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

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

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:

$$S_{ij} = \begin{cases} S^{00} & \text{if} \quad S_{ij}^n \equiv \boldsymbol{\alpha}_n^T Z_{ij}^{Sf} + \mathbf{a}_{nj}^T Z_{ij}^{Sr} + V_{ij} \le 0 \\ S^{01} & \text{if} \quad S_{ij}^n \ge 0 \text{ and } S_{ij}^c \equiv \boldsymbol{\alpha}_c^T Z_{ij}^{Sf} + \mathbf{a}_{cj}^T Z_{ij}^{Sr} + U_{ij} \le 0 \\ S^{11} & \text{if} \quad S_{ij}^n \ge 0 \text{ and } S_{ij}^c \ge 0 \end{cases}, \tag{6.5}$$

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

7. Application to KAHS Study

The baseline covariates that a preliminary analysis has shown to be useful for predicting strata membership include: the number of household members ($C_{1_{ij}}$), 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' hyperparameters 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 compliance within cluster correlation, 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, that is, $\mathbf{a}_{n1j} = \mathbf{a}_{c1j} = 0$. As a consequence, random effects are assumed to follow a uni-dimensional normal distribution, $a_{n0j} \mid \sigma_{a_n}^2 \sim N(0, \sigma_{a_n}^2)$ and $a_{c0j} \mid \sigma_{a_c}^2 \sim N(0, \sigma_{a_c}^2)$, and the conjugate prior distribution of their variances reduces from inverse-Wishart to inverse-gamma, $\sigma_{a_n}^2 \sim \mathrm{IG}(\eta_0^n, s_0^n)$ and $\sigma_{a_c}^2 \sim \mathrm{IG}(\eta_0^c, s_0^c)$, where we set $\eta_0^n = \eta_0^c = 0.01$ and $s_0^n = s_0^c = 0.01$.

As already said, we posit a binomial distribution for the potential outcomes, with probability $p_{i_j}(a, S_{ij}, \mathbf{C}_{i_j})$ being a function of the principal stratum, the encouragement condition and baseline covariates, as modeled in (6.2). We adopt as the logit link $g(\cdot)$. In the outcome model, we consider a subset of $\mathbf{C}_{i_j}$ 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: $\beta_{11}^{S_{ij}} = \beta_{12}^{S_{ij}} = \beta_{14}^{S_{ij}} = \beta_{15}^{S_{ij}} = \beta_{16}^{S_{ij}} = 0, \forall S_{ij} \in \{S^{00}, S^{01}, S^{11}\}$. In addition, we let the coefficients for baseline covariates be the same across strata, with the exception for the covariate that is also present in the interaction term: $\beta_{0k}^{S^{00}} = \beta_{0k}^{S^{01}} = \beta_{0k}^{S^{11}} = \beta_{0k}$, with $k = 1, 2, 4, 5$.

As with the principal strata model, between-cluster variation is taken into account by the inclusion of random intercepts, with

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 nonbuyers $(M = 0)$	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)

NOTE: 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.

the argument that the dependence 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: $\mathbf{b}_{1j} = 0, \forall j$. Finally, the model in (6.2) for the probability of the binomial potential outcome, can be rewritten as

$$\begin{aligned}\operatorname{logit}\left(p_{ij}\left(a, S_{ij}, \mathbf{C}_{i_j}\right)\right) &= \beta_{00}^{S_{ij}} + \beta_{02}C_{2_{ij}} + \beta_{03}^{S_{ij}}C_{3_{ij}} + \beta_{04}C_{4_{ij}} + \beta_{05}C_{5_{ij}} \\ &\quad + \beta_{06}C_{6_{ij}} + \beta_{10}^{S_{ij}}a + \beta_{11}^{S_{ij}}C_{3_{ij}}a + b_{0j} \\ b_{0j} &\sim N\left(0, \sigma_b^2\right). \end{aligned} \tag{7.1}$$

We postulate a multivariate normal prior for $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0})$, with hyperparameters $\boldsymbol{\mu}_{\beta 0} = \mathbf{0}$ and $\Lambda_{\beta 0} = 10\,\mathrm{I}$, and an inverse-gamma prior distribution for the variance of the random intercept, $\sigma_b^2 \sim \mathrm{IG}(\eta_0^b, s_0^b)$, setting $\eta_0^b = 0.01$ and $s_0^b = 0.01$.

Table 1 gives some basic information on 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 the assignment. The lack of perfect balance for some of them is handled by covariate 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 are always-takers and the remaining nonbuyers households are either never-takers or compliers. 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. These estimates are a result of the monotonicity assumption, which is arguably satisfied given that the presence of households, labeled as defiers, that would buy new bed nets at full prize and would not if offered loans, is implausible. The last row in Table 1 provides an ITT analysis, indicating that the encouragement intervention—being the offer of loans to buy bed nets randomly assigned at cluster level to the 11 farmers participating in the study and belonging to the same settlement of villages—results in a 44.9%$(= 0.0476/0.1060 \times 100)$ reduction of the risk of contracting 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 be interpreted causally because of the different compliance types involved in such contrast, due to the intermediate variable not being randomized.

Furthermore, we can make speculations on the mechanisms that in KAHS might give rise to the effect of the agricultural loan program for different types of households. Since no additional awareness campaign and no community-based interventions were provided to treated clusters, the effect for never-takers, that is, $\mathrm{DCE}(0, \mathbf{c})$, can be hypothesized to be mostly due to spillovers of the purchase of new bed nets by other households belonging to the same cluster. Conversely, we can assume that for always-takers spillover effects, if any, can explain only a small part of the effect of the encouragement intervention, whereas most of it likely follows from a greater number of bed nets they would buy under the loan program, thanks to the subsidized price and the deferred payment. This nonnegative dissociative effect $\mathrm{DCE}(1, \mathbf{c})$ is due to the particular choice of the binary treatment variable that only distinguishes the purchase of zero versus at least one new bed net between baseline and follow-up ($M_{ij} = 1$ if household i in cluster j has bought at least one more bed net and $M_{ij} = 0$ if no purchase has been made). As opposed to always-takers, complier households would not buy any new bed net if not encouraged by the loan program. Therefore, for compliers the intervention is presumed to reduce malaria by prompting them to get new bed nets they would not buy otherwise. This effect would be the individual treatment

Table 2. Estimated parameters for principal strata model (6.5).

	S^n Model		S^c Model	
	Mean	95% Interval	Mean	95% Interval
α coefficients				
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]

mediated effect, that is, $\text{iTME}^1(0, 1, \mathbf{c})$. Moreover, the purchase of bed nets in neighboring households can, in principle, also benefit complier households. This effect of the clustered intervention through spillovers would explain part, if not all, of the net encouragement effects for compliers, that is, $\text{NEE}^0(0, 1, \mathbf{c})$. These speculations only affect the interpretation of the estimated effects but do not alter the analysis.

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 (4) 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 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 (5), 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). On the basis of our previous hypotheses, this means that never-takers and compliers would benefit from the increased number of bed nets in the surroundings in the same way, that is, they would share the same effect of the program assigned in their cluster mainly through spillovers, if their bed net coverage were kept unchanged.

In the supplemental material, we provide details on how to exploit the homogeneity assumptions to make inference on the two causal mechanisms making use of a Bayesian imputation approach. In this approach, imputation of the required potential outcomes follows from these assumptions in a straightforward way. The adoption of one of the two homogeneity assumptions implies different steps in the final Bayesian imputation procedure.

7.1. 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 (6.5). 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, $C_{3_{ij}}$ 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 a complier (mean and 95% interval for α_{c4}: −2.289 [−4.540,−1.634]), as compared with being an always-taker. 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: households with higher coverage are less likely to buy new bed nets. Another expected result is that neighborhood baseline coverage reduces the probability of being a never-taker, probably because of a peer influence.

We have also derived estimates of the sample mean of the covariates within each principal stratum, that is, $\overline{C}_{hm_0m_1} = \sum_{i,\, j:S_{ij}=S^{m_0m_1}} C_{h_{ij}}/|S^{m_0m_1}|, \forall k \in \{1, \ldots, 5\}$ 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 previous findings, 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

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]

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]

NOTE: 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 $A_j = 0$, that is, $\overline{Y}(0)$. Both distributions are averaged over $\mathbf{C}_{ij}$ (or just over $C_{1_{ij}}$, $C_{2_{ij}}$, $C_{4_{ij}}$, and $C_{5_{ij}}$ when results are presented within household baseline coverage categories $\widetilde{C}_{4_{ij}}$), the clusters, and $\boldsymbol{\theta}$.

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.

The estimated variance of the random intercept $a_{n0\,j}$ included in the model for the conditional probability of being a never-taker versus being an always-taker or a complier (model for S^n_{ij}) is estimated to be 0.096 (95% quintiles: [0.017, 0.125]) reflecting in an intraclass correlation of 0.088 (Table 2). Similarly, the estimated variance of the random intercept $a_{c0\,j}$ of the model for S^c_{ij}, conditional on not being never-takers, is estimated to be 1.588 (95% quintiles: [0.035, 1.597]), reflecting in an intraclass correlation of 0.614: the proportion of never-takers does not seem to 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:

$$\widetilde{C}_{4_{ij}} = \big\{\text{Low coverage } \big(\text{if } C_{4_{ij}} \le 0.4\big), \text{ medium coverage } \big(\text{if } 0.4 < C_{4_{ij}} \le 0.8\big), \text{ high coverage } \big(\text{if } C_{4_{ij}} > 0.8\big)\big\}.$$

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 potential outcomes. The right column of Table 4 summarizes the predictive posterior distribution of malaria rates without encouragement, 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 or other protective behaviors. For always-takers, 95% intervals get wider as coverage augments due to the small proportion of always-takers in higher levels, hence no definite conclusion can be drawn. For compliers, posterior means seem to decrease with $\widetilde{C}_{4_{ij}}$, but still intervals cover zero making this pattern consistent to random fluctuation. Second, we compare principal strata: at all coverage levels, compliers are those households who would have a considerably higher risk of malaria infection if not encouraged, with an overall mean risk of 32.1% against 6% for never-taker and 8.4% for always-takers. This result may 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, the household coverage at follow-up would increase compared to baseline, even without the encouragement. 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, if not encouraged compliers seem to be the subpopulation most at risk of malaria. The reason might 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 possibly, for those with medium high coverage, old bed nets in bad physical integrity. The offer of loans to buy bed nets might make compliers more aware of their risk.

In any case, the different mean of potential outcomes under control encouragement between principal strata supports our hypothesis of assumption 4 of partial stochastic homogeneity of counterfactuals being implausible.

Table 5 concerns the estimated effects defined in Section 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

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]

NOTE: 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 $\widetilde{C}_{4_{ij}}$. The last block of rows concerns the estimated effect in the whole population.

outcomes (see supplemental materials). Results are based on 45,000 iterations, combining three chains, each run for 25,000 iterations, with a burn-in of 10,000 iterations. To check for convergence, for each effect we computed the potential scale reduction factor (Gelman 1996), giving a maximum value of 1.04, suggesting no evidence against convergence (details available from the authors upon request).

Consider principal causal effects, presented in the last block of columns. The estimated PCE for compliers is on average 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 is much larger than PCEs in the other principal strata, being PCE(0, 1, **c**) the sum of $NEE^0(0, 1, \mathbf{c})$ and $iTME^1(0, 1, \mathbf{c})$, that is, the effect of the loan program both through spillovers and through the purchase of bed nets by the household itself. The estimated PCE for always-takers, that is, DCE(1, **c**) is on average a reduction of the risk of infection of 6.2% (posterior mean), which is mainly due to the increased number of bed nets bought under the program. 95% credible intervals provide strong evidence of a beneficial effect of the encouragement for both compliers and always-takers.

For never-takers, instead, we find a negligible effect of the encouragement, that is, DCE(0), for all levels of coverage. The proportion of malaria cases at baseline and potential proportion under control encouragement have not suggested lack of knowledge and awareness of malaria for this subpopulation, but, on the contrary, never-takers are likely the most aware of preventive measures. 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 suggesting no evidence of spillover effects at any coverage level, at least for never-takers.

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 directly 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. Therefore, all the effect of the encouragement for this principal stratum would be through the purchase of new bed nets.

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, that is, $CACE^1$.

Average net encouragement effects in the whole population 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 compliers, we obtain an estimate of the individual treatment effect in the population given by −2.6% (95% interval: [−6.8%, 0.4%]).

All these results rely on both structural and modeling assumptions that were laid out in the previous sections. While some structural assumptions, such as cluster-level SUTVA (assumption 1) and unconfoundedness of the cluster encouragement assignment (assumption 3), plausibly hold by design, monotonicity (assumption 2), and the homogeneity assumptions (assumption 4 or 5) have been invoked based on subject-matter knowledge and are not directly testable from the data. However, monotonicity has a testable implication, which we have verified in our study: monotonicity is not falsified in our data because the number of units who buy new bed nets is significantly greater in the treatment arm than in the control arm. On the other hand, both homogeneity assumptions involve a priori counterfactuals that are never observed in this experiment on any subject and, therefore, do not generate any testable

implications. The plausibility can only be judged either by expert knowledge or by some related evidence in the data. For example, the fact that never-takers and compliers are observed to have the same conditional distribution of the potential outcomes $Y_{ij}(0, M_{ij}(0))$ can be viewed as a support of the assumption that they might also share the distribution of $Y_{ij}(1, M_{ij}(0))$ (Assumption 4). Also, because homogeneity assumptions are assumptions on a prior-counterfactuals, usual sensitivity analyses cannot be conducted for them, as any deviation from homogeneity would change the estimate of the net encouragement effect for compliers by exactly that same amount.

As far as modeling assumptions are concerned, to assess the sensitivity of posterior inference to the specification of the prior distribution, we have derived the distribution of our causal estimands using only the prior distribution over parameters. None of these distributions appear particularly informative for the causal estimands, thereby suggesting that the precision of the posterior distribution is mainly driven by the information in the data (Frangakis, Rubin, and Zhou 2002).

Regarding the models for the outcomes and the principal strata membership, the relative binomial distribution for the proportion of malaria cases and the linked probit model appear to be a reasonably flexible choice. As a model fit diagnostic, we have compared quantities that can be directly estimated from the data with their estimates derived from the estimated model. For example, the global intent-to-treat effect, estimated from the sample statistics, is equivalent to its estimate derived as a weighted average of principal stratum-specific effects (see Tables 1 and 5). Overall model fit could also be assessed using posterior predictive checks, which are however beyond the scope of the article.

8. Discussion

In this article, we provide a framework based on principal stratification approach to investigate the different mechanisms elicited in cluster encouragement designs. The core of this work concerns the proposal of new homogeneity assumptions allowing one to disentangle two different effects for the subpopulation of compliers, under violation of sequential ignorability. Even if we could assume sequential ignorability, which is stronger than our homogeneity assumptions, the characterization of principal strata and the estimation of causal mechanisms for different types of individuals is still an advantage of our proposed methodology over classical mediation analysis.

Principal causal effects themselves provide 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 their village reduces the risk of contracting malaria. It also suggests that those who would proceed anyway with the purchase of bed nets, 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. It might be the case that the potential beneficial effect of the bed nets in other households killing mosquitos is offset by the effect of diverting mosquitos to households without bed nets. The slightly detrimental effect for this subpopulation, especially with low coverage, even if intervals are too wide to draw definite conclusions, suggests the importance of investigating 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 within 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. 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 preventative measures. This characterization of principal strata can also help us understand whether and which homogeneity assumption is more plausible to untie the mediated and nonmediated effects among compliers.

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 under study, in the supplemental material we provide more general homogeneity assumptions. Each specific assumption enables the identification of a combination of the two effects $\text{NEE}^{\tilde{a}}$ and $\text{iTME}^{1-\tilde{a}}$, with $\tilde{a} = 0$ or $\tilde{a} = 1$, for always-takers, never-takers, and compliers or defiers. 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, and found no evidence of a net encouragement effect among compliers, at any coverage level. Therefore, spillover effects are negligible for both never-takers and compliers and most of the effect of the clustered loan program for compliers is achieved by making these households buy new bed nets. This conclusion is important in that it shows how the sole 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 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, but almost as much is given by the effect due to the subsidized price and deferred payment through the increased number of new bed nets among the always-takers, who constitute 41% of the population.

Since a negligible effect has been 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, etc.), we may want to exclude this subpopulation from the encouragement program. The lack of knowledge of strata membership would force one to exclude those units with higher probability of being never-takers, that is, those with higher baseline coverage. 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 to select appropriate suites of interventions.

Our analysis of KAHS has several limitations. First, the choice of a binary intermediate variable, although it sheds 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 require can be problematic if principal strata are highly clustered, due to mechanisms such as homophily or peer influence in the compliance behavior. 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. In any case, 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 may be less controversial.

Supplementary Materials

Section 1: Identifying Assumptions for Causal Mechanisms
Section 2: Controlled Net Encouragement Effects Within Principal Strata
Section 3: Average Treatment Effect
Section 4: Bayesian Inference
Section 5: Proofs of Other Equations

Acknowledgment

The authors thank Alessandra Mattei for helpful suggestions and Gunther Fink for providing the KAHS data.

Funding

This work is partially funded by PRIN 2012 grant and NIH grant ES017876.

References

Albert, J. (2008), "Mediation Analysis via Potential Outcomes Models," *Statistics in Medicine*, 27, 1282–1304. [511]

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., and 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. [512]

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–472. [511]

Barnard, J., Frangakis, C., Hill, J., and 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. [518]

Baron, R. M., and Kenny, D. A. (1986), "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182. [510]

Bartolucci, F., and Grilli, L. (2011), "Modelling Partial Compliance through Copulas in a Principal Stratification Framework," *Journal of the American Statistical Association*, 106, 469–479. [524]

Binka, F. N., Indome, F., and Smith, T. (1998), "Impact of Spatial Distribution of Permethrin-Impregnated Bednets on Child Mortality in Rural Northern Ghana," *American Journal of Tropical Medicine and Hygiene*, 59, 80–85. [512]

D'Alessandro, U., Olaleye, B. O., McGuire, W., Langerock, P., Bennett, S., Aikins, M. K., Thomson, M. C., Cham, M. K., Cham, B. A., and Greenwood, B. M. (1995), "Mortality and Morbidity From Malaria in Gambian Children After Introduction of an Impregnated Bednet Programme," *Lancet*, 345, 479–483. [512]

Dunn, G., and Bentall, R. (2007), "Modelling Treatment-Effect Heterogeneity in Randomized Controlled Trials of Complex Interventions (Psychological Treatments)," *Statistics in Medicine*, 26, 4719–4745. [511]

Elliott, M. R., Raghunathan, T. E., and Li, Y. (2010), "Bayesian Inference for Causal Mediation Effects Using Principal Stratification With Dichotomous Mediators and Outcomes," *Biostatistics*, 11, 353–372. [511,513]

Fink, G., and 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. [511,512]

Flores, C. A., and Flores-Lagunes, A. (2009a), "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment," IZA Discussion Paper No. 4237, Bonn, Germany: IZA. [511]

——— (2009b), "Nonparametric Partial and Point Identification of Net or Direct Causal Effects," American Economic Association, Annual Meeting Paper. [511]

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29. [511,513]

Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002), "Clustered Encouragement Design With Individual Noncompliance: Bayesian Inference and Application to Advance Directive Forms" (with discussion), *Biostatistics*, 3, 147–164. [511,514,517,518,523]

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffee, M., and Ten Have, T. R. (2009), "Mediation Analysis With Principal Stratification," *Statistics in Medicine*, 28, 1108–1130. [511,513]

Gelman, A. (1996), "Inference and Monitoring Convergence," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Boca Raton, FL: Chapman and Hall, pp. 131–143. [522]

Hafeman, D. M., and VanderWeele, T. J. (2011), "Alternative Assumptions for the Identification of Direct and Indirect Effects," *Epidemiology*, 22, 753–764. [510]

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., and 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. [512]

Hill, J., Waldfogel, J., and 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. [511]

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000), "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88. [511]

Howard, W. A., Omumbo, J., Nevill, C., Some, E. S., Donnelly, C. A., and 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, 357–360. [512]

Hudgens, M. G., and Halloran, M. E. (2008), "Towards Causal Inference With Interference," *Journal of the American Statistical Association*, 103, 832–842. [511]

Imai, K., Keele, L., and Tingley, D. (2010), "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 15, 309–334. [510]

Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71. [510]

Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476. [511]

Imbens, G. W., and Rubin, D. B. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," *Annals of Statistics*, 25, 305–327. [511]

Jin, H., and Rubin, D. B. (2008), "Principal Stratification for Causal Inference with Extended Partial Compliance," *Journal of the American Statistical Association*, 103, 101–111. [524]

Jo, B. (2008), "Causal Inference in Randomized Experiments With Mediational Processes," *Psychological Methods*, 13, 314–336. [511,517]

Jo, B., Asparouhov, T., and Muthén, B. O. (2008a), "Intention-to-Treat Analysis in Cluster Randomized Trials With Noncompliance," *Statistics in Medicine*, 27, 5565–5577. [511,517]

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., and Brown, C. H. (2008b), "Cluster Randomized Trials With Treatment Noncompliance," *Psychological Methods*, 13, 1–18. [511,517]

Lynch, K. G., Cary, M., Gallop, R., and Ten Have, T. R. (2008), "Causal Mediation Analysis for Randomized Trial," *Health Services and Outcomes Research Methodology*, 8, 57–76. [511]

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002), "A Comparison of Methods to Test Mediation and Other Intervening Variable Effects," *Psychological Methods*, 7, 83–104. [510]

Mattei, A., and Mealli, F. (2011), "Augmented Designs to Assess Principal Strata Direct Effects," *Journal of the Royal Statistical Society*, 73, 729–752. [511,513]

McDonald, C., Hiu, S., and Tierney, W. (1992), "Effects of Computer Reminders for Influenza Vaccination on Morbidity During Influenza Epidemics," *MD Computing*, 9, 304–312. [511]

Mealli, F., and Mattei, A. (2012), "A Refreshing Account of Principal Stratification," *The International Journal of Biostatistics*, 8, 246–254. [511,515]

Mealli, F., and 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. [511]

Morris, S., Flores, R., Olinto, P., and 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. [511]

Nevill, C. G., Some, E. S., Mung'ala, V. O., Mutemi, W., New, L., Marsh, K., Lengeler, C., and 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, 139–146. [512]

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. [511,513]

Pearl, J. (2001), "Direct and Indirect Effects," in *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, eds. J. S. Breese and D. Koller, San Francisco, CA: Morgan Kaufman, pp. 411–420. [510,515]

——— (2011), "Principal Stratification—A Goal or a Tool?," *International Journal of Biostatistics*, 7, Article 20. [511]

Robins, J. M., and Greenland, S. (1992), "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155. [510,515]

Rubin, B. D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non Randomized Studies," *Journal of Educational Psychology*, 66, 688–701. [510,512]

——— (1978), "Bayesian Inference for Causal Effects," *Annals of Statistics*, 6, 34–58. [510,513]

——— (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. [513]

——— (2004), "Direct and Indirect Causal Effects Via Potential Outcomes," *Scandinavian Journal of Statistics*, 31, 161–170. [510,511,516]

Schwartz, S. L., Li, F., and Mealli, F. (2011), "A Bayesian Semiparametric Approach to Intermediate Variables in Causal Inference," *Journal of the American Statistical Association*, 106, 1331–1344. [524]

Small, D. S. (2012), "Mediation Analysis Without Sequential Ignorability: Using Baseline Covariates Interacted With Random Assignment as Instrumental Variables," *Journal of Statistical Research*, 42, 89–101. [511]

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. [511]

Sommer, A., and Zeger, S. (1991), "On Estimating Efficacy From Clinical Trials," *Statistics in Medicine*, 10, 45–52. [511]

Tchetgen Tchetgen, E. J., and VanderWeele, T. J. (2012), "On Causal Inference in the Presence of Interference," *Statistical Methods in Medical Research*, 21, 55–75. [511]

TenHave, T. R., and Joffe, M. M. (2012), "A Review of Causal Estimation of Effects in Mediation Analyses," *Statistical Methods in Medical Research*, 21, 77–107. [511,516]

TenHave, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007), "Causal Mediation Analyses With Rank Preserving Models," *Biometrics*, 63, 926–934. [511]

VanderWeele, T. J. (2008), "Simple Relations Between Principal Stratification and Direct and Indirect Effects," *Statistics and Probability Letters*, 78, 2957–2962. [513,515]

——— (2010a), "Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects," *Epidemiology*, 21, 540–551. [511]

——— (2010b), "Direct and Indirect Effects for Neighborhood-Based Clustered and Longitudinal Data," *Sociological Research and Methods*, 38, 515–544. [515]

——— (2014), "A Unification of Mediation and Interaction: a Four-Way Decomposition," *Epidemiology*, 25, 749–761. [515]

VanderWeele, T. J., Hong, G., Jones, S. M., and 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, 469–482. [511]

VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Halloran, M. E. (2012), "Components of the Indirect Effect in Vaccine Trials: Identification of Contagion and Infectiousness Effects," *Epidemiology*, 23, 751–761. [511,512]