Received: 23 December 2021 | Accepted: 7 October 2022

DOI: 10.1111/biom.13782

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC METHODOLOGY

Statistical inference and power analysis for direct and spillover effects in two-stage randomized experiments

Zhichao Jiang[1] | Kosuke Imai[2] | Anup Malani[3,4]

[1]School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong, China

[2]Department of Government and Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

[3]Law School and Pritzker School of Medicine, University of Chicago, Chicago, Illinois, USA

[4]National Bureau of Economic Research, Cambridge, Massachusetts, USA

Correspondence
Zhichao Jiang, School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong 510275, China.
Email: jiangzhch7@mail.sysu.edu.cn

Funding information
Alfred P. Sloan Foundation, Grant/Award Number: 2020-13946

Abstract

Two-stage randomized experiments become an increasingly popular experimental design for causal inference when the outcome of one unit may be affected by the treatment assignments of other units in the same cluster. In this paper, we provide a methodological framework for general tools of statistical inference and power analysis for two-stage randomized experiments. Under the randomization-based framework, we consider the estimation of a new direct effect of interest as well as the average direct and spillover effects studied in the literature. We provide unbiased estimators of these causal quantities and their conservative variance estimators in a general setting. Using these results, we then develop hypothesis testing procedures and derive sample size formulas. We theoretically compare the two-stage randomized design with the completely randomized and cluster randomized designs, which represent two limiting designs. Finally, we conduct simulation studies to evaluate the empirical performance of our sample size formulas. For empirical illustration, the proposed methodology is applied to the randomized evaluation of the Indian National Health Insurance Program. An open-source software package is available for implementing the proposed methodology.

KEYWORDS

experimental design, interference between units, partial interference, spillover effects, statistical power

1 | INTRODUCTION

Much of the early causal inference literature relied upon the assumption that the outcome of one unit cannot be affected by the treatment assignment of another unit. Over the last two decades, however, researchers have made substantial progress by developing a variety of methodological tools to relax this assumption (e.g., Aronow & Samii, 2017; Forastiere et al., 2016; Hudgens & Halloran, 2008; Imai et al., 2021; Tchetgen Tchetgen & VanderWeele, 2012).

Two-stage randomized experiments have become increasingly popular when studying spillover effects. Under this experimental design, researchers first randomly assign clusters of units to different treatment assignment mechanisms, each of which has a different probability of treatment assignment. For example, one treatment assignment mechanism may randomly assign 80% of units to the treatment group, whereas another mechanism may only treat 40%. Then, within each cluster, units are randomized to the treatment and control conditions according to its selected treatment assignment mechanism. By comparing units who are assigned to the same treatment conditions but belong to different clusters with different treatment assignment mechanisms, one can infer how the treatment conditions of other units within the same cluster affect one's outcome. Two-stage randomized experiments are now frequently used in a number of disciplines, including economics (e.g., Angelucci & Di Maro, 2016), education (e.g., Rogers & Feller, 2018), political science (e.g., Sinclair et al., 2012), and public health (e.g., Benjamin-Chung et al., 2018).

The increasing use of two-stage randomized experiments in applied scientific research calls for the development of a general methodology for analyzing and designing such experiments. Building on the prior literature (e.g., Basse and Feller, 2018; Hudgens and Halloran, 2008; Imai et al., 2021), we consider various direct and spillover effects, and develop their unbiased point estimators and conservative variance estimators under the nonparametric randomization-based framework. This framework has also been used to study other types of randomized designs (e.g., Balzer et al., 2015, 2016). We also show how to conduct hypothesis tests and derive the sample size formulas for the estimation of these causal effects. The resulting formulas can be used to conduct power analysis when designing two-stage randomized experiments. Finally, we theoretically compare the two-stage randomized design with its two limiting designs, the completely randomized and cluster randomized designs. Through this comparison, we analyze the potential efficiency loss of the two-stage randomized design when no spillover effect exists.

We make several methodological contributions. First, the proposed causal quantities generalize those of Hudgens and Halloran (2008) to more than two treatment assignment mechanisms. We consider the joint estimation of the average direct and spillover effects to characterize the causal heterogeneity across different treatment assignment mechanisms. We also propose the average marginal direct effect (MDE) as a scalar summary of several average direct effects (ADE). Second, our variance estimators are guaranteed to be conservative, while those of Hudgens and Halloran (2008) are not when applied to our setting. Third, we develop hypothesis testing procedures and sample size formulas, which can be used when planning a two-stage randomized experiment. Fourth, we prove the equivalence relationships between the proposed randomization-based estimators and the least-square estimators. These results extend those of Basse and Feller (2018), in which the clusters have at most one treated unit. Finally, an open-source software package is available for implementing the proposed methodology (Huang et al., 2022).

In a closely related article, Baird et al. (2018) adopted a super population framework to study the randomized saturation design (a general form of two-stage randomized experiments), in which the proportion of treated units for each cluster is drawn from a distribution. The authors consider the assumptions about the structure of spillover effects that are similar to those made in this paper. However, Baird et al. imposed a specific variance–covariance structure for potential outcomes and derived the standard errors of the causal estimates from a saturated linear model. In contrast, we adopt the nonparametric randomization-based framework without imposing any variance–covariance structure for the potential outcomes although we consider simplifying conditions to facilitate the use of our method in practice. In addition, while their goal is to determine the optimal distribution of the treated proportion, we treat this distribution to be fixed and focus on the development of estimators, hypothesis testing procedures, and sample size formulas.

The remainder of the paper is organized as follows. Section 2 introduces our motivating study concerning the impact evaluation of the Indian National Health Insurance Program (Imai et al., 2021; Malani et al., 2021). Section 3 formally presents the two-stage randomized design and defines the three causal quantities of interest. In Section 4, we propose a methodology for statistical inference and power analysis. Section 5 revisits the health insurance study and applies the proposed methods. Finally, Section 6 provides concluding remarks. The Appendix presents simulation studies, establishes the equivalence relations between the regression-based and randomization-based inference, and compares the two-stage randomized design with the cluster and individual randomized designs. All proofs appear in the Web Appendix.

2 | RANDOMIZED EVALUATION OF THE INDIAN HEALTH INSURANCE PROGRAM

We describe the randomized evaluation of the Indian National Health Insurance Program, which serves as our motivating application. In 2008, the Indian government introduced its first national public health insurance scheme, Rastriya Swasthya Bima Yojana (RSBY). The goal was to provide insurance coverage for hospitalization to households below the poverty line. Subsequently, the government considered the expansion of the RSBY to some households above the poverty line.

We conducted a randomized control trial to assess whether the expansion of the RSBY increases access to hospitalization, and thus health. The experiment took place in two districts of Karnataka State, Gulbarga and Mysore. Gulbarge has a total of 918 villages with the village size varying from 0 to 2,428, while Mysore has 1,336 villages with the size ranging from 0 to 2,976. We selected 22% and 16% of the villages in Gulbarga and Mysore, respectively. This led to 11,089 households who had no pre-existing health insurance coverage and lived within 25 km of an

TABLE 1 The two-stage randomized design for the evaluation of the Indian Health Insurance Program

	Treatment assignment mechanisms		
	1	2	3
Treatment assignment proportion	90%	70%	50%
Number of villages	285	88	63
Number of households	5512	1553	1170

RSBY empaneled hospital. The households in the treatment group were offered an opportunity to enroll in the RSBY, whereas those in the control group were able to buy the RSBY at the usual government price.

The evaluation was conducted using the two-stage randomized design shown in Table 1. In the first stage, a total of 436 villages are randomly assigned to three treatment assignment mechanisms, yielding 258, 88, and 63 villages for treatment assignment mechanisms 1, 2, and 3, respectively. Treatment assignment mechanisms 1, 2, and 3 correspond to the treatment assignment probabilities of 90%, 70%, and 50%, respectively. In the second stage of randomization, households were assigned to the treatment within each village according to the treatment assignment probability chosen in the first stage. Households were informed of the opportunities to enroll in RSBY from April to May, 2015. Approximately 18 months later, we carried out a survey and measured a variety of outcomes about the health and financial conditions of the household members. For more details about the experiment, see Imai et al. (2021), and Malani et al. (2021).

Both direct and spillover effects are of interest. The direct effect quantifies how much the household members would benefit from their own receipt of the program benefits. In contrast, the spillover effect characterizes how the treatment of other households affects one's outcomes, possibly through the replacement of informal insurance by formal insurance and the efficient use of limited resources in local hospitals. Moreover, the heterogeneity in the direct and spillover effects is also of interest. For example, a greater treatment assignment probability may cause the overcrowding of local hospitals, leading to a lower direct effect.

3 | EXPERIMENTAL DESIGN AND CAUSAL QUANTITIES OF INTEREST

We now formally describe the two-stage randomized experimental design and define the causal quantities of interest using the potential outcomes framework (e.g., Neyman, 1923; Rubin, 1974).

3.1 | Assumptions

Suppose that we have a total of J clusters and each cluster j has n_j units. Let N represent the total number of units, that is, $N = \sum_{j=1}^{J} n_j$. Under the two-stage randomized design, we first randomly assign clusters to different treatment assignment mechanisms, and then assign a certain proportion of individual units within a cluster to the treatment condition by following the treatment assignment mechanism selected at the first stage of randomization. Let A_j denote the treatment assignment mechanism chosen for cluster j, which takes a value in $\mathcal{M} = \{1, 2, \ldots, m\}$. Let $\boldsymbol{A} = (A_1, A_2, \ldots, A_J)$ denote the vector of treatment assignment mechanisms for all J clusters and $\boldsymbol{a} = (a_1, a_2, \ldots, a_J)$ represent the vector of realized assignment mechanisms. We assume complete randomization such that a total of J_a clusters are assigned to the assignment mechanism $a \in \mathcal{M}$, where $\sum_{a=1}^{m} J_a = J$.

The second stage of randomization concerns the treatment assignment for each unit within cluster j based on the assignment mechanism A_j. Let Z_{ij} be the binary treatment assignment variable for unit i in cluster j, where $Z_{ij} = 1$ and $Z_{ij} = 0$ imply that the unit is assigned to the treatment and control conditions, respectively. Let $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{n_j j})$ be the vector of assigned treatments for the n_j units in the cluster and $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{n_j j})$ be the vector of realized assignments. Then, $\Pr(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid A_j = a)$ represents the distribution of the treatment assignment, when cluster j is assigned to the assignment mechanism $A_j = a$. We assume complete randomization such that a total of n_{jz} units in cluster j are assigned to the treatment condition $z \in \{0, 1\}$ where $n_{j0} + n_{j1} = n_j$. Finally, let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_J)$ be the vector of assigned treatments for all the N units in the population and $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_J)$ be the vector of realized assignments. We now formally define the two-stage randomized design.

Assumption 1. (Two-stage randomization)

(1) Complete randomization of treatment assignment mechanisms across clusters:

$$\Pr(\boldsymbol{A} = \boldsymbol{a}) = \frac{J_1! \cdots J_m!}{J!}$$

for all $\boldsymbol{a}$ such that $\sum_{j=1}^{J} \mathbf{1}(a_j = a') = J_{a'}$ for $a' \in \mathcal{M}$.

(2) Complete randomization of treatment assignment across units within each cluster:

$$\Pr(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid A_j = a) = \frac{1}{\binom{n_j}{n_{j1}}}$$

for all $\boldsymbol{z}_j$ such that $\sum_{i=1}^{n_j} z_{ij} = n_{j1}$.

Next, we introduce the potential outcomes. For unit i in cluster j, let $Y_{ij}(\boldsymbol{z})$ be the potential value of the outcome if the assigned treatment vector for the entire sample is $\boldsymbol{z}$, where $\boldsymbol{z}$ is an N-dimensional vector. The observed outcome is given by $Y_{ij} = Y_{ij}(\boldsymbol{Z})$. This notation implies that the outcome of one unit may be affected by the treatment assignment of any other unit in the sample.

Unfortunately, it is impossible to learn about causal effects without additional assumptions because each unit has 2^N possible potential outcome values. Thus, following the literature (Hudgens and Halloran, 2008; Sobel, 2006), we assume that the potential outcome of one unit cannot be affected by the treatment assignment of another unit in other clusters while allowing for possible interference between units within a cluster.

Assumption 2 (No interference between clusters).

$$Y_{ij}(\boldsymbol{z}) = Y_{ij}(\boldsymbol{z}') \quad \text{for any } \boldsymbol{z}, \boldsymbol{z}' \text{ with } \boldsymbol{z}_j = \boldsymbol{z}'_j.$$

Assumption 2, which is known as the partial interference assumption in the literature, partially relaxes the standard assumption of no interference between units (Rubin, 1990). This assumption reduces the number of potential outcome values for each unit from 2^N to 2^{n_j}.

Lastly, we rely upon the stratified interference assumption proposed by Hudgens and Halloran (2008) to further reduce the number of potential outcome values.

Assumption 3 (Stratified interference).

$$Y_{ij}(\boldsymbol{z}_j) = Y_{ij}(\boldsymbol{z}'_j) \quad \text{if} \quad z_{ij} = z'_{ij} \text{ and } \sum_{i=1}^{n_j} z_{ij} = \sum_{i=1}^{n_j} z'_{ij}.$$

Assumption 3 implies that the outcome of one unit depends on the treatment assignment of other units only through the number of those who are assigned to the treatment condition within the same cluster. The assumption has been commonly used in the literature (e.g., Liu & Hudgens, 2014; Miles et al., 2019; Tchetgen Tchetgen & VanderWeele, 2012). It is a reasonable simplification of the interference structure and is directly motivated by two-stage randomization which varies the proportion of treated units within a cluster.

As pointed out by Hudgens and Halloran (2008), although the identification of the direct and indirect effects does not require Assumption 3, a valid variance estimator is unavailable without an additional assumption. A more general form of Assumption 3 is exposure mappings, which require the potential outcome to depend on a known function of treatment conditions (e.g., Bargagli Stoffi et al., 2020; Forastiere et al., 2016, 2021; Sävje et al., 2021; VanderWeele et al., 2013). While it is relatively straightforward to extend our results regarding statistical inference under such settings (Aronow & Samii, 2017), sample size and power calculation will be more complicated. Therefore, we maintain Assumption 3 throughout this paper. Under Assumptions 2 and 3, we can simplify the potential outcome as a function of one's own treatment and the treatment assignment mechanism of its cluster, that is, $Y_{ij}(\boldsymbol{z}) = Y_{ij}(z, a)$.

3.2 | Direct effect

Under the above assumptions, we now define the main causal quantities of interest. The first quantity is the direct effect of the treatment on one's own outcome. We define the unit-level direct effect for unit i in cluster j as

$$\mathrm{ADE}_{ij}(a) = Y_{ij}(1, a) - Y_{ij}(0, a)$$

for $a = 1, \ldots, m$. This quantity may depend on the treatment assignment mechanism a due to the possible spillover effect from other units' treatments. The direct effect quantifies how the treatment of a unit affects its outcome under a specific assignment mechanism. This unit-level direct effect can be aggregated, leading to the definition of the cluster-level direct effect,

$$\mathrm{ADE}_j(a) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathrm{ADE}_{ij}(a) = \overline{Y}_j(1, a) - \overline{Y}_j(0, a),$$

where $\overline{Y}_j(z, a) = 1/n_j \cdot \sum_{i=1}^{n_j} Y_{ij}(z, a)$. We can further aggregate this quantity and obtain the population-level direct effect,

$$\mathrm{ADE}(a) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{ADE}_j(a) = \overline{Y}(1, a) - \overline{Y}(0, a), \qquad (1)$$

where $\overline{Y}(z, a) = 1/J \cdot \sum_{j=1}^{J} \overline{Y}_j(z, a)$. The direct effects depend on the treatment assignment mechanisms; we denote them by a column vector, $\mathrm{ADE} = (\mathrm{ADE}(1), \ldots, \mathrm{ADE}(m))^\top$.

3.3 | Marginal direct effect

With m treatment assignment mechanisms, we have a total of m direct effects $\mathrm{ADE}(a)$ for $a = 1, \ldots, m$. Although such direct effects are informative about how the treatment of a unit affects its own outcome given different treatment assignment mechanisms, researchers may be interested in having a single quantity that summarizes all the direct effects. We define the unit-level MDE by marginalizing the direct effects over the treatment assignment mechanisms,

$$\mathrm{MDE}_{ij} = \sum_{a=1}^{m} q_a\{Y_{ij}(1,a) - Y_{ij}(0,a)\}.$$

The weight q_a is the proportion of the clusters assigned to treatment assignment mechanism a, which equals J_a/J under Assumption 1. Based on the unit-level effect, we define the cluster-level MDE and the population-level MDE as

$$\mathrm{MDE}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathrm{MDE}_{ij}, \quad \mathrm{MDE} = \frac{1}{J}\sum_{j=1}^{J}\mathrm{MDE}_j. \qquad (2)$$

We emphasize that the MDE, unlike the ADE, depends on the distribution of treatment assignment mechanism q_a. Thus, a different value of the design parameter can alter the interpretation of MDE.

3.4 | Spillover effect

In two-stage randomized experiments, another causal quantity of interest is the spillover effect, which quantifies how one's treatment affects the outcome of another unit. Under Assumptions 2 and 3, we define the unit-level spillover effect on the outcome as

$$\mathrm{ASE}_{ij}(z;a,a') = Y_{ij}(z,a) - Y_{ij}(z,a'),$$

which compares the potential outcomes under two different assignment mechanisms, a and a', while holding one's treatment assignment constant at z. We then define the spillover effects on the outcome at the cluster and population levels,

$$\mathrm{ASE}_j(z;a,a') = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathrm{ASE}_{ij}(z;a,a'),$$

$$ASE(z;a,a') = \frac{1}{J}\sum_{j=1}^{J}\mathrm{ASE}_j(z;a,a').$$

The spillover effects depend on both the treatment condition and treatment assignment mechanisms; we denote them by $\mathrm{ASE} = (\mathrm{ASE}(1;1,2), \mathrm{ASE}(1;2,3), \dots, \mathrm{ASE}(1;m-1,m), \quad \mathrm{ASE}(0;1,2), \mathrm{ASE}(0;2,3), \dots, \mathrm{ASE}(0;m-1,m))$, which consists of the spillover effects comparing adjacent treatment assignment mechanisms for both the treatment and control conditions.

We give equal weight to each cluster in the quantities defined above (see Hudgens & Halloran, 2008), while Basse and Feller (2018) assigned an equal weight to each unit. For example, Basse and Feller (2018) defined the direct effect as

$$\mathrm{ADE}(a) = \sum_{j=1}^{J}\frac{n_j}{N}\cdot\mathrm{ADE}_j(a) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathrm{ADE}_{ij}(a).$$

While our analysis focuses on the cluster-weighted quantities rather than individual-weighted quantities, our method can be generalized to any weighting scheme.

4 | A GENERAL METHODOLOGY FOR TWO-STAGE RANDOMIZED EXPERIMENTS

We next develop a general methodology for the direct and spillover effects introduced above. We show how to estimate these quantities, compute the randomization-based variance, and conduct hypothesis tests. We also derive the sample size formulas for testing the direct and spillover effects.

Formally, define $\overline{Y} = (\overline{Y}(1,1), \overline{Y}(0,1), \dots, \overline{Y}(1,m), \overline{Y}(0,m))^\top$, which is a $2m$-dimensional column vector with the $(2a-1)$th and $2a$th elements representing the treatment and control potential outcomes under treatment assignment mechanism a, respectively, for $a = 1, \dots, m$. The direct, marginal direct, and spillover effects can all be written as linear transformations of $\overline{Y}$. Our methodological development will exploit these linear transformations.

In particular, let e_l denote the $2m$-dimensional column vector whose lth element is equal to 1 with other elements being equal to 0. Then, the direct effect can be written as $\mathrm{ADE} = C_1\overline{Y}$, where $C_1 = (e_1 - e_2, e_3 - e_4, \dots, e_{2m-1} - e_{2m})^\top$ is an $m \times 2m$ matrix with the ath row representing the contrast in $\mathrm{ADE}(a)$ for $a = 1, \dots, m$. Similarly, the MDE can be written as $\mathrm{MDE} = C_2\overline{Y}$, where $C_2 = (J_1, -J_1, J_2, -J_2, \dots, J_m, -J_m)^\top/J$. Lastly, the spillover effect can be written as $\mathrm{ASE} = C_3\overline{Y}$, where $C_3 = (C_{31}, C_{30})^\top$ with $C_{31} = (e_1 - e_3, e_3 - e_5, \dots, e_{2m-3} - e_{2m-1})^\top$ and $C_{30} = (e_2 - e_4, e_4 - e_6, \dots, e_{2m-2} - e_{2m})^\top$. That is, the ath column in C_{31} and C_{30} represents the contrast in $\mathrm{ASE}(1;a,a+1)$ and $\mathrm{ASE}(0;a,a+1)$, respectively, for $a = 1, \dots, m-1$.

Under Assumptions 2 and 3, our setting is similar to a split-plot design in the sense that the treatment and the treatment assignment mechanism can be viewed as the interventions at the sub-plot and whole-plot levels, respectively. Therefore, we leverage this connection and use the results in the split-plot design developed in Zhao and Ding (2022) to obtain the unbiased estimation, variances of the estimators, and asymptotic properties of the estimators. We then develop hypothesis testing procedures and sample size formulas based on these results.

4.1 | Unbiased estimation

Hudgens and Halloran (2008) proposed unbiased estimators of the average direct and spillover effects. Here, we present analogous estimators for the three causal quantities defined above. Define

$$\widehat{Y}_j(z) = \frac{\sum_{i=1}^{n_j} Y_{ij}\mathbf{1}(Z_{ij}=z)}{\sum_{i=1}^{n_j}\mathbf{1}(Z_{ij}=z)} \quad \text{and}$$

$$\widehat{Y}(z,a) = \frac{\sum_{j=1}^{J}\widehat{Y}_j(z)\mathbf{1}(A_j=a)}{\sum_{j=1}^{J}\mathbf{1}(A_j=a)},$$

where $\widehat{Y}_j(z)$ is the average outcome under treatment condition z in cluster j, and $\widehat{Y}(z,a)$ is the average of $\widehat{Y}_j(z)$ in clusters with treatment assignment mechanism a. The following theorem gives the unbiased estimators of the ADE, MDE, and ASE.

Theorem 1 (Unbiased estimation). *Define* $\widehat{Y} = (\widehat{Y}(1,1),\widehat{Y}(0,1),\dots,\widehat{Y}(1,m),\widehat{Y}(0,m))$. *Under Assumptions 1, 2, and 3,* $\widehat{Y}$ *is unbiased for* $\overline{Y}$, *that is,* $\mathbb{E}(\widehat{Y}) = \overline{Y}$. *Therefore,* $\widehat{\mathrm{ADE}} = C_1\widehat{Y}$, $\widehat{\mathrm{MDE}} = C_2\widehat{Y}$, *and* $\widehat{\mathrm{ASE}} = C_3\widehat{Y}$ *are unbiased for* ADE, MDE, *and* ASE, *respectively, that is,* $\mathbb{E}(\widehat{\mathrm{ADE}}) = \mathrm{ADE}$, $\mathbb{E}(\widehat{\mathrm{MDE}}) = \mathrm{MDE}$, $\mathbb{E}(\widehat{\mathrm{ASE}}) = \mathrm{ASE}$.

We note that the theory of simple random sampling implies $\mathbb{E}\{\widehat{Y}_j(z) \mid A_j = a\} = \overline{Y}_j(z,a)$. Therefore, it is straightforward to show that $\mathbb{E}\{\widehat{Y}(z,a)\} = \overline{Y}(z,a)$ and hence $\mathbb{E}(\widehat{Y}) = \overline{Y}$.

4.2 | Variance

Hudgens and Halloran (2008) derived the variances of $\widehat{\mathrm{ADE}}(a)$ and $\widehat{\mathrm{ASE}}(z;a',a)$ under stratified interference (Assumption 3). However, this is not sufficient for obtaining the variance of our causal quantities, which require the covariance between the elements in $\widehat{Y}$. We first derive the covariance matrix of $\widehat{Y}$ and then use it to obtain the covariance matrix of ADE, MDE, and ASE.

The covariance matrix of $\widehat{Y}$ consists of the variance of $\widehat{Y}(z,a)$ and the covariance between $\widehat{Y}(z,a)$ and $\widehat{Y}(z',a')$. Define,

$$\sigma_j^2(z,z';a,a') = \frac{1}{n_j-1}\sum_{i=1}^{n_j}\{Y_{ij}(z,a)-\overline{Y}_j(z,a)\}\{Y_{ij}(z',a')-\overline{Y}_j(z',a')\},$$

$$\sigma_b^2(z,z';a,a') = \frac{1}{J-1}\sum_{j=1}^{J}\{\overline{Y}_j(z,a)-\overline{Y}(z,a)\}\{\overline{Y}_j(z',a')-\overline{Y}(z',a')\},$$

where $\sigma_j^2(z,z';a,a')$ is the within-cluster covariance between $Y_{ij}(z,a)$ and $Y_{ij}(z',a')$, and $\sigma_b^2(z,z';a,a')$ is their between-cluster covariance. When $a=a'$, $\sigma_j^2(z,z';a,a')$ reduces to $\sigma_j^2(z,z';a)$ and $\sigma_b^2(z,z';a,a')$ equals $\sigma_b^2(z,z';a)$. When $z=z'$, $\sigma_j^2(z,z';a,a')$ reduces to $\sigma_j^2(z;a,a')$ and $\sigma_b^2(z,z';a,a')$ equals $\sigma_b^2(z;a,a')$. Lastly, when $z=z'$ and $a=a'$, $\sigma_j^2(z,z';a,a')$ reduces to $\sigma_j^2(z,a)$ and $\sigma_b^2(z,z';a,a')$ equals $\sigma_b^2(z,a)$. We denote $S_b = (\sigma_b^2(z,z';a,a'))_{2m\times 2m}$ and $S_j = (\sigma_j^2(z,z';a,a'))_{2m\times 2m}$ as the between and within cluster covariance matrix of $(Y_{ij}(1,1),Y_{ij}(0,1),\dots,Y_{ij}(1,m),Y_{ij}(0,m))$.

Let $0_{m\times n}$ and $1_{m\times n}$ be the $m\times n$ matrices of zeros and ones, respectively, whereas I_m is the $m\times m$ identity matrix. Use $\otimes$ and $\circ$ to denote the Kronecker and Hadamard products of matrices, respectively. Denote

$$H = \mathrm{diag}(J/J_1,\dots,J/J_m)\otimes 1_{2\times 2} - 1_{2m\times 2m},$$

$$H_j = \mathrm{diag}(J/J_1,\dots,J/J_m)\otimes\left\{\mathrm{diag}(n_j/n_{j1},n_j/n_{j0}) - 1_{2\times 2}\right\}.$$

The following theorem gives the covariance matrix of $\widehat{Y}$.

Theorem 2 (Variance–covariance matrix). *Under Assumptions 1–3, we have*

$$\mathrm{cov}(\widehat{Y}) = J^{-1}(H\circ S_b) + J^{-2}\sum_{j=1}^{J} n_j^{-1}(H_j\circ S_j).$$

The multiplication facilitates the development of sample size formulas in Section 4.5. Theorem 2 implies that the covariance matrices of $\widehat{\mathrm{ADE}}$, $\widehat{\mathrm{MDE}}$, and $\widehat{\mathrm{ASE}}$ are $\mathrm{var}\{\widehat{\mathrm{ADE}}\} = C_1DC_1^{\top}/J$, $\mathrm{var}\{\widehat{\mathrm{MDE}}\} = C_2DC_2^{\top}/J$, $\mathrm{var}\{\widehat{\mathrm{ASE}}\} = C_3DC_3^{\top}/J$, where $D = J\mathrm{cov}(\widehat{Y})$.

Because we cannot observe $Y_{ij}(1,a)$ and $Y_{ij}(0,a)$ simultaneously, no unbiased estimator exists for $\sigma_j^2(1,0;a)$. This implies that no unbiased estimation of D is possible. Following the idea of Hudgens and Halloran (2008), we propose a conservative estimator. Define

$$\widehat{\sigma}_b^2(z,a) = \frac{1}{J_a-1}\sum_{j=1}^{J}\left\{\widehat{Y}_j(z)-\widehat{Y}(z,a)\right\}^2\mathbf{1}(A_j=a),$$

$$\widehat{\sigma}_b^2(1,0;a) = \frac{1}{J_a-1}\sum_{j=1}^{J}\left\{\widehat{Y}_j(1)-\widehat{Y}(1,a)\right\}\left\{\widehat{Y}_j(0)-\widehat{Y}(0,a)\right\}\mathbf{1}(A_j=a),$$

where $\widehat{\sigma}_b^2(z,a)$ represents the between-cluster sample variance of $Y_{ij}(z,a)$, and $\widehat{\sigma}_b^2(1,0;a)$ denotes the between-cluster sample covariance between $Y_{ij}(1,a)$ and $Y_{ij}(0,a)$. The following theorem provides a conservative variance estimator, which is exactly unbiased when the cluster-level average potential outcome, that is, $\overline{Y}_j(z,a)$, does not vary across clusters.

Theorem 3 (Conservative estimator of variance). *Let $\widehat{D}$ be a 2m by 2m block diagonal matrix with the ath matrix ($a = 1, \dots, m$) on the diagonal*

$$\widehat{D}_a = \frac{J}{J_a}\begin{pmatrix} \widehat{\sigma}^2_b(1,a) & \widehat{\sigma}^2_b(1,0;a) \\ \widehat{\sigma}^2_b(1,0;a) & \widehat{\sigma}^2_b(0,a)\end{pmatrix}.$$

Then, $\widehat{D}$ is a conservative estimator for D, that is, $\mathbb{E}\{\widehat{D}\} - D$ is a positive semi-definite matrix. It is an unbiased estimator for D when the cluster-level average potential outcomes, that is, $\overline{Y}_j(z,a)$, is constant across clusters.

The covariance matrix estimator $\widehat{D}$ estimates $\text{var}\{\widehat{Y}(z,a)\}$ and $\text{cov}\{\widehat{Y}(1,a),\widehat{Y}(0,a)\}$ by their corresponding between-cluster sample variance and covariance, $\widehat{\sigma}^2_b(z,a)$ and $\widehat{\sigma}^2_b(1,0;a)$, while replacing $\text{cov}\{\widehat{Y}(1,a),\widehat{Y}(0,a')\}$ with 0. Theorem 3 implies the following conservative variance estimators for ADE, MDE, and ASE, $\widehat{\text{var}}\{\widehat{\text{ADE}}\} = C_1\widehat{D}C_1^\top/J$, $\widehat{\text{var}}\{\widehat{\text{MDE}}\} = C_2\widehat{D}C_2^\top/J$, $\widehat{\text{var}}\{\widehat{\text{ASE}}\} = C_3\widehat{D}C_3^\top/J$. Similar to $\widehat{D}$, these estimators are unbiased if $\overline{Y}_j(z,a)$ are the same across clusters.

Note that alternative conservative variance estimators exist with different conditions for unbiasedness (Mukerjee et al., 2018). In particular, Hudgens and Halloran (2008) proposed the following conservative variance estimator for each ADE(a),

$$\frac{1}{J_a}\left(1-\frac{J_a}{J}\right)\{\widehat{\sigma}^2_b(1,a)+\widehat{\sigma}^2_b(0,a)-2\widehat{\sigma}^2_b(1,0;a)\}$$
$$+\frac{1}{JJ_a}\sum_{j=1}^{J}\left\{\frac{\widehat{\sigma}^2_j(1)}{n_{j1}}+\frac{\widehat{\sigma}^2_j(0)}{n_{j0}}\right\}\mathbf{1}(A_j=a),$$

where $\widehat{\sigma}^2_j(z) = 1/(n_{jz}-1)\cdot\sum_{i=1}^{n_j}\{Y_{ij}-\widehat{Y}_j(z)\}^2\mathbf{1}(Z_{ij}=z)$ represents the within-cluster sample variance of $Y_{ij}(z)$. They show that it is a conservative estimator of the variance of ADE(a), and is unbiased if the unit-level direct effects, $Y_{ij}(1,a)-Y_{ij}(0,a)$, do not vary within each cluster.

In practice, this variance estimator is generally smaller than the a-th diagonal element of $\widehat{\text{var}}\{\widehat{\text{ADE}}\}$. However, its conservativeness property holds only for the variance of each ADE(a). No similar estimator can be obtained for the covariance matrix of $\widehat{\text{ADE}}$. For example, replacing the diagonal elements of $\widehat{\text{var}}\{\widehat{\text{ADE}}\}$ with Hudgens and Halloran (2008)'s estimators does not yield a conservative estimator for $\text{var}\{\widehat{\text{ADE}}\}$. Therefore, we recommend using Hudgens and Halloran (2008)'s estimator when the variance of ADE(a) alone is of interest, whereas our proposed estimator should be used when the joint distribution of ADE is of interest.

4.3 | Asymptotic normality of the estimators

To conduct statistical inference and power analysis, we study the asymptotic properties of the estimators. We state the regularity conditions for finite-population asymptotics.

Condition 1. Denote $\overline{Y^4_j(z,a)} = n_j^{-1}\sum_{i=1}^{nj} Y^4_{ij}(z,a)$. As J goes to infinity, for $z = 0,1$ and $a = 1,\dots,m$,

(a) J_a/J has a limit in (0,1); $\epsilon \le n_{jz}/n_j \le 1-\epsilon$ for $j = 1,\dots,J$, and some $\epsilon \in (0,1/2)$;
(b) $\max_j |\overline{Y}_j(z,a)-\overline{Y}(z,a)|^2/J = o(1)$;
(c) $\overline{Y}$ has a finite limit; $S_b = O(1)$ and $J^{-1}\sum_{j=1}^{J} n_j^{-1}(H_j \circ S_j) = O(1)$;
(d) $J^{-2}\sum_{j=1}^{J}\overline{Y^4_j(z,a)} = o(1)$.

From Theorem 2, Conditions 1(a) and (b) imply that the covariance matrix of $\widehat{Y}$ is at the order of J^{-1}, which guarantees the consistency of $\widehat{Y}$ for estimating $\overline{Y}$. Conditions 1(c) and (d) hold as long as Y_i is bounded. Condition 1 requires only J to go to infinity and thus can incorporate both scenarios when the cluster size is fixed or goes to infinity.

Theorem 4 (Asymptotic normality). *Under Assumptions 1–3, and Condition 1, we have $\sqrt{J}(\widehat{Y}-\overline{Y}) \xrightarrow{d} N(0, D^*)$, where D^* is the limiting value of D.*

4.4 | Hypothesis testing

We consider testing the following three null hypotheses of no direct effect, no MDE, and no spillover effect, H_0^{de} : ADE = 0, H_0^{mde} : MDE = 0, H_0^{se} : ASE = 0. Because ADE, MDE, and ASE are linear transformations of $\overline{Y}$, we focus on a more general null hypothesis,

$$H_0 : C\overline{Y} = 0, \tag{3}$$

where C is a constant contrast matrix with full row rank. By setting C to C_1, C_2, and C_3, H_0 becomes H_0^{mde}, H_0^{se}, and H_0^{se}, respectively. We propose the following Wald-type test statistic,

$$T = J(C\widehat{Y})^\top(C\widehat{D}C^\top)^{-1}(C\widehat{Y}), \tag{4}$$

where the covariance matrix of $\widehat{Y}$ is replaced with its conservative estimator $\widehat{D}/J$. Unfortunately, T does not follow a χ^2 distribution asymptotically with the conservative covariance matrix estimator.

Theorem 5 (Asymptotic distribution of the test statistic). *Suppose that Assumptions 1–3, and Condition 1 hold, and the rank of C is k. Under the null hypothesis in Equation (3), the asymptotic distribution of the test statistic T defined in Equation (4) is stochastically dominated by the χ^2 distribution with k degrees of freedom, that is,* $\Pr(T \geq t) \leq \Pr\{X \geq t\}$ *for any constant t where $X \sim \chi^2(k)$.*

With a pre-specified significance level α, we can reject H_0 if $T > \chi^2_{1-\alpha}(k)$ where $\chi^2_{1-\alpha}(k)$ represents the $(1-\alpha)$ quantile of the χ^2 distribution with k degrees of freedom. Theorem 5 implies that this rejection rule controls the type I error asymptotically.

We can use the following three Wald-type test statistics for the direct, marginal direct, and spillover effects, respectively,

$$T_{\text{de}} = J(C_1\widehat{Y})^\top(C_1\widehat{D}C_1^\top)^{-1}(C_1\widehat{Y}), \tag{5}$$

$$T_{\text{mde}} = J(C_2\widehat{Y})^\top(C_2\widehat{D}C_2^\top)^{-1}(C_2\widehat{Y}), \tag{6}$$

$$T_{\text{se}} = J(C_3\widehat{Y})^\top(C_3\widehat{D}C_3^\top)^{-1}(C_3\widehat{Y}). \tag{7}$$

Theorem 5 implies that under the corresponding null hypothesis, the asymptotic distributions of T_{de}, T_{mde}, and T_{se} are stochastically dominated by a χ^2 distribution with the degrees of freedom equal to m, 1, and $2(m-1)$, respectively.

4.5 | Sample size formula

When planning a two-stage randomized experiment, we may wish to determine the sample size needed to detect a certain effect size with a given statistical power $(1-\beta)$ and a significance level (α). The sample size depends on the number of clusters and cluster sizes. In two-stage randomized experiments, however, the cluster sizes are often fixed. Therefore, we derive the required number of clusters of fixed sizes that ensures sufficient power to detect a deviation from the null hypothesis.

General formulation.
We begin by considering a general alternative hypothesis,

$$H_1 : C\overline{Y} = x, \tag{8}$$

where C is a $k \times 2m$ matrix of full row rank $(k \leq 2m)$ and x is a vector of constants. With the test statistic given in Equation (4), the required number of clusters J should satisfy

$$\text{pr}\{J(C\widehat{Y})^\top(C\widehat{D}C^\top)^{-1}(C\widehat{Y}) \geq \chi^2_{1-\alpha}(k) \mid C\overline{Y} = x\} \geq 1-\beta. \tag{9}$$

However, because $\widehat{D}$ is a conservative estimator for D, $J(C\widehat{Y})^\top(C\widehat{D}C^\top)^{-1}(C\widehat{Y})$ follows a generalized chi-square distribution instead of a standard chi-square distribution asymptotically, rendering it difficult to directly solve Equation (9) for J.

Fortunately, based on the properties of the generalized chi-square distribution, the following theorem gives a conservative sample size formula.

Theorem 6 (General sample size formula). *Consider a statistical hypothesis test with level α, where the null and alternative hypotheses are given in Equations (3) and (8), respectively. We reject the null hypothesis if $T > \chi^2_{1-\alpha}(k)$ where the test statistic T is defined in Equation (4) and k is the rank of C. Then, the number of clusters required for this hypothesis test to have the statistical power of $(1-\beta)$ is given by*

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(k), 1-\beta, k)}{x^\top\{C\mathbb{E}(\widehat{D})C^\top\}^{-1}x},$$

where $s^2(q, 1-\beta, k)$ represents the non-centrality parameter of the non-central χ^2 distribution with k degrees of freedom, whose β quantile is equal to q.

In practice, we must compute $s^2(\chi^2_{1-\alpha}(k), 1-\beta, k)$ numerically. Based on Theorem 6, we can obtain the sample size formula for the direct, marginal direct, and spillover effects by setting k to m, 1, and $2(m-1)$, respectively.

Simplification
The practical difficulty of the sample size formula in Theorem 6 is that it requires the specification of many parameters in $\mathbb{E}(\widehat{D})$ and the value of vector x in the alternative hypothesis. Thus, we consider the further simplification of the sample size formula to facilitate its application by reducing the number of parameters to be specified by researchers.

Assumption 4 (Simplification). We make the following simplifying assumptions:

(1) The within-cluster variances of $Y_{ij}(z,a)$ are the same across different clusters, different treatments, and different treatment assignment mechanisms: $\sigma_j^2(z,a) = \sigma_w^2$ for all z, a;

(2) The between-cluster variances of $Y_{ij}(z,a)$ are the same across different treatments and different treatment assignment mechanisms: $\sigma_b^2(z,a) = \sigma_b^2$ for all z and a;
(3) The within-cluster and between-cluster correlation coefficients between $Y_{ij}(1,a)$ and $Y_{ij}(0,a)$ are the same and non-negative: $\sigma_j^2(1,0;a) = \sigma_{j'}^2(1,0;a') \geq 0$ and $\sigma_b^2(1,0;a) = \sigma_b^2(1,0;a') \geq 0$ for all j, j', a, and a'.

Baird et al. (2018) also made simplifying assumptions to reduce the number of parameters. The authors, however, use the super population framework to derive the optimal design parameters rather than the sample size formulas as done in this paper.

Under these simplifying conditions, we can write $\sigma_j^2(1,0;a) = \rho\sigma_w^2$ and $\sigma_b^2(1,0;a) = \rho\sigma_b^2$, where $\rho \geq 0$ is the within-cluster and between-cluster correlation coefficient between $Y_{ij}(1,a)$ and $Y_{ij}(0,a)$. We can also rewrite σ_w^2 and σ_b^2 as $\sigma_w^2 = (1-r)\sigma^2$ and $\sigma_b^2 = r\sigma^2$, where $\sigma^2 = \sigma_w^2 + \sigma_b^2$ represents the total variance of $Y_{ij}(z,a)$ and $r = \sigma_b^2/(\sigma_w^2 + \sigma_b^2)$ is the intracluster correlation coefficient with respect to $Y_{ij}(z,a)$. Denote $D_0^* = \text{diag}(D_{01}^*, D_{02}^*, \ldots, D_{0m}^*)$ with

$$D_{0a}^* = \frac{1}{q_a}\begin{pmatrix} r + \frac{(1-p_a)(1-r)}{\bar{n}p_a} & \rho\left(r - \frac{1-r}{\bar{n}}\right) \\ \rho\left(r - \frac{1-r}{\bar{n}}\right) & r + \frac{p_a(1-r)}{\bar{n}(1-p_a)} \end{pmatrix}$$

for $a = 1, \ldots, m$, where p_a is the treated proportion under treatment assignment mechanism a and $\bar{n}$ is the harmonic mean of n_j defined as $\bar{n} = J/\sum_{j=1}^J \frac{1}{n_j}$. When $n_j = n$ for all j, $\bar{n} = n$. Thus, D_0^* is a $2m \times 2m$ block diagonal matrix with D_{0a}^* being the ath block for $a = 1, \ldots, m$.

We derive the sample size formula for the direct effect under Assumption 4. To reduce the number of parameters in the alternative hypothesis H_1 : ADE $= x$, we consider the alternative hypothesis about the direct effects across m treatment assignment mechanisms:

$$H_1^{\text{de}} : |\text{ADE}(a)| = \mu \quad \text{for all } a. \tag{10}$$

The following theorem gives the sample size formula for rejecting the null hypothesis H_0 : ADE $= 0$, with respect to the alternative hypothesis in Equation (10).

Theorem 7 (Simplified sample size formula for direct effects). *Consider a statistical hypothesis test with level α where the null hypothesis is H_0^{de} : ADE $= 0$ and the alternative hypothesis is given in Equation (10). We reject the null hypothesis if $T_{de} > \chi^2_{1-\alpha}(m)$, where the test statistic T_{de} is defined in Equation (5). Under Assumption 4, the number of clusters required for this test to have the statistical power of $1-\beta$ is given by*

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m)\cdot\sigma^2}{\mu^2}\cdot \frac{1}{\sum_{a=1}^m \left\{(1,-1)D_{0a}^*(1,-1)^\top\right\}^{-1}}. \tag{11}$$

Moreover, if $r \geq 1/(n+1)$, then the required number of clusters is given by

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(m), 1-\beta, m)\cdot\sigma^2}{\mu^2}\cdot \frac{1}{\sum_{a=1}^m \{(1,-1)D_{0a}(1,-1)^\top\}^{-1}}, \tag{12}$$

where $D_{0a} = q_a^{-1}\text{diag}(r + (np_a)^{-1}(1-p_a)(1-r), r + \{n(1-p_a)\}^{-1}p_a(1-r))$.

To apply Equation (11), one needs to specify (p_a, q_a) based on the study design and $(\rho, r, \sigma^2, \bar{n})$ based on prior information (e.g., pilot studies). Because the sample size formula depends on the cluster sizes only through their harmonic mean, the formula can be applied regardless of whether the cluster sizes are given as fixed or random. Since ρ is the correlation coefficient between potential outcomes under different treatment conditions, it is an unidentifiable parameter. Therefore, we provide a more conservative sample size formula in Equation (12) that does not involve ρ. The condition $r \geq 1/(n+1)$ is easily satisfied so long as the cluster size is moderate or large. Under this condition, if J satisfies Equation (12), then it also satisfies Equation (11).

Next, we derive the sample size formula for the MDE under Assumption 4. Because the MDE is a scalar, we continue to use the alternative hypothesis considered above, that is, H_1 : MDE $= \mu$. The following theorem gives the sample size formula.

Theorem 8 (Simplified sample size formula for marginal direct effect). *Consider a statistical hypothesis test with level α, where the null hypothesis is H_0^{mde} : MDE $= 0$ and the alternative hypothesis is H_1^{mde} : MDE $= \mu$. We reject the null hypothesis if $T_{mde} > \chi^2_{1-\alpha}(1)$, where T_{mde} is the test statistic defined in Equation (6). Under Assumption 4, the number of clusters required for the test to have the statistical power of $1-\beta$ is given by*

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(1), 1-\beta, 1)\cdot\sigma^2}{\mu^2}\cdot \sum_{a=1}^m q_a^2\{(1,-1)D_{0a}^*(1,-1)^\top\}. \tag{13}$$

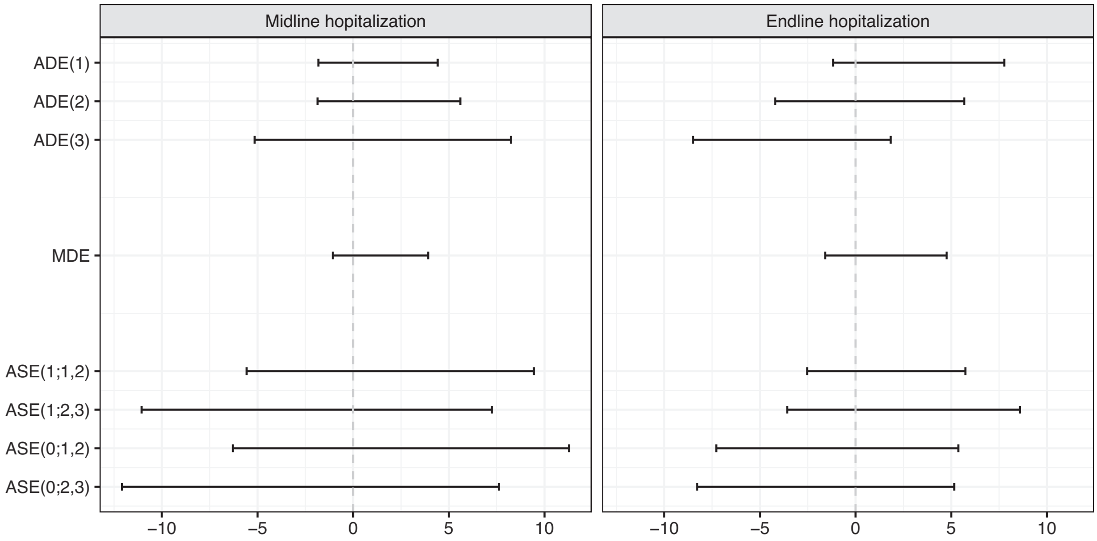

FIGURE 1 Estimated average direct, marginal direct, and spillover effects for the two outcomes of interest, midline hospitalization and endline hospitalization (percentage points). The top three lines are the average direct effects (ADE) under the three treatment assignment mechanisms; the middle line is the marginal direct effect (MDE); the bottom two lines are the average spillover effects (ASE) comparing the adjacent treatment assignment mechanisms under the treatment and control conditions. 95% confidence intervals as well as point estimates are shown.

Moreover, if $r \geq 1/(n+1)$, then the number of clusters required is given by

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(1), 1-\beta, 1)\cdot\sigma^2}{\mu^2}\cdot \sum_{a=1}^{m} q_a^2\{(1,-1)D_{0a}(1,-1)^\top\}. \tag{14}$$

Similar to Theorem 7, the application of Equation (13) requires the specification of both (p_a, q_a, n) and (ρ, r, σ^2), while the more conservative formula given in Equation (14) does not depend on ρ.

Finally, we derive the sample size formula for the spillover effect under Assumption 4. To reduce the number of parameters in the alternative hypothesis H_1 : ASE $= x$, we consider the following alternative hypothesis about the spillover effects across different treatment conditions and treatment assignment mechanisms,

$$H_1^{\mathrm{se}} : \max_{a\neq a'} |\mathrm{ASE}(z;a,a')| = \mu \quad \text{for all } z. \tag{15}$$

The next theorem gives the sample size formula.

Theorem 9 (Simplified sample size formula for spillover effects). *Consider a statistical hypothesis test with level α, where the null hypothesis is H_0^{se} : $\mathrm{ASE}(z;a,a') = 0$ for all z and $a \neq a'$ and the alternative hypothesis given in Equation (15). We reject the null hypothesis if $T_{se} > \chi^2_{1-\alpha}(2(m-1))$, where the test statistic T_{se} is defined in Equation (7). Under Assumption 4, the number of clusters required for the test to have the statistical power $1-\beta$ is given by*

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(2(m-1)), 1-\beta, 2(m-1))\cdot\sigma^2}{\mu^2\cdot\min_{s\in\mathcal{S}} s^\top\{C_3 D_0^* C_3^\top\}^{-1}s}, \tag{16}$$

where $\mathcal{S}$ is the set of $s = (\mathrm{ASE}(0;1,2), \mathrm{ASE}(0;2,3), \ldots, \mathrm{ASE}(0;m-1,m), \mathrm{ASE}(1;1,2), \mathrm{ASE}(1;2,3), \ldots, \mathrm{ASE}(1;m-1,m))$ satisfying $\max_{a\neq a'}|\mathrm{ASE}(z;a,a')| = 1$ for $z = 0, 1$.

In Appendix S4, we show how to numerically compute the denominator of Equation (16) using quadratic programming. Unlike Theorems 7 and 8, we cannot obtain a more conservative sample size formula by setting ρ to 0. Nonetheless, we use the following formula that does not involve ρ and evaluate its performance in our simulation study given in Appendix S5,

$$J \geq \frac{s^2(\chi^2_{1-\alpha}(2(m-1)), 1-\beta, 2(m-1))\cdot\sigma^2}{\mu^2\cdot\min_{s\in\mathcal{S}} s^\top\{C_3 D_0 C_3^\top\}^{-1}s}, \tag{17}$$

where $D_0 = \mathrm{diag}(D_{01}, D_{02}, \ldots, D_{0m})$.

5 | EMPIRICAL ANALYSIS

In this section, we analyze the data from the randomized experiment of the Indian National Health Insurance

TABLE 2 The required number of clusters for detecting the causal effects of certain sizes with the statistical power 0.8 at the significance level 0.05

	$\lvert\mathrm{ADE}(a)\rvert$ = 5%	MDE = 5%	$\max_{a\neq a'} \lvert\mathrm{ASE}(z; a, a')\rvert$ = 5%
Midline hospitalization	803	585	2230
Endline hospitalization	400	323	857

Abbreviations: ADE, average direct effects; MDE, marginal direct effect.

Program described in Section 2. We focus on two health outcomes: midline and endline hospitalizations. Figure 1 shows the estimated direct, marginal direct, and spillover effects for midline and endline hospitalizations with their 95% confidence intervals. For midline hospitalization (left panel), we find all of the estimated ADE to be positive under the three treatment assignment mechanisms but statistically insignificant. Little heterogeneity in the direct effects means that the estimated MDE is similar to the three ADE. The spillover effects of treatment mechanism 1 versus 2 are estimated to be positive, while the spillover effects of treatment mechanism 2 versus 3 are estimated to be negative. All of these spillover effects, however, are not statistically significant.

For endline hospitalization (right panel), the estimated MDE is similar to that for midline hospitalization. However, heterogeneity exists across different treatment assignment mechanisms. The estimated ADE is positive under treatment assignment mechanism 1 but negative under treatment assignment mechanism 3, with the difference between them being 6.6 percentage points (95% CI: $[-0.2, 13.5]$). This may suggest that enrolling in the RSBY leads to a reduction in hospitalization in the long run, but only when the treatment proportion is not large. The spillover effects are positive under the treatment condition and negative under the control condition, but they are not distinguishable from zero.

Next, we consider a hypothetical scenario, in which a researcher uses this experiment as a pilot study for planning a future experiment. The goal is to compute the sample size required for detecting certain effect sizes at statistical power 0.8 and significance level 0.05. For each outcome, we consider three null hypotheses: $|\mathrm{ADE}(a)| =$ 5 percentage points (pp.) for all a, $\mathrm{MDE} = 5$pp., and $\max_{a\neq a'} |\mathrm{ASE}(z; a, a')| = 5$pp. for all z. Note that the total variance is $\sigma^2 = 0.175$ for midline hospitalization and $\sigma^2 = 0.180$ for endline hospitalization; the intracluster correlation coefficient is $r = 0.42$ for midline hospitalization and $r = 0.11$ for endline hospitalization.

Table 2 presents the results. We find that a greater sample size is required for the midline hospitalization than for the endline hospitalization. The reason is that the intracluster correlation coefficient is much larger for the midline hospitalization. In addition, a much greater sample size is required for detecting the spillover effects than the direct effects. This is because only a small proportion of the entire sample (i.e., 15%) is allocated to treatment assignment mechanism 2. Leading to a larger required overall sample size for detecting the spillover effects.

6 | CONCLUDING REMARKS

In this paper, we introduced a general methodology for analyzing and planning two-stage randomized experiments. Future research should address several remaining methodological challenges. First, many experiments suffer from attrition, which leads to missing outcome data for some units. It is of interest to deal with such a complication in the presence of spillover effects. Second, it is often believed that spillover effects arise from interactions among a relatively small number of units. How to explore this causal heterogeneity is an important question to be addressed. Third, the standard two-stage randomized design can be extended to sequential experimentation, allowing researchers to examine how spillover effects evolve over time. Finally, it is of interest to develop an optimal policy that exploits spillover effects. The two-stage randomized design, or its extensions, may be able to shed light on the construction of such cost-effective policies.

ACKNOWLEDGMENT

This work was supported by Alfred P. Sloan Foundation (grant no. 2020-13946).

DATA AVAILABILITY STATEMENT

The data that support the findings in the paper are available in the Supporting Information. See also Jiang et al. (2022).

ORCID

Zhichao Jiang https://orcid.org/0000-0002-8571-0217
Kosuke Imai https://orcid.org/0000-0002-2748-1022
Anup Malani https://orcid.org/0000-0002-2594-5778

REFERENCES

Angelucci, M. & Di Maro, V. (2016) Programme evaluation and spillover effects. *Journal of Development Effectiveness*, 8(1), 22–43.

Aronow, P. & Samii, C. (2017) Estimating average causal effects under general interference. *Annals of Applied Statistics*, 11(4), 1912–1947.

Baird, S., Bohren, J.A., McIntosh, C. & Ozler, B. (2018) Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5), 844–860.

Balzer, L.B., Petersen, M.L., van der Laan, M.J. & Collaboration, S. (2016) Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Statistics in Medicine*, 35(21), 3717–3732.

Balzer, L.B., Petersen, M.L., van der Laan, M.J. & Consortium, S. (2015) Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. *Statistics in Medicine*, 34(6), 999–1011.

Bargagli-Stoffi, F.J., Tortú, C. & Forastiere, L. (2020) Heterogeneous treatment and spillover effects under clustered network interference. *arXiv preprint arXiv:2008.00707*.

Basse, G. & Feller, A. (2018) Analyzing multilevel experiments in the presence of peer effects. *Journal of the American Statistical Association*, 113(521), 41–55.

Benjamin-Chung, J., Arnold, B.F., Berger, D., Luby, S.P., Miguel, E., Colford Jr, J.M. & Hubbard, A.E. (2018) Spillover effects in epidemiology: parameters, study designs and methodological considerations. *International Journal of Epidemiology*, 47(1), 332–347.

Forastiere, L., Airoldi, E.M. & Mealli, F. (2021) Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534), 901–918.

Forastiere, L., Mealli, F. & VanderWeele, T.J. (2016) Identification and estimation of causal mechanisms in clustered encouragement designs: disentangling bed nets using Bayesian principal stratification. *Journal of the American Statistical Association*, 111(514), 510–525.

Huang, K., Jiang, Z. & Imai, K. (2022) RCT2: R package for designing and analyzing two-stage randomized experiments. Available at the Comprehensive R Archive Network. https://CRAN.R-project.org/package=RCT2. [Accessed 18 October 2022].

Hudgens, M.G. & Halloran, M.E. (2008) Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 832–842.

Imai, K., Jiang, Z. & Malai, A. (2021) Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, 116(534), 632–644.

Jiang, Z., Imai, K., Malani, A. (2022) Replication Data for: Statistical Inference and Power Analysis for Direct and Spillover Effects in Two-Stage Randomized Experiments. https://doi.org/10.7910/DVN/GASZXA, Harvard Dataverse, V1.

Liu, L. & Hudgens, M.G. (2014) Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505), 288–301.

Malani, A., Holtzman, P., Imai, K., Kinnan, C., Miller, M., Swaminathan, S., Voena, A., Woda, B. & Conti, G. (2021) Effect of health insurance in India: a randomized controlled trial. Technical Report Working Paper 29576, National Bureau of Economic Research.

Miles, C.H., Petersen, M. & van der Laan, M.J. (2019) Causal inference when counterfactuals depend on the proportion of all subjects exposed. *Biometrics*, 75(3), 768–777.

Mukerjee, R., Dasgupta, T. & Rubin, D.B. (2018) Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113(522), 868–881.

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

Rogers, T. & Feller, A. (2018) Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, 2, 335–342.

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

Rubin, D.B. (1990) Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science*, 5, 472–480.

Sävje, F., Aronow, P.M. & Hudgens, M.G. (2021) Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2), 673–701.

Sinclair, B., McConnell, M. & Green, D.P. (2012) Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4), 1055–1069.

Sobel, M.E. (2006) What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476), 1398–1407.

Tchetgen Tchetgen, E.J. & VanderWeele, T.J. (2012) On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1), 55–75.

VanderWeele, T.J., Hong, G., Jones, S.M. & Brown, J.L. (2013) Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention. *Journal of the American Statistical Association*, 108(502), 469–482.

Zhao, A. & Ding, P. (2022) Reconciling design-based and model-based causal inferences for split-plot experiments. *The Annals of Statistics*, 50(2), 1170–1192.

SUPPORTING INFORMATION

Web Appendices, and data and codes for reproducing the tables and figures referenced in Sections 2 and 5 are available with this paper at the Biometrics website on Wiley Online Library.

Data S1

How to cite this article: Jiang, Z., Imai, K. & Malani, A. (2023) Statistical inference and power analysis for direct and spillover effects in two-stage randomized experiments. *Biometrics*, 79, 2370–2381. https://doi.org/10.1111/biom.13782