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

Laura Forastiere [*] [1], *Fabrizia Mealli*[1], *and Tyler J. VanderWeele*[2]

[1] *University of Florence* , [2] *Harvard School of Public Health*

Supplemental Material

1. Identifying Assumptions for Causal Mechanisms

Here we review the sequental ignorability assumption and delineate the conditions that would be needed in the particular setting of CEDs. We then provide a generalization of homogeneity assumptions (4) and (5), where each specification yields identification of $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ and the corresponding $iTME^{1-\tilde{a}}(m_0, m_1, \mathbf{c})$, with a specific value $\tilde{a} = 0, 1$ and for one of the two principal strata with $m_0 \neq m_1$. For each assumption we outline a comparison with sequential ignorability.

1.1 Sequential Ignorability

Sequential ignorability consists of two assumptions. We report here their expression in the setting of cluster-level interventions.

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

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

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

[*] Forastiere@disia.unifi.it

intermediate variable:

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

Substantially assumption (SI.I) rules out the presence of unmeasured confounders of the relationships of A_j with M_{ij} and Y_{ij}, while assumption (SI.II) prohibits unmeasured confounders of the relationships between A_j and Y_{ij} as well as measured or unmeasured confounders of the same relationships affected by the encouragement A_j. Assumptions (SI.I) and (SI.II) yield the following identification:

$$\mathrm{E}\left[Y_{ij}\left(a, M_{ij}(\tilde{a})\right) \mid \mathbf{C}_{ij} = \mathbf{c}\right] = \sum_{m=0}^{1} \mathrm{E}\left[Y_{ij} \mid A_j = a, M_{ij} = m, \mathbf{C}_{ij} = \mathbf{c}\right] \times P\left(M_{ij} = m \mid A_j = \tilde{a}, \mathbf{C}_{ij} = \mathbf{c}\right) \tag{1.1}$$

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

1.2 HOMOGENEITY ASSUMPTIONS

Assumption 4b. *Partial Stochastic Homogeneity of the Counterfactuals across Principal Strata*
Partial stochastic homogeneity of the counterfactuals across principal strata is said to be assumed if for specific values of $a, \tilde{a}, m \in \{0,1\}$ if the following conditional independence holds:

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

If assumption (4b) holds for a certain value of m and a certain value of $\tilde{a}$, with $a = \tilde{a}$, then the potential outcome $Y_{ij}(\tilde{a}, M_{ij}(\tilde{a}))$ is independent of $M_{ij}(1-\tilde{a})$, conditioning on levels of covariates $\mathbf{C}_{ij}$ and on strata where $M_{ij}(\tilde{a}) = m$. In this particular case the assumption can be supported from the data if the distribution of outcomes under encouragement status $A_j = \tilde{a}$, within levels of covariates, is the same for the two strata that share the same potential value of the treatment receipt $M_{ij}(\tilde{a}) = m$. When $a \neq \tilde{a}$, (4b) is an assumption on the distribution of potential outcomes of the form $Y_{ij}(a, M_{ij}(\tilde{a}))$, hence it is neither testable nor can find support in the data. However if assumption 4b holds for a certain value of m and a certain value of $\tilde{a}$, with $a = \tilde{a}$, we can also assume that it is valid for $a \neq \tilde{a}$.

The main result that follows from assumption (4b) is that if it is deemed valid for for specific values of $\tilde{a}$, a and m, then the two principal strata that share the same potential value $M_{ij}(\tilde{a}) = m$ present equal conditional mean of the potential outcome $Y_{ij}\big(a, M_{ij}(\tilde{a})\big)$:

$$\mathrm{E}\big[Y_{ij}\big(a, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = m, M_{ij}(1-\tilde{a}) = m_{1-\tilde{a}}, \mathbf{C}_{ij} = \mathbf{c}\big] = \mathrm{E}\big[Y_{ij}\big(1, M_{ij}(\tilde{a})\big) \mid M_{ij}(\tilde{a}) = M_{ij}(1-\tilde{a}) = m, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{1.2}$$

Theorem 1b. If assumption (4b) holds for $\tilde{a} = 0$, $a = 1$ and a specific value of $m \in \{0,1\}$, the net encouragement effect $NEE^0(m, m_1, \mathbf{c})$ for the principal stratum S^{mm_1}, with $M_{ij}(0) = m$ and $M_{ij}(1) = m_1 \neq m$, within levels of covariates, is given by:

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

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

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

If assumption (4b) holds for $\tilde{a} = 1$, $a = 0$ and a specific value of $m = 0, 1$, the net encouragement effect $NEE^1(m_0, m, \mathbf{c})$ for the stratum $S^{m_0 m}$, with $M_{ij}(0) = m_0 \neq m$ and $M_{ij}(1) = m$, within levels of covariates, is given by:

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

Consequently, the individual treatment mediated effect $iTME^0(m_0, m, \mathbf{c})$ for the stratum S^{mm_1}, with $M_{ij}(0) = m_0 \neq m$ and $M_{ij}(1) = m$, within levels of covariates, is given by the following difference:

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

Proof. We show here the proof for the first part of the theorem relative to NEE^0. The proof simply uses the implication of assumption (4b) shown in (1.2), concerning homogeneity in terms of conditional mean:

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

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

Corollary 1. *If assumption* (4b) *holds for* $\tilde{a} = 0$, $a = 1$ *and* $\forall m \in \{0,1\}$, *the population mean of the counterfactual*

$Y_{ij}(1, M_{ij}(0))$, *within levels of covariates, can be estimated using the following result:*

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

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

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

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

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

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

where second equality follows from assumption (4b) and the consequent homogeneity in (1.2) for the two strata sharing the same potential value $M_{ij}(0) = m$. The last equality uses the fact that $Y_{ij}(1, m) = Y_{ij}(1)$ for strata where $M_{ij}(1) = m$. □

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

Remark

Assumption (4b) differs from the assumption of conditional unconfoundedness of the treatment receipt in (SI.II) in a substantial way. (4b) assumes that, conditioning on levels of covariates, a potential outcome of the form $Y_{ij}\big(a, M_{ij}(\tilde{a})\big)$ only depends on one of the two potential values of the treatment receipt, precisely the one that we are assuming to keep fixed with the hypothetical intervention on M_{ij}, namely $M_{ij}(\tilde{a})$, and is instead independent of the other potential treatment receipt. On the contrary, the second assumption of sequential ignorability (SI.II) requires the independence of the potential outcome from both potential values of the treatment receipt, so that it makes possible to extrapolate information across strata relying on the observed, instead of the potential, values of the treatment received. This substantial difference can be better understood if we express the identification formula (1.1), following from the sequential ignorability, in terms of principal strata:

$$\mathrm{E}\big[Y_{ij}\big(1, M_{ij}(0)\big) \mid \mathbf{C}_{ij} = \mathbf{c}\big] = \sum_{m=0}^{1} \left(\sum_{m_0=m}^{1-m} \left(\mathrm{E}\big[Y_{ij}(1) \mid S_{ij} = S^{m_0 m}, \mathbf{C}_{ij} = \mathbf{c}\big] \frac{\pi_{m_0 m}(\mathbf{c})}{\pi_{mm}(\mathbf{c}) + \pi_{1-mm}(\mathbf{c})} \right) \sum_{m_1=m}^{1-m} \pi_{mm_1}(\mathbf{c}) \right) \tag{1.3}$$

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

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

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

by virtue of unconfoundedness of the encouragement assignement (SI.I)

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

by virtue of unconfoundedness of the treatment receipt (SI.II)

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

again by virtue of unconfoundedness of the encouragement assignement (SI.I)

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

again by virtue of unconfoundedness of the treatment receipt (SI.II)

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

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

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

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

□

If we now compare the identification result in corollary (1), yield by the homogeneity assumption (4b), with the identification result in equation (1.3), yield by the sequential ignorability assumptions (SI.I) and (SI.II), we can see that, in the latter, for all the strata where $M_{ij}(0) = m$, information on the mean of the counterfactual $Y_{ij}\big(1, M_{ij}(0)\big)$ for is taken from the mean value of the potential outcome $Y_{ij}(1)$ for those units where the potential value of the treatment received under $A_j = 1$, instead of $A_j = 0$, $M_{ij}(1)$, equals m. On the contrary, in (1), for the principal strata where $M_{ij}(0) = m$ and $M_{ij}(1) = m_1 \neq m$ information on the a priori counterfactual is borrowed just from those strata where $M_{ij}(0) = M_{ij}(1) = m$, who are the only ones for whom the mean value can be estimated from the data thanks to of the equality $Y_{ij}(1, M_{ij}(0)) \equiv Y_{ij}(1, M_{ij}(1)) \equiv Y_{ij}(1)$. For instance, when there are no defiers, this means to say that sequential ignorability allows to estimate $Y_{ij}(1, M_{ij}(0))$ for always-takers, where $M_{ij}(0) = 1$, not only from the values of $Y_{ij}(1) = Y_{ij}(1,1)$ for that sub-population but also borrowing information from the values of $Y_{ij}(1) = Y_{ij}(1,1)$ for compliers, whereas assumption (4b) does not use this extrapolation across these two strata.

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

Assumption 5b. *Partial Homogeneity of the Mean Difference between Counterfactuals across Principal Strata* Partial homogeneity of the mean difference between counterfactuals is said to be assumed if, for specific values of $\tilde{a} \in \{0,1\}$ and $m \in \{0,1\}$, the following identity holds:

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

In words, it states that the mean difference between potential outcomes under the two encouragement conditions and intervening to set the treatment receipt of each unit to the value it would take if A_j were set to $\tilde{a}$, i.e. $M_{ij}(\tilde{a}) = m$, is independent of the potential value of the treatment receipt under the opposite encouragement status, $M_{ij}(1-\tilde{a})$.

Theorem 2b. If assumption (5b) is satisfied for a certain value of $\tilde{a} \in \{0,1\}$ and a specific value of $m \in \{0,1\}$, the net encouragement effect $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$, within levels of covariates, for the principal stratum $S^{m_0 m_1}$ where $M_{ij}(\tilde{a}) = m_{\tilde{a}} = m$, is given by:

$$NEE^{\tilde{a}}(m_0, m_1, \mathbf{c}) \equiv DCE(m_0, \mathbf{c})(1-\tilde{a}) + DCE(m_1, \mathbf{c})(\tilde{a}) = DCE(m_{\tilde{a}}, \mathbf{c}) \tag{1.4}$$

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

Proof. The proof is accomplished by using the definition of $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ in (4.5):

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

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

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

Now the proof simply proceeds by applying assumption (5b) twice

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

□

Remark

Assumption (5b) differs from the assumption of conditonal ignorability of the treatment receipt in (SI.II) on three main provisions. First, the latter states a stochastic independence whereas the former is an assumption about independence in terms of the expected value. Second, conditonal ignorability of the treatment receipt concerns separately each counterfactual, whereas (5b) concerns a difference between pairs of counterfactuals. Third, a way to interpret (SI.II) is saying that the counterfactual $Y_{ij}(a, m)$ does not depend either on $M_{ij}(\tilde{a})$ or on $M_{ij}(1 - \tilde{a})$, conditioning on levels of covariates and the observed encouragement, so that information on $Y_{ij}(a, m)$, for for all units, can be extrapolated from $Y_{i'j'}(a)$ for all those units with $M_{i'j'}(a) = m$, regardless of the values of $M_{ij}(a)$, $M_{ij}(1 - a)$ and $M_{i'j'}(1 - a)$. Conversely, partial homogeneity assumption (5b) is solely based on the independence of the mean difference between potential outcomes $Y_{ij}(1, m)$ and $Y_{ij}(0, m)$ from $M_{ij}(1 - \tilde{a})$, conditioning on covariates but more important on $M_{ij}(\tilde{a}) = m$, with specific values of $a, \tilde{a}$ and m. This means that extrapolation across strata is only carried out for the a priori counterfactual $Y_{ij}(a, m)$ for those whose compliance behavior is given by $M_{ij}(\tilde{a}) = m$ and $M_{ij}(1-\tilde{a}) \neq m$ from $DCE(m, \mathbf{c})$ for the principal stratum with the same value m of treatment receipt under both encouragement conditions, i.e. $M_{ij}(\tilde{a}) = M_{ij}(1 - \tilde{a}) = m$. For these three reason we can conclude that assumption (5b) of partial homogeneity is a much weaker assumption that the second of the sequential ignorability assumptions. Mixing information across strata with the same behavior under a specific encouragement assignment seems more reasonable that mixing across all the principal strata, especially when these strata are most likely very different because of the presence of latent characteristics.

Furthermore, note that the first two differences between assumptions (5b) and (SI.II) also apply to a comparison between assumptions (5b) and (4b). Intuitively in general it is more plausible to assume homogeneity in terms of a mean difference rather that a stochastic homogeneity of each specific counterfactual.

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

Corollary 2. *If either assumption* (4b) *holds for a value of* $\tilde{a} = 0$ *and both* $a = 0$ *and* $a = 1$ *and* $\forall m \in 0,1$*, or assumption* (5b) *holds for a value of* $\tilde{a} = 0$ *and* $\forall m \in 0,1$*, the population net encouragement effect* $NEE^0(\mathbf{c})$*, within levels of covariates, is given by:*

$$NEE^0(\mathbf{c}) = \sum_{(m_0,m_1)} NEE^0(m_0,m_1,\mathbf{c})\pi_{m_0m_1}(\mathbf{c}) = \sum_{m=0}^{1}\left(DCE(m,\mathbf{c})\sum_{m_1=m}^{1-m}\pi_{mm_1}(\mathbf{c})\right) \tag{1.5}$$

If either assumption 4b holds for a value of $\tilde{a} = 1$ *and both* $a = 0$ *and* $a = 1$ *and* $\forall m \in 0,1$*, or assumption 5b holds for a value of* $\tilde{a} = 1$ *and* $\forall m \in 0,1$*, the population net encouragement effect* $NEE^1(\mathbf{c})$*, within levels of covariates, is given by:*

$$NEE^1(\mathbf{c}) = \sum_{(m_0,m_1)} NEE^1(m_0,m_1,\mathbf{c})\pi_{m_0m_1}(\mathbf{c}) = \sum_{m=0}^{1}\left(DCE(m,\mathbf{c})\sum_{m_0=m}^{1-m}\pi_{m_0m}(\mathbf{c})\right) \tag{1.6}$$

Proof. The proof of the corollary simply follows from equation (1.2) applied for the specified values of $a, \tilde{a}$ and m and from theorem 2b, by performing a weighted average over all four principal strata.

□

Both assumptions (4b) and (5b) provide the possibility of a generalization of the information on one potential outcome or the net encouragement effect from a stratum S^{mm} to the stratum $S^{m_0m_1}$ with $M_{ij}(\tilde{a}) = m$ and $M_{ij}(1-\tilde{a}) \neq m$, as stated by theorems (1b) and (2b). As a fair consequence of this generalization, the estimation of the individual treatment mediated effect for strata with $M_{ij}(0) \neq M_{ij}(1)$ in this stratum is straightforward and given by the difference between the estimated principal causal effect and net encouragement effect: $iTME^{1-\tilde{a}}(m_0,m_1,\mathbf{c}) = PCE(m_0,m_1,\mathbf{c}) - NEE^{\tilde{a}}(m_0,m_1,\mathbf{c})$.

Corollary 3. *If either assumption* (4b) *holds for a specific value of $\tilde{a}$,* $\forall m \in 0,1$ *and both* $a = 0$ *and* $a = 1$ *or assumption* (5b) *holds for a specific value of $\tilde{a}$ and* $\forall m \in 0,1$*, the individual treatment mediated effect in the whole population is given by the weighted sum over the compliers and the defiers, as reported in* (4.10).

$$iTME^{1-\tilde{a}}(\mathbf{c}) = \sum_{m_0 \neq m_1} \Big(PCE(m_0, m_1, \mathbf{c}) - DCE(m_{\tilde{a}}, \mathbf{c})\Big)\pi_{m_0 m_1}(\mathbf{c}) \tag{1.7}$$

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

2. Controlled Net Encouragement Effects within Principal Strata

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

$$CNEE^{m}(m_0, m_1\mathbf{c}) := \mathrm{E}\big[Y_{ij}(1,m) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] - \mathrm{E}\big[Y_{ij}(0,m) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big] \tag{2.1}$$

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

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

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

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

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

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

□

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

Corollary 4. *If either assumption* (4b) *hold for a specific value of* $\tilde{a} \in \{0,1\}$*, both* $a = 0$ *and* $a = 1$ *and a specific value of* $m \in \{0,1\}$*, or assumption* (5b) *holds for specific values of* $\tilde{a} \in \{0,1\}$ *and* $m \in \{0,1\}$*, then the controlled net encouragement effect, within level of covariates, for the stratum* $S^{m_0 m_1}$ *where*$M_{ij}(\tilde{a}) = m_{\tilde{a}} = m$ *and* $M_{ij}(1-\tilde{a}) = m_{1-\tilde{a}} \neq m$*, setting the treatment receipt to* $m_{\tilde{a}}$*, is equal to the corresponding controlled net encouragement effect for the stratum* $S^{m_{\tilde{a}} m_{\tilde{a}}}$ *where both* $Mij(\tilde{a}) = M_{ij}(1-\tilde{a}) = m_{\tilde{a}} = m$*.*

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

As a final result we can claim that, if assumptions (4b) or (5b) are satisfied for both encouragement conditions, $\tilde{a} = 0$ and $\tilde{a} = 1$, the controlled net encouragement effect $\mathit{CNEE}^{m}(m_0, m_1\mathbf{c})$ is the same for all the strata with at least one of the potential values $M_{ij}(0)$ or $M_{ij}(1)$ equal to m.

3. Average Treatment Effect

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

$$\begin{aligned} \mathit{ATE}^{a}(\mathbf{c}) &:= \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid \mathbf{C}_{ij} = \mathbf{c}\big] \\ &= \sum_{(m_0 = m_1)} \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big]\pi_{m_0 m_1} \\ &\quad + \sum_{(m_0 \neq m_1)} \mathrm{E}\big[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}\big]\pi_{m0 m_1} \end{aligned} \tag{3.1}$$

where the last expression simply expands the definition taking an average of the specific average treatment effects within the different principal strata. Referring to the definition in (3.1) we have to make two main considerations. First, we can see that the average treatment effects in general depends on the specific value of a we consider for the encouragement condition, while we compare the two scenarios where the treatment is or is not taken. The possible difference between $\mathrm{ATE}^{0}(\mathbf{c})$ and $ATE^{1}(\mathbf{c})$ is due to the interaction between the encouragement and the individual treatment uptake on the outcome. In clustered encouragements it can also be due to the interaction of the individual treatment uptake with other behavioral changes in other subjects in the same cluster. Second, unfortunately the empirical data do not provide any information on the treatment effect for principal strata where the treatment uptake is unaffected by the encouragement assignment if $M_{ij}(0) = M_{ij}(1) = 0$ because there is no individual information on the counterfactual $Y_{ij}(a,1)$ and vice versa for the symmetric stratum. The only strata where we could learn something about the treatment effect are those where $M_{ij}(0) \neq M_{ij}(1)$.

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

$$CACE^{a}(\mathbf{c}) := \mathrm{E}\left[Y_{ij}(a,1) - Y_{ij}(a,0) \mid S_{ij} = S^{01}, \mathbf{C}_{ij} = \mathbf{c}\right] \tag{3.2}$$

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

Nevertheless, when exclusion restriction assumptions are violated, if assumption (4) or (5) hold, the resulting identification of the individual treatment mediated effect $iTME^{1}(0,1,\mathbf{c})$ will also yield identification of $CACE^{1}(\mathbf{c})$, given the following equality:

$$CACE^{1}(\mathbf{c}) \equiv iTME^{1}(0,1,\mathbf{c}) \tag{3.3}$$

When we are evaluating a new treatment that cannot be randomized and we use the encouragement design as an instrument, the effect of primary interest is $CACE^{0}(\mathbf{c})$. In that case assumptions similar to (4) and

(5) are needed (see the supplemental material for a generalization of the assumptions). Alternatively, when the treatment effect has already been assessed in previous experiments, that is $CACE^0(\mathbf{c})$ is already known, and an encouragement, designed to increase or decrease its uptake, is the intervention of interest, the estimated $CACE^1(\mathbf{c})$ will give insight into how the encouragement itself changes the effect of the treatment on the outcome. This is the case when the treatment is the purchase of new bed nets.

4. Bayesian Inference

Let A_j^{obs} be the observed encouragement assigned to cluster j. Assuming that all the potentially observable information for each cluster is in the random vector $\left(A_j, \mathbf{C}_j, \mathbf{M}_j^{obs}, \mathbf{M}_j^{mis}, \mathbf{Y}_j^{obs}, \mathbf{Y}_j^{mis}\right)$, where each vector with subscript j contains the corresponding variable for all the units in cluster j, whereas we denote with superscript *obs* and *mis*, respectively, the observed and missing but observable potential outcomes, that is: $\mathbf{Y}_j^{obs} \equiv \mathbf{Y}_j(A_j)$, $\mathbf{Y}_j^{mis} \equiv \mathbf{Y}_j(1-A_j)$, $\mathbf{M}_j^{obs} \equiv \mathbf{M}_j(A_j)$ and $\mathbf{M}_j^{mis} \equiv \mathbf{M}_j(1-A_j)$. As extensively discussed, counterfactuals of the form $Y_{ij}\left(a, M_{ij}(\tilde{a})\right)$ are never observable unless $M_{ij}(\tilde{a}) \equiv M_{ij}(a)$. Under assumptions (4) or (5) presented above, all the causal estimands depend solely on the observable potential outcomes ${Y_{ij}}^{obs}$ and ${Y_{ij}}^{mis}$ of individuals belonging to each principal stratum. Therefore we can assume that all the missing information required for each cluster is contained in the vectors $\left(\mathbf{M}_j^{mis}, \mathbf{Y}_j^{mis}\right)$.

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

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

where we have collected each set of parameters such that $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{S^{00}}, \boldsymbol{\beta}^{S^{11}}, \boldsymbol{\beta}^{S^{01}}\right)$,

$\boldsymbol{b} = \left(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_J\right)$, and $\Sigma_a = \left(\Sigma_{a_n}, \Sigma_{a_c}\right)$.

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

$$p(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}, \mathbf{M}^{obs}, \ \mathbf{C}\ , \ \mathbf{A}) \propto \ p(\boldsymbol{\theta}) \int \int \prod_{j=1}^{J} p\big(\mathbf{Y}_j^{obs}, \mathbf{M}_j^{obs}, \mathbf{Y}_j^{mis}, \mathbf{M}_j^{mis}, \ \mathbf{C}_j \mid \boldsymbol{\theta}\big) d\mathbf{Y}_j^{mis} d\mathbf{M}_j^{mis} \tag{4.1}$$

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

$$p(\boldsymbol{\theta}, \mathbf{S} \mid \mathbf{Y}^{obs}, \ \mathbf{C}\ , \ \mathbf{A}) \propto \ p(\boldsymbol{\theta}) \prod_{j=1}^{J} p\big(\mathbf{Y}_j^{obs}, \mathbf{S}_j, \ \mathbf{C}_j \mid \boldsymbol{\theta}\big) \tag{4.2}$$

which follows from the assumed independence between the potential outcomes

The second term in (4.2) is the complete-data likelihood function, which results in the likelihood function of a finite mixture model with known membership, unlike the observed likelihood where the strata membership is unknown. The complete-data likelihood function, namely $\mathscr{L}(\boldsymbol{\theta}; \mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C}) := p(\mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C} \mid \boldsymbol{\theta})$, can be factorized in $p(\mathbf{Y}^{obs} \mid \mathbf{S}, \mathbf{C}\ \boldsymbol{\theta}) p(\mathbf{S} \mid \mathbf{C}, \ \boldsymbol{\theta}) p(\mathbf{C} \mid \boldsymbol{\theta})$. Letting $\delta_{ij}(S^{m_0 m_1}) = \delta(S^{m_0 m_1}, S_{ij})$ be 1 if $S_{ij} = S^{m_0 m_1}$ and 0 otherwise, we can write:

$$\mathscr{L}(\boldsymbol{\theta}; \mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C}) = \prod_{j=1}^{\mathrm{J}} \prod_{i=1}^{N_j} \sum_{m_0 m_1} \delta_{ij}(S^{m_0 m_1}) p\big(Y_{ij} \mid A_j, S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij}, \ \boldsymbol{\theta}\big) \times P\big(S_{ij} = S^{m_0 m_1} \mid \mathbf{C}_{ij}, \boldsymbol{\theta}\big) p(\mathbf{C}_{ij} \mid \boldsymbol{\theta}) \tag{4.3}$$

where assumption of consistency (1) has been used to express the distribution of the observed potential outcome in terms of the distribution of the observed values. The two models involved in the likelihood for Y_{ij} and S_{ij} have already been defined in (6.1) and (6.4) respectively. The complete-data likelihood allows the full conditional distributions $p(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}, \mathbf{S}, \mathbf{C}\ , \ \mathbf{A})$ and $p(\mathbf{S} \mid \mathbf{Y}^{obs}, \ \mathbf{C}\ , \ \mathbf{A}, \boldsymbol{\theta})$ to be analytically tractable. Therefore, the joint posterior distribution of $\big(\boldsymbol{\theta}, \mathbf{S}\big)$ motivates a two-stage Gibbs-sampling strategy that first samples the missing strata memberships S_{ij}, thereby allowing assessment of the distributions of Y_{ij}conditional on the complete data consisting of subpopulations without mixture components. This approach is well know as *Data Augmentation* scheme (Tanner & Wong, 1987). See the supplemental material for the detailed Gibbs-Sampling procedure.

4.1 PRIOR SPECIFICATION

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

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})\prod_j p(\mathbf{b}_j \mid \Sigma_b)p(\Sigma_b)p(\boldsymbol{\alpha}_n)p(\boldsymbol{\alpha}_c)\prod_j p(\mathbf{a}_{nj} \mid \Sigma_{a_n})p(\Sigma_{a_n})p(\mathbf{a}_{cj} \mid \Sigma_{a_c})p(\Sigma_{a_c}) \tag{4.4}$$

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

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{\mu}_{\beta 0}, \Lambda_{\beta 0}\right) \tag{4.5}$$

whereas the random effects are modeled independently for each cluster

$$\mathbf{b}_j \mid \Sigma_b \sim N(\mathbf{0}, \Sigma_b) \tag{4.6}$$

with the covariance matrices following an inverse-Wishart distribution:

$$\Sigma_b \sim IW\left(\eta_0^b, \eta_0^b S_0^b\right) \tag{4.7}$$

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

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

$$\boldsymbol{\alpha}_n \sim N\left(\boldsymbol{\mu}_{\alpha 0}^n, \Lambda_{\alpha 0}^n\right) \quad \boldsymbol{\alpha}_c \sim N\left(\boldsymbol{\mu}_{\alpha 0}^c, \Lambda_{\alpha 0}^c\right) \tag{4.8}$$

as well as for the random effects

$$\mathbf{a}_{nj} \mid \Sigma_{a_n} \sim N\left(\mathbf{0}, \Sigma_{a_n}\right) \quad \mathbf{a}_{cj} \mid \Sigma_{a_c} \sim N\left(\mathbf{0}, \Sigma_{a_c}\right) \tag{4.9}$$

with an inverse-Wishart prior for covariances matrices

$$\Sigma_{a_n} \sim IW\left(\eta_0^n, \eta_0^n S_0^n\right) \quad \Sigma_{a_c} \sim IW\left(\eta_0^c, \eta_0^c S_0^c\right) \tag{4.10}$$

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

4.2 Imputation Approach for Finite Population Effects

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

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

$$f_{m_0 m_1}(a, \mathbf{c}) = p\big(Y_{ij}(a) \mid S_{ij} = S^{m_0 m_1}, \mathbf{C}_{ij} = \mathbf{c}, \mathscr{O}\big) \tag{4.11}$$

At each iteration $k = 1, \ldots, K$ of the MCMC, samples from the posterior distribution of PCE for each principal stratum $S^{m_0 m_1}$ are drawn as follows:

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

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

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

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

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

Let us now turn to the analysis of mechanisms. As depicted in (4.7), for principal strata of the type S^{mm}, i.e. never-takers and always-takers, there is no effect through a change in the treatment received and principal causal effects are called dissociative causal effects, $\widehat{DCE}(m, \mathbf{c}) = \widehat{PCE}(m_0, m_1, \mathbf{c})$, as they are entirely net encouragement effects. On the contrary for the stratum S^{01} of compliers, which is, under monotonicity, the only stratum where the treatment is affected by the encouragement, the overall effect of the encouragement comprises both individual treatment effect from the net encouragement effect. With sequential ignorability (SI.II) not holding, disentangling these two effects for this stratum can be accomplished under one of the two assumptions (4) or (5). In general we can separate the derivation of $NEE^0(0, 1, \mathbf{c})$ into two three steps. The first two steps involve, respectively, the counterfactual $Y_{ij}(0) = Y_{ij}(0, M_{ij}(0))$ and $Y_{ij}(1, M_{ij}(0))$, whereas the third step concerns the mean difference.

3. For each unit being a complier at iteration k, the potential outcome $Y_{ij}^k(0)$ is derived as follows: if assumption (4) holds, $Y_{ij}^k(0)$ is simply taken from Y_{ij}^{obs} or Y_{ij}^{mis}, depending on A_j^{obs}; if assumption (5)

holds, in order to follow the identification result in theorem 2, ${Y_{ij}}^k(0)$ is imputed from the predictive posterior distribution of $Y_{ij}(0)$ for never-takers, given his values of covariates $\mathbf{C}_{ij}$:

$$
{Y_{ij}}^k(0) : \begin{cases} \text{3a. } \textit{if assumption } 4\text{: } {Y_{ij}}^k(0) = {Y_{ij}}^{obs} \cdot (1 - A_j^{obs}) + {Y_{ij}}^{k,mis} \cdot A_j^{obs} \\ \\ \text{3b. } \textit{if assumption } 5\text{: } {Y_{ij}}^k(0) \sim f_{00}(0, \mathbf{C}_{ij}) \end{cases} \qquad \forall i,j : S_{ij}^k = S^{01}
$$

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

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

5. $NEE^{k,0}$ for compliers is computed by taking the average, within levels of covariates, of the difference between the two imputed potential outcomes:

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

Again subgroup analysis based on covariates might require some restrictions.

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

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

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

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

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

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

The first part of the algorithm concerns the imputation of potential outcomes and hence of causal estimands from their posterior predictive distributions. Imputation of principal causal effects, net encouragement effects and individual treatment effects follows the procedure outlined in section 4.2 under assumption (4b) (5b).

1. The missing outcome ${Y_{ij}}^{mis} = Y_{ij}1 - A_j$ for each unit is drawn from the likelihood function $f_{m_0 m_1}(1 - A_j \mid \mathbf{C}_{ij}, \boldsymbol{\theta}^{k})$, as defined by the model (6.1). In addition, for each complier, i.e. with strata indicator $S_{ij} = S^{01}$, we draw two random samples, ${Y_{ij}}^{k}(\tilde{a})$ and ${Y_{ij}}^{k}(1-\tilde{a}, M_{ij}(\tilde{a}))$, as described in section 4.2. Finally, $PCE(m_0, m_1, \mathbf{c})$ and $NEE^{\tilde{a}}(m_0, m_1, \mathbf{c})$ for all three individual principal strata and $iTME^{1-\tilde{a}}(0, 1, \mathbf{c})$ for compliers are derived.
2. The vector of parameters $\boldsymbol{\beta}$ of the outcome model is drawn from its full conditional distribution $p\left(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{S}, \mathbf{Z}^{Yf}, \mathbf{Z}^{Yr}, \boldsymbol{b}\right)$. This is accomplished by a random walk Metropolis-Hastings algorithm with a normal proposal distribution, whose covariance matrix is a scaled version of an initial estimate.
3. Cluster-specific $\boldsymbol{b}_j$ are drawn independently for each cluster from their posterior distribution $p\left(\boldsymbol{b}_j \mid \boldsymbol{\beta}, \mathbf{Y}_j, \mathbf{Z}_j^{Yr}, \mathbf{Z}_j^{Yf}\right)$. Another step of random walk Metropolis-Hastings is used for the purpose, with a normal proposal distribution, a likelihood derived from the binomial regression model in (6.1) and (6.2) and a normal prior distribution given in (4.6), where the prior covariance matrix Σ_b is drawn at the previous iteration from its own posterior distribution.
4. The drawing of the covariance matrix Σ_b of the random effects is from the Inverse-Wishart posterior distri-

bution, derived as the posterior distribution of a covariance matrix of multivariate normal random variable, $\boldsymbol{b}_j$ in this case, with Inverse-Wishart prior as defined in (4.7).

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

5. The vectors of parameters $\boldsymbol{\alpha}_n$ and $\boldsymbol{\alpha}_c$ of the strata membership model are drawn independently from their normal posterior distribution $p\left(\boldsymbol{\alpha}_n \mid \mathbf{S}^n, \mathbf{Z}^{Sf}, \mathbf{Z}^{Sr}, \mathbf{a}_n\right)$ and $p\left(\boldsymbol{\alpha}_c \mid \mathbf{S}^c, \mathbf{Z}^{Sf}, \mathbf{Z}^{Sr}, \mathbf{a}_c\right)$ computed from their likelihood resulting from the linear models of the latent variables S_{ij}^n and S_{ij}^c in (6.5), and their prior distributions in (4.8). This time Bayesian regression are run with offsets $\mathbf{a}_{nj}^T Z_{ij}^{Sr}$ and $\mathbf{a}_{cj}^T Z_{ij}^{Sr}$ respectively.
6. According to distributional assumptions presented above, cluster-specific random effects $\mathbf{a}_{nj}$ and $\mathbf{a}_{cj}$ are drawn independently for each cluster from their normal posterior distributions $p\left(\mathbf{a}_{nj} \mid \mathbf{S}_j^n, \mathbf{Z}_j^{Sr}, \mathbf{Z}_j^{Sf}, \boldsymbol{\alpha}_n\right)$ and $p\left(\mathbf{a}_{cj} \mid \mathbf{S}_j^c, \mathbf{Z}_j^{Sr}, \mathbf{Z}_j^{Sf}, \boldsymbol{\alpha}_c\right)$ derived from the linear regression model in (6.5), this time with offsets $\boldsymbol{\alpha}_n^T \mathbf{Z}_{ij}^{Sf}$ and $\boldsymbol{\alpha}_c^T \mathbf{Z}_{ij}^{Sf}$, and normal prior distribution given in (4.9), where the prior covariance matrices Σ_{a_n} and Σ_{a_c} come from the previous iteration.
7. As with outcome random effects, the drawing of the covariance matrices of the strata model random effects, Σ_{a_n} and Σ_{a_c} , is from the Inverse-Wishart posterior distributions $p(\Sigma_{a_n} \mid \mathbf{a}_n)$ and $p(\Sigma_{a_c} \mid \mathbf{a}_c)$, derived as the posterior distribution of a covariance matrix of multivariate normal random variable, in this case $\mathbf{a}_{nj}$ and $\mathbf{a}_{cj}$, with Inverse-Wishart prior as defined in (4.10).
8. Given the fixed effects, the random effects and the observed data, the vector of latent strata membership $\mathbf{S}$ has to be generated from its full conditional distribution $p(\mathbf{S} \mid \mathbf{Y}, \mathbf{M}, \mathbf{A}, \mathbf{C}, \boldsymbol{\theta})$, which this time depends as well on the vector of individual mediator $\mathbf{M}$ being the principal stratum defined based on the potential mediators. This is the typical data augmentation step of the principal strata framework. As far as the average effects for each individual principal stratum are concerned, within each cluster the strata memberships of the unit are independent and hence strata indicators can be drawn independently from the conditional

distribution factorized as:

$$
\begin{aligned}
p(S_{ij} = & S^{m_0 m_1} \mid Y_{ij}, M_{ij}, A_j, \mathbf{C}_{ij}, \boldsymbol{\theta}) \\
&= \frac{p\left(Y_{ij} \mid S_{ij} = S^{m_0 m_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m_0 m_1}}, \mathbf{b}_j^{S^{m_0 m_1}}\right) p(S_{ij} = S^{m_0 m_1} \mid M_{ij}, A_j, \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a})}{\sum_{S^{m'_0 m'_1}} p(Y_{ij} \mid S_{ij} = S^{m'_0 m'_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m'_0 m'_1}}, \mathbf{b}_j^{S^{m'_0 m'_1}}) p(S_{ij} = S^{m'_0 m'_1} \mid M_{ij}, A_j, \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a})} \\
&= \frac{p\left(Y_{ij} \mid S_{ij} = S^{m_0 m_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m_0 m_1}}, \mathbf{b}_j^{S^{m_0 m_1}}\right) p(S_{ij} = S^{m_0 m_1} \mid \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a}) I\big(M_{ij}(A_j) = M_{ij}\big)}{\sum_{S^{m'_0 m'_1}} p(Y_{ij} \mid S_{ij} = S^{m'_0 m'_1}, A_j, \mathbf{C}_{ij}, \boldsymbol{\beta}^{S^{m'_0 m'_1}}, \mathbf{b}_j^{S^{m'_0 m'_1}}) p(S_{ij} = S^{m'_0 m'_1} \mid \mathbf{C}_{ij}, \boldsymbol{\alpha}, \mathbf{a}) I\big(M_{ij}(A_j) = M_{ij}\big)}
\end{aligned}
\tag{4.12}
$$

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

9.

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

$$
\begin{aligned}
S^n_{ij} &\sim \begin{cases} N_-(\boldsymbol{\alpha}_n Z^{Sf}_{ij} + \mathbf{a}^T_{nj} Z^{Sr}_{ij}, 1) I(S^n_{ij} \leq 0) & \text{if } S_{ij} = S^{00}_{ij} \\ N_+(\boldsymbol{\alpha}_n Z^{Sf}_{ij} + \mathbf{a}^T_{nj} Z^{Sr}_{ij}, 1) I(S^n_{ij} > 0) & \text{if } S_{ij} = S^{01}_{ij} \text{ or } S_{ij} = S^{11}_{ij} \end{cases} \\
S^c_{ij} &\sim \begin{cases} N(\boldsymbol{\alpha}_c Z^{Sf}_{ij} + \mathbf{a}^T_{cj} Z^{Sr}_{ij}, 1) & \text{if } S_{ij} = S^{00}_{ij} \\ N_-(\boldsymbol{\alpha}_c Z^{Sf}_{ij} + \mathbf{a}^T_{cj} Z^{Sr}_{ij}, 1) I(S^c_{ij} \leq 0) & \text{if } S_{ij} = S^{01}_{ij} \\ N_+(\boldsymbol{\alpha}_c Z^{Sf}_{ij} + \mathbf{a}^T_{cj} Z^{Sr}_{ij}, 1) I(S^c_{ij} > 0) & \text{if } S_{ij} = S^{11}_{ij} \end{cases}
\end{aligned}
\tag{4.13}
$$

5. PROOFS OF OTHER EQUATIONS

PROOF OF EQUATION 4.7

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

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

With similar manipulations we yield the second result:

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

PROOF OF EQUATION 4.8

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

PROOF OF EQUATION 3.3 (Supplemental Material)

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