Using Structural-Nested Models To Estimate The Effect Of Cluster-Level Adherence On Individual-Level Outcomes With A Three-Armed Cluster-Randomized Trial

By: Babette A. Brumback, Zhulin He, Mansi Prasad, Matthew C. Freeman, and **Richard Rheingans**

Abstract

Much attention has been paid to estimating the causal effect of adherence to a randomized protocol using instrumental variables to adjust for unmeasured confounding. Researchers tend to use the instrumental variable within one of the three main frameworks: regression with an endogenous variable, principal stratification, or structural-nested modeling. We found in our literature review that even in simple settings, causal interpretations of analyses with endogenous regressors can be ambiguous or rely on a strong assumption that can be diffi- cult to interpret. Principal stratification and structural-nested modeling are alternative frameworks that render unambiguous causal interpretations based on assumptions that are, arguably, easier to interpret. Our interest stems from a wish to estimate the effect of cluster-level adherence on individual-level binary outcomes with a three-armed cluster-randomized trial and polytomous adherence. Principal stratification approaches to this problem are quite challenging because of the sheer number of principal strata involved. Therefore, we developed a structural-nested modeling approach and, in the process, extended the methodology to accommodate cluster- randomized trials with unequal probability of selecting individuals. Furthermore, we developed a method to implement the approach with relatively simple programming. The approach works quite well, but when the structural-nested model does not fit the data, there is no solution to the estimating equation. We investigate the performance of the approach using simulated data, and we also use the approach to estimate the effect on pupil absence of school-level adherence to a randomized water, sanitation, and hygiene intervention in western Kenya.

Brumback, B. A., He, Z., Prasad, M., Freeman, M. C. and **Rheingans, R.** (2014), Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial. Statist. Med., 33: 1490–1502. doi: 10.1002/sim.6049

Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial

Babette A. Brumback,[a*†] Zhulin He,[b] Mansi Prasad,[a] Matthew C. Freeman[c] and Richard Rheingans[d]

Much attention has been paid to estimating the causal effect of adherence to a randomized protocol using instrumental variables to adjust for unmeasured confounding. Researchers tend to use the instrumental variable within one of the three main frameworks: regression with an endogenous variable, principal stratification, or structural-nested modeling. We found in our literature review that even in simple settings, causal interpretations of analyses with endogenous regressors can be ambiguous or rely on a strong assumption that can be difficult to interpret. Principal stratification and structural-nested modeling are alternative frameworks that render unambiguous causal interpretations based on assumptions that are, arguably, easier to interpret. Our interest stems from a wish to estimate the effect of cluster-level adherence on individual-level binary outcomes with a three-armed cluster-randomized trial and polytomous adherence. Principal stratification approaches to this problem are quite challenging because of the sheer number of principal strata involved. Therefore, we developed a structural-nested modeling approach and, in the process, extended the methodology to accommodate cluster-randomized trials with unequal probability of selecting individuals. Furthermore, we developed a method to implement the approach with relatively simple programming. The approach works quite well, but when the structural-nested model does not fit the data, there is no solution to the estimating equation. We investigate the performance of the approach using simulated data, and we also use the approach to estimate the effect on pupil absence of school-level adherence to a randomized water, sanitation, and hygiene intervention in western Kenya.

Keywords: unmeasured confounding; instrumental variables; structural-nested models; complex survey data; adherence; three-armed randomized clinical trial

1. Introduction

In our collaboration to assess the impact of a school-based water, sanitation, and hygiene (WASH) intervention on pupil absence in Nyanza Province, Kenya [1], one of the goals was to estimate the effect of the received components of the intervention, as distinct from the effect of the randomly assigned components. The nature of the study design and the primary research questions spurred our interest in statistical methods to estimate the effect of cluster-level adherence on individual-level binary outcomes with a three-armed cluster-randomized trial and polytomous adherence [2]. Much attention has been paid to estimating the causal effect of adherence to a randomized protocol using the randomization

[a]*Department of Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.*
[b]*National Institutes of Statistical Sciences, 19 T.W. Alexander Drive, PO Box 14006, Research Triangle Park, NC 27709, U.S.A.*
[c]*Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.*
[d]*Department of Environmental and Global Health, University of Florida, Gainesville, FL 32611, U.S.A.*
**Correspondence to: Babette A. Brumback, Department of Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.*
[†]*E-mail: brumback@ufl.edu*

assignment as an instrumental variable to adjust for unmeasured confounding. Though not without variation, researchers tend to use instrumental variables within one of three main frameworks: regression with an endogenous variable [3–9], principal stratification [10–21], or structural-nested modeling [22–30].

Although much attention has been paid to instrumental variables estimation with two-armed trials, we found very few examples of analyses with three-armed trials. We found two examples in which the investigators made use of the principal stratification framework [19, 20], but our application has an even more complex structure that lends itself better to a structural-nested modeling approach. The principal stratification approach encounters the difficulty that as the numbers of randomized treatments and adherence categories increase, even more so do the number of principal strata, leading either to nonidentifiability of parameters or to the need for a complex network of modeling assumptions. For comparison, we also apply an approach based on regression with an endogenous variable. Structural-nested models (SNMs) were introduced by Robins [22, 23] and further developed for binary and more general outcomes by Vansteelandt and Goetghebeur [25]. Korhonen, Laird, and Palmgren [26] developed and applied SNMs for time-to-event outcomes. Hernan and Robins [24] provided an accessible introduction, and the models have proven to be useful for adjusting estimated causal effects of adherence for unmeasured confounding [25–29]. SNMs provide estimates of the effect of observed adherence versus a reference level of adherence, conditional on the observed adherence level (sometimes referred to in simple settings as the effect of treatment on the treated). Vansteelandt *et al.* [30] recently offered a review of the use of SNMs with binary outcomes and pointed out that, in some instances, the estimating equation has no solution.

An additional complication of our application is that the study design utilized unequal probabilities of selecting individuals. We did not find any examples in the literature of applying SNMs with cluster-randomized trials or with sampling weights. One of the additional problems we encounter in this setting is the need to adjust for measured individual-level confounders of the effect of randomization on the outcome. We borrowed an idea from Cain *et al.* [31] in a simpler setting: they used weighting to adjust for the individual-level confounders, followed by weighted estimation with the instrumental variable. We thus treat this problem by weighting the sampled data with a product of two component weights: the first weight adjusts the sample so that individuals have equal probability of selection, and the second weight further adjusts it by removing the association between individual level confounders and randomization. Then we apply Newton's method for estimating the parameters of a weighted generalized structural-nested mean model, using an easily programmed algorithm. The sampling distribution can be approximated using survey standard errors via either the bootstrap or jackknife for complex survey data [32–35] or a sandwich estimator for complex survey data [32-33,36]. On the basis of our literature search, we believe that our methodology for a cluster-randomized trial with a complex sampling design, our simple method of computation, and our use of the jackknife are new developments in the methodology and application of SNMs. Furthermore, we apply and compare three different structural-nested modeling approaches with estimating causal relative risks with the school-based WASH data—the first based on a linear SNM, the second on a logistic SNM, and the third based on a loglinear SNM. We are unaware of previous attempts to compare the three approaches in terms of a common estimand.

The paper is organized as follows. In the next section, we introduce the school-based WASH intervention study. In Section 3, we explain the endogenous regressor framework. In Section 4, we present our method of estimation with weighted generalized structural-nested mean models, including our simple method of computation. In Section 5, we conduct a simulation study of the method to show that it is generally quite robust. In Section 6, we apply the method to the school-based WASH trial, and in Section 7, we conclude with a discussion.

2. The school-based water, sanitation, and hygiene intervention study

The school-based WASH intervention randomized public primary schools nested in three geographical strata to one of the three study arms: water treatment and hygiene (hand-washing) promotion (WH), additional sanitation improvement that included latrine construction (WH C S), or control. We assessed pupil absence at follow-up on a subset of pupils within each school. Because pupils were selected into the study with unequal probabilities, sampling weights needed to be incorporated into the analysis. Results of the intent-to-treat analysis presented in Freeman *et al.* [1] suggested that the school-based WASH components can improve school attendance, particularly for girls; therefore, in the present paper, we focus only on girls. Like Freeman *et al.* [1], we will also restrict our attention to two geographical strata (Rachuonyo and Suba). The third (Nyando/Kisumu) stratum experienced unusually low absence

at follow-up in all three arms, possibly because of political reasons, and thus because of effect-measure modification that needs to be analyzed separately.

As with many such studies, adherence at schools to the randomly assigned intervention components was far from perfect. The program did not achieve adherence for one or more of the three supplied components (W, H, or S) in many intervention schools, and fortunately, some control schools provided one or more of those components independent of program activity. For the purpose of analysis, we dichotomized the measure of adherence for each of the three components as adequate or not; therefore, overall adherence has eight levels. Ideally, we would be able to estimate the causal effects of seven of these levels versus the reference category of no W, no H, and no S. If there were no confounding, simple comparisons would be sufficient; if all confounders were measured, we could adjust the comparisons using logistic regression for complex survey data. However, one of the advantages of a randomized study is that randomized assignment can be used as an instrumental variable [3], provided certain assumptions hold, and thereby, we can also adjust for unmeasured confounders. However, as the school-based WASH trial only had three randomization arms, we are restricted by most instrumental variables methods either to making just two comparisons versus the reference category or to reducing the dimension of estimated parameters down to three in some other way, for example, by assuming that the comparisons are linearly related on the logit scale.

The next two sections present the statistical methods we used to analyze the school-based WASH intervention study. First, we considered the endogenous regression framework, and then we turned to SNMs.

3. Estimation within the endogenous regressor framework

In this section, we present how to use the endogenous regressor framework to adjust for unmeasured confounding of the effect of adherence in the context of a cluster-randomized trial. We let Z_i be a multinomial random variable denoting the randomized treatment arm, which will serve as an instrumental variable [3] in our analysis. For ease of exposition, in this section, we assume that we have obtained a simple random sample of individuals from a population in which students are randomly assigned to clusters, and then we randomized clusters to the intervention groups. Let Y_{ij} be the outcome for individual j in cluster i, and let A_i denote the adherence of cluster i, with reference level $A_i = 0$.

Using the endogenous regressor framework, one posits a regression representing the effect of A_i on Y_{ij}, such as

$$Y_{ij} = \gamma + A_{vi}\beta + E_{ij} \tag{1}$$

where $E(E_{ij}) = 0$ and A_{vi} is a vector function of A_i (perhaps denoting dummy variables, e.g. when A_i is multinomial) that equals zero when $A_i = 0$. The variable A_{vi} is then specified as an endogenous regressor because it is correlated with E_{ij}, because of unmeasured confounding. Finally, one assumes that the instrumental variable Z_i is independent of E_{ij}, because of randomization. Let Z_{vi} be a vector function of Z_i, which includes the intercept. The preceding assumptions imply that the estimating equation

$$\sum_i \sum_j Z_{vi}^T \left(Y_{ij} - A_{vi}\beta - \gamma \right) = 0 \tag{2}$$

is unbiased (i.e., the left hand side has mean zero) and, therefore, that when it can be solved uniquely for β and γ, the effect of A_i on Y_{ij} can be estimated consistently.

To interpret β causally and precisely, a potential outcome framework is helpful. We assume that the potential outcomes $Y_{ij}(a; z)$ to randomization with $Z_i = z$ and subsequent adherence $A_i = a$ are well defined for each participant and do not depend on z, so that $Y_{ij}(a; z) = Y_{ij}(a)$. We further assume that they satisfy the consistency assumption, $Y_{ij}(a) = Y_{ij}(A_i) = Y_{ij}$, when A_i is observed to equal a. Probably the most natural interpretation of β stems from an implicit formulation of the marginal structural model (MSM) [37]

$$E(Y_{ij}(a)) = \gamma + a_v\beta \tag{3}$$

where a_v relates to a just as A_{vi} relates to A_i. When Y_{ij} is binary, $a_v\beta = E\left(Y_{ij}(a) - Y_{ij}(0)\right)$ represents the causal effect of adherence at level a relative to the reference level in terms of a risk difference.

We next seek to determine conditions under which estimating equation (2) is an unbiased estimating equation for β. Let $E_{ij}(a) = Y_{ij}(a) - a_v\beta - \gamma$ and observe that $E_{ij} = E_{ij}(A_i)$. Because we assume

that the potential outcomes exist at baseline, our study design implies that $E_{ij}.a/ \perp Z_i$ for all a (where $\perp$ denotes independence). However, because it depends on A_i, E_{ij} is not generally independent of Z_i, unless $E_{ij}.a/$ is constant in a (which implies a constant effect of adherence across individuals, i.e., no effect modifiers). Thus, if $E_{ij}.a/$ is not constant in a, then a major assumption required by the endogenous regressor framework is violated, and in turn, equation (2) may be biased. With binary outcomes, requiring $E_{ij}.a/$ to be constant in a equivalently constrains the causal risk differences to equal $-1, 0$, or 1. Even with continuous outcomes, the assumption is implausible. Therefore, interpretation of ψ in terms of model (3) may be problematic.

However, if we weaken the requirement that $E_{ij} \perp Z_i$, and assume only that $E(Z_{vt}^T E_{ij}) = 0$, equation (2) is trivially unbiased. But this latter assumption is not implied by randomization and the MSM at (3); it needs further justification. Indeed, not even $E.E_{ij}/ = 0$ is implied by randomization and the MSM. In Appendix 7, we present a simple example for which the assumptions of randomization and the MSM are satisfied, but for which $E.E_{ij}/ \neq 0$ and $E.Z_i E_{ij}/ \neq 0$. By attempting to construct examples, one comes to appreciate that the assumption $E(Z_{vt}^T E_{ij}) = 0$ is difficult to comprehend.

One can, however, find alternative causal interpretations of the probability limit of the solution to the instrumental variables (IV) estimating equation (2) on the basis of other assumptions. For a relatively simple adherence structure, principal stratification leads to one type of interpretation. SNMs, which we develop in Section 4, are readily applied to more complex adherence structures and lead to another type of interpretation.

Problems with the endogenous regressor framework have surfaced repeatedly in the literature. Angrist, Imbens, and Rubin [11] noted ambiguity of the interpretation of E_{ij} in the endogenous regressor framework, but they did not formalize the consequences. Bang and Davis [38] observed bias in the IV estimating equation for estimating the ψ of (3), but they did not attempt to understand the cause of that bias. It is not difficult to find examples in the literature where researchers working within the endogenous regressor framework either are vague about the interpretation of the instrumental variables analysis or do not justify the assumption that $E(Z_{vt}^T E_{ij}) = 0$ using potential outcomes, or both [3–8]. We also found examples where researchers are using potential outcomes to interpret their estimates but nevertheless effectively assume that $E_{ij}.a/$ or a related latent variable is constant in a [9, 39, 40].

4. Estimation with weighted generalized structural-nested mean models

An alternative framework for using an instrumental variable to adjust for unmeasured confounders is based on an SNM. The structural-nested mean model incorporates the same potential outcomes introduced in the previous section. In this section, we also address the additional complications introduced by the complex sampling design of the cluster-randomized trial. Because of the randomization of clusters rather than individuals, Z_i is not necessarily independent of individual-level covariates. Suppose we could have randomized all clusters in the population and observed both cluster-level adherence and individual-level outcomes, so that Z_i, A_i, a set of measured individual-level covariates X_{ij}, and the potential outcomes $Y_{ij}.a/$ for all a are defined for all individuals in the population. Besides the assumptions that the potential outcomes are well-defined at baseline and that $Y_{ij}.A/ = Y_{ij}$, our methodology requires two additional assumptions. The first is as follows.

Assumption 1

Conditional on X_{ij}, the population distribution of $Y_{ij}.0/$ does not depend on Z_i; that is, $P^p(Y_{ij}.0/ | Z_i; X_{ij}) = P^p(Y_{ij}.0/ | X_{ij})$, where $P^p.V/$ is the probability that V equals its observed value on the basis of the distribution of the population data.

Let $W_{ij1} = P^p.Z_i/ / P^p(Z_i | X_{ij})$. Define $P^{W_1}.Y_{ij}.0/; Z_i; X_{ij}/ = P^p(Y_{ij}.0/; Z_i; X_{ij}) W_{ij1}$; by Assumption 1, $P^{W_1}.Y_{ij}.0/; Z_i; X_{ij}/ = P^p.Y_{ij}.0/; X_{ij}/P^p.Z_i/$. This weighted distribution reflects the distribution of the population data we would have observed if we could have randomized schools so that the distribution of X_{ij} was the same at each level of Z_i (e.g., by paired matching or frequency matching [41] of schools); note that for this distribution, $Y_{ij}.0/ \perp Z_i$. Thus, Assumption 1 implies that $E^{W_1}(Y_{ij}.0/ | Z_i) = E^{W_1}(Y_{ij}.0/) = E^p(Y_{ij}.0/)$, where $E^{W_1}.V | C/$ is the conditional expectation of V given C with respect to the weighted distribution $P^{W_1}.V | C/$ and $E^p.V/$ is the expectation of V with respect to the population distribution $P^p.V/$.

Next, let $P^{W_1}(Y_{ij}.0/; Y_{ij}.1/; Y_{ij}.2/; Z_i; A_i; X_{ij}) = P^p(Y_{ij}.0/; Y_{ij}.1/; Y_{ij}.2/; Z_i; A_i; X_{ij}) W_{ij1}$. Our second assumption is as follows.

Assumption 2

$h\,E^{W_1}\,Y_{ij}.a/\mathrm{j}A_i \mathrm{D}\, a; Z_i \quad \mathrm{D}\; h\,E^{W_1}\,Y_{ij}.0/\mathrm{j}A_i \mathrm{D}\, a; Z_i \quad \mathrm{C}\, a_v\,\phi$, where a_v was defined in the previous section and $h.\cdot/$ is a canonical link corresponding to a generalized linear model [42], such as $h.p/ \mathrm{D} \log .p=.1 - p//$, $h.p/ \mathrm{D} \log.p/$, or $h.p/ \mathrm{D} p$.

Assumption 2 states that the distribution of potential outcomes in the population satisfies a weighted generalized structural-nested mean model. It also implies that an unweighted structural-nested mean model of the same form holds for the counterfactual population data obtained by randomizing clusters so that the distribution of X_{ij} is the same at each level of Z_i. When $h.p/ \mathrm{D} p$, $a_v\phi$ represents a risk difference, for $h.p/ \mathrm{D} \log.p/$, a log relative risk, and for $h.p/ \mathrm{D} \log.p=.1\text{-}p//$, a log odds ratio. Our focus, however, is on estimating the relative risks $RR.a/ \mathrm{D}\; E^{W_1}\,Y_{ij}.a/\mathrm{j}A_i \mathrm{D}\, a \;=E^{W_1}\,Y_{ij}.0/\mathrm{j}A_i \mathrm{D}\, a$ for all a, which represent the effects of cluster-level adherence levels unconditionally on Z_i, had we randomized the clusters so that the distribution of X_{ij} was the same at each level of Z_i.

Let $W_{ij\,2}$ be the inverse probability that individual j from cluster i was selected into the actual study. $W_{ij\,2}$ may not be constant across individuals because, first, the chance that a cluster is selected into the study may vary, and, second, the chance that an individual within a cluster was selected for observation may also vary. Let $W_{ij} \mathrm{D} W_{ij1}W_{ij\,2}$. Let $\mu.A_i\,; Z_i \mathrm{I}\, ry/$ be a parametric model for $E^{W_1}.Y_{ij}\mathrm{j}A_i; Z_i/$ with parameter ry. When A_i is multinomial, one could use the saturated model $\mu.A_i\,; Z_i \mathrm{I}\, ry/ \mathrm{D} g.A_{vi}ry_1 \mathrm{C} Z_{vi}ry_2 \mathrm{C} A_i * Z_i ry_3/$, where A_{vi} and Z_{vi} were defined in the preceding section and $A_i * Z_i$ represents a multidimensional interaction that saturates the model. In our application, we will use this saturated model. Let D_i^T be a function of A_i and Z_i; we will let $D_i - .A_{vi}\,; Z_{vi}; A_i * Z_i/^T$. Under Assumptions 1 and 2, and assuming $\mu.A_i\,; Z_i \mathrm{I}\, ry/$ is correctly specified, we can consistently estimate $.\phi; ry/$ by solving the estimating equations

$$\begin{aligned} \dagger_i \dagger_j W_{ij} Z_{vi}^T \Big[h^{-1}\{h.\mu.A_i; Z_i \mathrm{I}\, ry// - A_{vi}\phi\} - {}_{\epsilon} \Big] &\mathrm{D}\, 0 \\ \dagger_i \dagger_j W_{ij} D_i^T \; Y_{ij} - \mu\,.A_i; Z_i \mathrm{I}\, ry/ \;&\mathrm{D}\, 0 \end{aligned} \tag{4}$$

for $.\phi; ry; {}_{\epsilon}/$. The first estimating equation is unbiased conditional on Z_i, because

$$E^{W_1}_{A_i \mathrm{j} Z_i}\Big[h^{-1}\Big\{ h\,E^{W_1}\,Y_{ij}\mathrm{j}A_i; Z_i \quad - A_{vi}\phi \Big\} \Big] \mathrm{D}\, E^{W_1}\,Y_{ij}.0/\mathrm{j}Z_i \;\mathrm{D}\, E^{W_1}\,Y_{ij}.0/ \; - {}_{\epsilon}$$

where the first equality follows from Assumption 2, and the second equality from Assumption 1 (and note that we substituted $E^{W_1}\,Y_{ij}\mathrm{j}A_i\,; Z_i$ in place of $\mu.A_i\,; Z_i\,; \mathrm{I}\, ry/$). The second equation is unbiased conditional on A_i and Z_i provided $\mu.A_i\,; Z_i \mathrm{I}\, ry/$ is correctly specified; if one uses a saturated model, this is automatic. In both cases, observe that if we had data from the entire population available, the equations would be unbiased with W_{ij1} in place of W_{ij}. However, because of the complex sampling design, we need to use W_{ij}. With the resulting estimates ϕ and $\eta 0$, one can use the model of Assumption 2 and $\mu\,.A_i\,; Z_i \mathrm{I}\, \eta 0/$ to solve for $E^{W_1}\,Y_{ij}.0/\mathrm{j}A_i \mathrm{D}\, a; Z_i$, and consequently for $E^{W_1}\,Y_{ij}.0/\mathrm{j}A_i \mathrm{D}\, a$ (by averaging with respect to the weighted distribution of Z_i given A_i) and $RR.a/$.

If we use a generalized linear model for $\mu\,.A_i; Z_i; \mathrm{I}\, ry/$ with a canonical link function $g^{-1}.\cdot/$, the second equation can be solved using weighted generalized linear model software (e.g. PROC GLM in SAS). If we furthermore let $h.\cdot/ \mathrm{D} g^{-1}.\cdot/$ and substitute $\eta 0$ for ry into the first equation, it reduces to

$$\dagger_i \dagger_j W_{ij} Z_{vi}^T \mathrm{f} g\,.D_i\eta 0 - A_{vi}\phi/ - {}_{\epsilon}\mathrm{g} \mathrm{D}\, 0 \tag{5}$$

which can be solved iteratively using Newton's method by linearizing $g\,.D_i\eta 0 - A_{vi}\phi/$ about a current estimate of ϕ and then solving a weighted version of equation (2) using weighted instrumental variables software (e.g. PROC SYSLIN in SAS). For example, with $g.x/ - \exp.x/=.1 \mathrm{C} \exp.x//$, and ϕ^t the current estimate of ϕ, we linearly approximate $g\,.D_i\eta 0 - A_{vi}\phi/$ as $Y_i^* - A_{vi}^*\phi$, where $A_{vi}^* - A_{vi} g\,.D_i\eta 0 - A_{vi}\phi/\,\{1 - g\,.D_i\eta 0 - A_{vi}\phi/\}$ and $Y_i^* - g\,.D_i ry_i - A_{vi}\phi/ \mathrm{C}\, A_{vi}^*\phi$. Then we solve the equation

$$\dagger_i \dagger_j W_{ij} Z_{vi}^T \; \{Y_i^* - A_{vi}^*\phi - {}_{\epsilon}\} \;\mathrm{D}\, 0 \tag{6}$$

to find the next estimate $\phi^{t\mathrm{C}1}$ of ϕ. Equation (6) can be solved using weighted instrumental variables software, with Y_i^* as the outcome, A_{vi}^* as the endogenous regressor, Z_{vi} as the instrument, and W_{ij} as the weights.

If instead of specifying $h.p/$ as $\log .p=.1 - p//$, we let $h.p/ \mathrm{D} p$ in Assumption 2 (and we let $g.x/ \mathrm{D} x$), the second equation can be solved using weighted linear regression, and the first equation becomes

$$\dagger_i \dagger_j W_{ij} Z_{vi}^T . D_{ij} 0 - A_{vi} \hat{} - \acute{} / \mathrm{D} 0 \quad (7)$$

which can be solved without iteration using weighted instrumental variables software, with $D_{ij} 0$ as the outcome, A_{vi} as the endogenous regressor, Z_{vi} as the instrument, and W_{ij} as the weights. We note that the resulting estimators of $\hat{}$ and $\acute{}$ are identical to the solutions to the weighted version of equation (2), that is, to

$$\dagger_i \dagger_j W_{ij} Z_{vi}^T \ Y_{ij} - A_{vi} \hat{} - \acute{} \ \mathrm{D} 0$$

However, the interpretation of $\hat{}$ and $\acute{}$ and the underlying assumptions required for that interpretation differ markedly from the weighted structural-nested modeling framework to the weighted endogenous regression framework. In the former framework, we require Assumptions 1 and 2, and $a_v \hat{} \mathrm{D} E^{W_1} Y_{ij} .a/ - Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} a$ is a risk difference conditional on $A_i \mathrm{D} a$. In the latter framework, $a_v \hat{} \mathrm{D} E^{W_1} \ Y_{ij} .a/ - Y_{ij} .0/$ is an unconditional risk difference, and we require the Assumption 1 as well as the opaque assumption that $E^{W_1} \ Z_{vi} E_{ij} \ \mathrm{D} 0$. Although much less opaque, Assumption 2 within the structural-nested modeling framework may not be plausible. It requires us to believe that $E^{W_1} Y_{ij} .a/ - Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} a; Z_i$ does not depend on Z_i. Even if we did not have all of the complications of a cluster-randomized trial with a complex sampling design, but instead were working with a simpler design .so that W_1 was constant/, any baseline covariate V_{ij} may have a different distribution conditional on $A_i \mathrm{D} a$ and $Z_i \mathrm{D} \acute{}_1$ than conditional on $A_i \mathrm{D} a$ and $Z_i \mathrm{D} \acute{}_2$. Consider V_{ij} to be gender, for example, in the context of the school-based WASH study. There may be proportionally more girls in the schools with adherence at level (1) in the control group than there are in the schools with adherence at level (1) in the WH group. In this case, Assumption 2 would require us to believe that even though gender is imbalanced across those two groups, the effect of adherence at level (1) versus level (0) within those two groups would be the same. If gender was not an effect-modifier, this would be plausible. But because the intent-to-treat analysis in Freeman *et al.* [1] suggested that gender is an effect-modifier, in Section 6, we analyze the school-based WASH data for girls only (observing that nothing in our methods development precluded us from working entirely within a subpopulation defined by baseline covariates.) However, we must hope that there is not another imbalance of baseline covariates that renders Assumption 2 implausible. We note that in much less complex applications, for example, those in which there is randomization to a placebo or active treatment, such that adherence is binary, and no one in the placebo group has access to the active treatment, Assumption 2 is tautologous.

We have observed in practice that if Assumption 2 does not fit the data, estimation can be problematic. One reason Assumption 2 may be incorrect is an improper choice of $h.\cdot/$. When $h.p/ \mathrm{D} p$ and Y_i is binary, we have observed that the estimated $E \ Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} a$ may fall outside the range of $Œ0; 1]$, sometimes entailing estimated risk differences outside the possible range of $Œ-1; 1]$. The example in Appendix 7 illustrates this problem. For the observed data in that example (with any choice of n, e.g., $n \mathrm{D} 500$), the estimated $E \ Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1 \mathrm{D} -0:25$ when we assume $h.p/ \mathrm{D} p$. The estimated $E.Y_{ij} \mathrm{j} A_i \mathrm{D} 1/ \mathrm{D} 0:25$, and so the estimated risk difference is 0.5, but it is meaningless because of the negativity of $E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1/$. For the counterfactual data in that example, we verified that Assumption 2 is incorrect for all three choices of $h.\cdot/$, by calculating (i) $E.Y_{ij} - Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \mathrm{D} -0:56$, whereas $E.Y_{ij} - Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1/ \mathrm{D} -0:35$; (ii) $\log \ E.Y_{ij} \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \ - \log \ E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \ \mathrm{D} -1:50$, whereas $\log \ E \ Y_{ij} \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1 \ - \log \mathrm{f} E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1/ \mathrm{g} \mathrm{D} -0:74$; and (iii) $\mathrm{logitf} E.Y_{ij} \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \mathrm{g} - \mathrm{logitf} E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \mathrm{g} \mathrm{D} -2:60$, whereas $\mathrm{logit} \ E.Y_{ij} \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1/ \ - \mathrm{logit} \ E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1/ \ \mathrm{D} -1:48$.

An interesting next question is whether there is any choice of $h.\cdot/$ that renders Assumption 2 correct for the counterfactual plus observed data in Appendix 7. To answer this question, we calculated $h \ E.Y_{ij} \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \ \mathrm{D} 0:16$, $h \ E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 0/ \ \mathrm{D} 0:72$, $h \ E.Y_{ij} \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1/ \ \mathrm{D} 0:3226$, and $h \ E.Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} 1; Z_i \mathrm{D} 1/ \ \mathrm{D} 0:6774$. Choosing $h.\cdot/$ to make Assumption 2 correct would require $h.:16/ - h.:72/ \mathrm{D} h.:3226/ - h.:6774/$, which is impossible if $h.\cdot/$ is monotonic and increasing, as are canonical link functions. This leads to one final question, which is whether there is any choice of $h.\cdot/$ in Assumption 2 that fits the observed data in Appendix 7. We have already determined that $h.p/ \mathrm{D} p$ does not fit the observed data, because the resulting estimate $E \ Y_{ij} .0/ \mathrm{j} A_i \mathrm{D} a$ is out of range. However, we found that both $h.p/ \mathrm{D} \log.p/$ and

$h.p/ \mathrm{D} \log \mathrm{f}p=.1 - p/\mathrm{g}$ fit the observed data, with estimated $E.Y_{ij}.0/\mathrm{j}A_i \mathrm{D} 1/ \mathrm{D} 0{:}1250$ and $0{:}1124$, respectively. This highlights the predominantly untestable nature of Assumption 2. If the estimated value for $E.Y_{ij}.0/\mathrm{j}A_i \mathrm{D} 1/$ falls outside the range of $Œ0; 1]$, then we will doubt Assumption 2, although it could be that Assumption 2 holds but the data set is an outlying one (more on this in Section 4.) However, if the estimated value for $E.Y_{ij}.0/\mathrm{j}A_i \mathrm{D} 1/$ is inside $Œ0; 1]$, then Assumption 2 may be true or it may be false, as we have just seen.

For the observed data reported in Appendix 7, when $h.p/ \mathrm{D} \log.p/$ is selected, the iterative algorithm fails to converge, and a grid search confirms that there is no solution to the estimating equations. For that same data set, setting $h.p/ \mathrm{D} p$ leads to an estimated $E\ Y_{ij}.0/\mathrm{j}A_i \mathrm{D} a$ of $2{:}33$ and an estimated causal risk difference of $-1{:}83$, which is outside the possible range. Setting $h.p/ \mathrm{D} \log\mathrm{f}p=.1 - p/\mathrm{g}$ leads to an estimated $E\ Y_{ij}.0/\mathrm{j}A_i \mathrm{D} a$ of $0{:}93$; as $E\ Y_{ij}\mathrm{j}A_i \mathrm{D} a$ is estimated at $0{:}5$, this leads to an estimated causal relative risk of $0{:}54$.

When $h.p/ \mathrm{D} \log \mathrm{f}p=.1 - p/\mathrm{g}$, we have also observed that the iterative algorithm may fail to converge, again because there is no solution to the estimating equations at (4). We observed this as part of the simulation study reported in Section 5. Data that caused this to happen are reported in Appendix 7.

Constructing confidence intervals

The estimating equations at (4) are of the form $U.()/ \mathrm{D} \dagger_{h\mathrm{D}1}^{H} \dagger_{c\mathrm{D}1}^{C_h} U_{hc}.()/ \mathrm{D} 0$, where $()$ is a vector of parameters, c indexes primary sampling units (PSUs; e.g., the schools), and h indexes the primary strata. $U_{hc}.()/$ is a sum of weighted estimating equations, with the weighted components each having an expected value of zero unconditionally, but not conditionally upon stratum h. Thus, $U_{hc}.()/$ does not generally have a zero expectation. The parameter $()$ characterizes a superpopulation consisting of an infinite number of groups within the primary strata. One can use a sandwich estimator of variance for the $\hat{()}$, which solves $U.()/ \mathrm{D} 0$. The sandwich estimator is based on a Taylor series linearization [32, 33, 36] and has the form

$$\widehat{\mathrm{var}}\,\hat{()} \mathrm{D} \left\{ \mathrm{r}U\,\hat{()} \right\}^{-1} V\,\hat{()} \left\{ \mathrm{r}U\,\hat{()}^{T} \right\}^{-1} \tag{8}$$

where $\mathrm{r}U.()/$ is the gradient of $U.()/$ with respect to $()$, and

$$V\,\hat{()} \mathrm{D} \dagger_{h\mathrm{D}1}^{H} \mathrm{f}C_h=.C_h\text{-}1/\mathrm{g} \dagger_{c\mathrm{D}1}^{C_h} \left\{ U_{hc}\,\hat{()} - U_{h:}\,\hat{()} \right\} \left\{ U_{hc}\,\hat{()} - U_{h:}\,\hat{()} \right\}^{T} \tag{9}$$

where $U_{h:}\,\hat{()} \mathrm{D} .1=C_h/\dagger_{c\mathrm{D}1}^{C_h} U_{hc}\,\hat{()}$. By the law of large numbers and the central limit theorem, $\hat{()}$ is approximately distributed as multivariate normal with mean $()$ and variance $\widehat{\mathrm{var}}\,\hat{()}$.

Unfortunately, even though our estimate of $\hat{()}$ is easy to program, the sandwich estimator of variance is not. An alternative estimator of $\widehat{\mathrm{var}}\,\hat{()}$ that is much easier to program is the bootstrap or jackknife for complex survey data [32–35]. The bootstrap resamples PSUs within each stratum with replacement and reestimates. Let $\hat{()}^{b}$ be an estimate of $()$ based on the data from the b^{th} bootstrap sample, then

$$\widehat{\mathrm{var}}_B\,\hat{()} \mathrm{D} \mathrm{f}1=.B - 1/\mathrm{g} \dagger_{b\mathrm{D}1}^{B} \left\{ \hat{()}^{b} - .1=B/\dagger_{b\mathrm{D}1}^{B} \hat{()}^{b} \right\}_{2} \tag{10}$$

where B is the total number of bootstrap samples, is the bootstrap estimator of variance, which can be used with a normal approximation to produce confidence intervals. For estimating confidence intervals for functions $r.()/$ of $()$ such as relative risks, we use the normal approximation to the log of $r\,\hat{()}$.

Unfortunately, in practice, we found the bootstrap to readily generate samples for which the logistic SNM did not fit, in the sense that the estimating equation had no solution. We therefore turned instead to the jackknife. Let $\hat{()}^{hc}$ be an estimate of $()$ based on deleting the c^{th} PSU within stratum h. The jackknife estimator of variance we used is

$$\widehat{\mathrm{var}}_J\,\hat{()} \mathrm{D} \dagger_{h\mathrm{D}1}^{H} \mathrm{f}.C_h - 1/=C_h\mathrm{g} \dagger_{c\mathrm{D}1}^{C_h} \left\{ \hat{()}^{hc} - \hat{()} \right\}^{2} \tag{11}$$

Again, for estimating confidence intervals for functions $r.()/$ of $()$ such as relative risks, we use the normal approximation to the log of $r\,\hat{()}$.

5. Simulation study

We conducted two sets of simulations, the first based on a logistic SNM, with $h.p/ = \log.p=.1 - p//$ in Assumption 2 and the second based on a loglinear SNM, with $h.p/ = \log.p/$ in Assumption 2. For each set of simulations, we simulated data sets with 400 observations that satisfied Assumptions 1 and 2. Our aim was to investigate bias of the estimators of $RR.a/$ as well as of the jackknife estimator of variance of those estimators. For both sets of simulations, we let $Z_i = 0; 1;$ or 2 with equal probability. Then we generated A_i dependent on Z_i as follows. We let $A_i = 0; 1; 2$ with probabilities $3=4; 1=8; 1=8$ when $Z_i = 0$, probabilities $1=8; 3=4; 1=8$ when $Z_i = 1$, and probabilities $1=8; 1=8; 3=4$ when $Z_i = 2$. We assumed an unclustered design with simple random sampling, and we generated $Y_i.0/$ according to $P.Y_i.0/ = 1|A_i; Z_i/$ as specified in Table I. For that distribution, $P.Y_i.0/|Z_i/ = 0:22/$, so that $Y_i.0/ \perp Z_i$ and Assumption 1 is satisfied. We then generated Y_i according to either the logistic or loglinear SNM, satisfying Assumption 2. For the logistic SNM, we let $\text{logit}.E.Y_i|A_i = a; Z_i//$-$\text{logit}.E.Y_i.0/|A_i = a; Z_i//$ equal $\log.2/$ for $a = 1$ and 2 $\log.2/$ for $a = 2$. We presented the resulting probabilities $P.Y_i = 1|A_i; Z_i/$ used to generate the observed data in Table I as $P^{1}.Y_i.a/ = 1|A_i = a; Z_i/$. For the loglinear SNM, we let $\log.E.Y_i|A_i = a; Z_i// - \log.E.Y_i.0/|A_i = a; Z_i//$ equal $\log.1:5/$ for $a = 1$ and $\log.2/$ for $a = 2$. We presented the resulting probabilities $P.Y_i = 1|A_i; Z_i/$ in Table I as $P^{2}.Y_i.a/ = 1|A_i = a; Z_i/$.

For both SNMs, we calculated $E.Y_i.0/|A_i = 1/ = 0:2125$ and $E.Y_i.0/|A_i = 2/ = 0:2333$. For the logistic SNM, we further calculated $E.Y_i|A_i = 1/ = 0:35$ and $E.Y_i|A_i = 2/ = 0:5417$. Therefore, $RR.1/ = 1:647$ and $RR.2/ = 2:321$, so that $\log.RR.1// = 0:499$ and $\log.RR.2// = 0:842$. For the loglinear SNM, we further calculated $E.Y_i|A_i = 1/ = 0:3188$ and $E.Y_i|A_i = 2/ = 0:4667$. Therefore, $RR.1/ = 1:5$ and $RR.2/ = 2$, so that $\log.RR.1// = 0:405$ and $\log.RR.2// = 0:693$.

We used equation (4) with $h.p/ = g^{-1}.p/$ for estimation with the easily programmed iterative algorithm, where we set $h.p/ = \text{logit}.p/$ for the logistic SNM simulation and $h.p/ = \log.p/$ for the loglinear SNM simulation. Note that the weights W_i in estimating equation (4) can be set equal to one for this simulation. We then estimated $RR.1/$ and $RR.2/$ as well as the jackknife estimator of variance of those estimators, as described in Section 3.

We found that, rarely, the logistic SNM simulation would generate a data set for which there is no solution to estimating equation (4)—we double checked this with a grid search. We simulated until three such data sets were generated. Those data sets were generated at the 734th, 2481st, and 2563rd simulation. Appendix 7 presents the three data sets as well as the expected data set under the model. We observe that the departures of the observed frequencies from the expected frequencies are not that drastic, which surprised us. Using the negative binomial distribution, we estimated the probability of a data set with no solution to the estimating equation at 0.12%.

To assess the bias of our estimating procedure, we simulated 1000 data sets. For the logistic SNM, one of the data sets led to an estimating equation with no solution. For the remaining 999 data sets, we estimated $\log.RR.1//$ at 0.534 with a standard error of 0.013 and $\log.RR.2//$ at 0.874 with a standard error of 0.013. For the loglinear SNM, none of the 1000 data sets led to an estimating equation with

Table I. Specification of probability distributions for the simulation study of Section 4.

		$Z_i = 0$	$Z_i = 1$	$Z_i = 2$
$A_i = 0$	$P.Y_i.0/ = 1\|A_i = 0; Z_i/$	1/5	1/4	1/3
$A_i = 1$	$P.Y_i.0/ = 1\|A_i = 1; Z_i/$	1/4	1/5	1/4
	$P^{1}.Y_i.1/ = 1\|A_i = 1; Z_i/$	2/5	1/3	2/5
	$P^{2}.Y_i.1/ = 1\|A_i = 1; Z_i/$	3/8	3/10	3/8
$A_i = 2$	$P.Y_i.0/ = 1\|A_i = 2; Z_i/$	1/3	1/3	1/5
	$P^{1}.Y_i.2/ = 1\|A_i = 2; Z_i/$	2/3	2/3	1/2
	$P^{2}.Y_i.2/ = 1\|A_i = 2; Z_i/$	2/3	2/3	2/5

The distribution of $P.Y_i.0/ = 1|A_i = a; Z_i/$ is identical for the logistic and the loglinear SNM simulations.

The distribution of the observed data $P.Y_i.a/ = 1|A_i = a; Z_i/$ is given as $P^{1}.Y_i.a/ = 1|A_i = a; Z_i/$ for the logistic SNM and as $P^{2}.Y_i.a/ = 1|A_i = a; Z_i/$ for the loglinear SNM.

no solution. We estimated log.*RR.1//* at 0.420 with a standard error of 0.013 and log.*RR.2//* at 0.731 with a standard error of 0.014. Comparing these values with the truth given earlier, we observe that our estimators are biased slightly high because of the finite sample size.

To study the performance of the jackknife, we simulated 500 data sets as discussed earlier and computed the jackknife confidence intervals for each one. We found that a 95% confidence interval for the coverage for *RR.1/* was *94:2 - 2:0*% for the logistic SNM and *98:4 - 1:1*% for the loglinear SNM. The coverage for *RR.2/* was *95:6 - 1:8*% for the logistic SNM and *94:4 - 2:0*% for the loglinear SNM. Thus, the jackknife performs well.

6. Analysis of the school-based water, sanitation, and hygiene intervention

For the school-based WASH analysis, we defined adherence A_i as an ordinal variable representing three levels. Specifically, we defined reference level (0) as inadequate degrees of water treatment, hygiene promotion, or sanitation improvement; level (1) as an adequate degree of exactly one of those three components; and level (2) as an adequate degree of two or more of those components. We were interested in the effect of adherence on school absence; we let Y_{ij} indicate the absence outcome. We needed to adjust for individual-level confounding by grade level X_{ij}; to do so, we estimated $1=W_{ij1}$ using a baseline category logit model [43]. The inverse probabilities of selection into the study represent W_{ij2}. Table II summarizes the observed data for the study. The final column represents the weighted relative frequency of the row, using $W_{ij} \mathrm{D} W_{ij1}W_{ij2}$ as the weight. Table III summarizes the observed number of schools in each intervention and adherence category.

We used linear, logistic, and loglinear SNMs to analyze the effect of intervention adherence on absenteeism for the school-based WASH trial. Validity of our analysis requires Assumptions 1 and 2; we furthermore assumed $h.p/ \mathrm{D} g^{-1}.p/$, so that we could use the simple method of computing. Our colleagues were most interested in $RR.1/ \mathrm{D} E^{W_1}\ Y_{ij}\mathrm{j}A_i \mathrm{D}\ 1\ = E^{W_1}\ Y_{ij}.0/\mathrm{j}A_i \mathrm{D}\ 1$ and

Table II. Summary of the school-based WASH study data.

Y_{ij}	A_i	Z_i	*WRF*
0	0	Control	0.1817
1	0	Control	0.0738
0	0	WH	0.0185
1	0	WH	0.0050
0	0	WHCS	0.0142
1	0	WHCS	0.0013
0	1	Control	0.0632
1	1	Control	0.0147
0	1	WH	0.1289
1	1	WH	0.0325
0	1	WHCS	0.0473
1	1	WHCS	0.0129
0	2	Control	0
1	2	Control	0
0	2	WH	0.1309
1	2	WH	0.0175
0	2	WHCS	0.2025
1	2	WHCS	0.0552

Y_{ij} indicates absence of student j in school i; A_i denotes adherence level (0,1, or 2); Z_i denotes randomization level (control, WH, or WHCS); *WRF* denotes the weighted relative frequency of the row, using the final weight, which is the product of the confounding adjustment weight and the sampling weight.

Table III. The observed number of schools in each intervention (Z_i) and adherence (A_i) category in the school-based WASH study.

	$A_i = 0$	$A_i = 1$	$A_i = 2$
$Z_i = Control$	22	6	0
$Z_i = WH$	3	14	12
$Z_i = WH \cap S$	2	5	22

Table IV. Estimated relative risks *RR.1/* and *RR.2/* and 95% confidence intervals using the linear, logistic, and loglinear SNM approaches.

Approach	*RR.1/*	*RR.2/*
Linear	0.45 (0.42, 0.49)	0.66 (0.63, 0.70)
Logistic	0.41 (0.19, 0.89)	0.69 (0.32, 1.51)
Loglinear	0.40 (0.36, 0.44)	0.72 (0.67, 0.77)

$RR.2/ = E^{W_1}(Y_{ij}|A_i = 2) = E^{W_1}(Y_{ij}.0|A_i = 2/$. In words, *RR.a/* measures the effect of school-level adherence at level a versus level *0* on individual-level absence in terms of a relative risk, among schools observed to have adherence at level a. Table IV presents estimates and 95% confidence intervals on the basis of the jackknife. For the logistic SNM, we also derived and programmed the sandwich estimator of the 95% confidence interval, which was (0.19, 0.86) for *RR.1/* and (0.41, 1.17) for *RR.2/*, similar to the jackknife estimates. In Table IV, we observe that the logistic SNM leads to very wide confidence intervals as compared with the loglinear or linear SNMs. However, all three methods yield similar point estimates. Our colleagues hypothesized that increased adherence to intervention components would reduce absenteeism, and we observed this to be the case. The relative risk is further away from one for the $A_i = 1$ group than it is for the $A_i = 2$ group, but this is due to the different estimates of $E^{W_1}(Y_{ij}.0|A_i = a/$. For example, for the logistic SNM, the estimate for $a = 1$ was 0.49, whereas that for $a = 2$ was 0.26. Therefore, more reduction in risk of absenteeism was possible in the schools with $A_i = 1$, and more reduction was achieved.

For completeness, we also applied the endogenous regressor approach. For this, we estimated $RD.1/ = E^{W_1}(Y_{ij}.1/ - Y_{ij}.0//$ at *-0:24* with a 95% confidence interval of (*-0:21; -0:27*), and $RD.2/ = E^{W_1}(Y_{ij}.2/ - Y_{ij}.0//$ at *-0:09* with a 95% confidence interval of (*-0:08; -0:10*). We estimated $E^{W_1}(Y_{ij}.0//$ at *0:32*. Qualitatively, the results agree with the linear and loglinear SNM approaches, and quantitatively, the unconditional risk differences for the endogenous regressor approach agree identically with the conditional risk differences $E^{W_1}(Y_{ij}.a/ - Y_{ij}.0|A_i = a$ we computed for the linear SNM. However, the estimated expected value of $Y_{ij}.0/$ is very different for the endogenous regressor approach; particularly as for it, we are estimating an unconditional expectation, whereas for the SNM approaches, we are estimating a conditional expectation.

7. Discussion

We have developed methods and software on the basis of SNMs for the analysis of multi-armed cluster-randomized trials with unequal probabilities of sampling individuals. We have applied the methods to analyze the effect of adherence in the school-based WASH study. In the process, we reviewed the relevant literature and critiqued the endogenous regression framework. We developed and applied weighted generalized SNMs to implement our analysis. We showed that computation is straightforward using an iterative application of weighted instrumental variable software and a jackknife method of variance estimation. Software in SAS is available upon request.

In our investigation, we learned that none of the methods for analyzing the effect of adherence is ideal. With the endogenous framework, one needs to assume that either no effect modifiers exist or an opaque assumption holds. With the principal stratification framework, there are too many principal strata in our application. With the SNM framework, one needs to assume no effect modification in a weak sense, but still that may be undesirable. We explained that if effect modifiers are thought to be present, we can

stratify on them, as we did with gender. Another option is to incorporate them as continuous covariates into the SNM—see the Appendix of Hernan and Robins [24]—but that might lead to even more difficulties in terms of non-existent solutions to the SNM estimating equation.

Our simulation study validated our methodology, but it also demonstrated that for certain data sets, the estimating equation has no solution. This is a problem that deserves further study, for two reasons. First, perhaps one could predict from the data set whether the estimating equation has no solution, before applying the algorithm. Second, one might try to adapt the methodology so that an estimate could be obtained for any data set.

Appendix A

An example, including observed data $Z_i; A_i; Y_{ij}$ as well as counterfactual data and error terms, which satisfies the marginal structural model and randomization assumptions but not the endogenous regressor assumptions. One can calculate $E.E_{ij}/ = E\ E_{ij}.A_i/ = -0{:}138 \neq 0$ and $E\ Z_i E_{ij} = -0{:}048 \neq 0$. In this example, $E\ Y_{ij}.1/ - Y_{ij}.0/ = 0{:}6 - 0{:}4 = 0{:}2$. Randomization holds, in that $E.Z_i/ = 0{:}5$ and $E\ E_{ij}.0/ = E.E_{ij}.1// = 0$. The last column represents the frequency of the row divided by the total sample size, n.

Z_i	A_i	$Y_{ij}.0/$	$Y_{ij}.1/$	Y_{ij}	$E_{ij}.0/$	$E_{ij}.1/$	$E_{ij}.A_i/$	freq/n
0	0	0	0	0	-0:6	-0:4	-0:6	0.05
0	0	0	1	0	-0:6	0.6	-0:6	0.08
0	0	1	0	1	0.4	-0:4	0.4	0.04
0	0	1	1	1	0.4	0.6	0.4	0.08
0	1	0	0	0	-0:6	-0:4	-0:4	0.05
0	1	0	1	1	-0:6	0.6	0.6	0.02
0	1	1	0	0	0.4	-0:4	-0:4	0.16
0	1	1	1	1	0.4	0.6	0.6	0.02
1	0	0	0	0	-0:6	-0:4	-0:6	0.05
1	0	0	1	0	-0:6	0.6	-0:6	0.05
1	0	1	0	1	0.4	-0:4	0.4	0.04
1	0	1	1	1	0.4	0.6	0.4	0.05
1	1	0	0	0	-0:6	-0:4	-0:4	0.05
1	1	0	1	1	-0:6	0.6	0.6	0.05
1	1	1	0	0	0.4	-0:4	-0:4	0.16
1	1	1	1	1	0.4	0.6	0.6	0.05

Appendix B

An example of observed data for which the linear SNM estimates the causal risk difference outside the possible range at $-1{:}83$, whereas for the loglinear SNM the iterative algorithm fails to converge. We set $n = 500$.

Z_i	A_i	Y_i	freq/n
0	0	0	0.13
0	0	1	0.12
0	1	0	0.07
0	1	1	0.18
1	0	0	0.1
1	0	1	0.09
1	1	0	0.21
1	1	1	0.10

Appendix C

Three data sets generated according to the logistic SNM in the simulation study of section 4, but for which there is no solution to the estimating equation. Column 4 represents the expected frequency (total

sample size is 400) for the model used for simulation, and columns 5–7 represent the observed frequencies within the data sets. Surprisingly, the departures of the observed frequencies from the expected frequencies are not extreme.

Z_i	A_i	Y_i	E(freq)	freq 1	freq 2	freq 3
0	0	0	80	81	79	84
0	0	1	20	18	12	14
0	1	0	10	14	9	9
0	1	1	6.6667	7	8	6
0	2	0	5.5556	3	9	7
0	2	1	11.1111	3	8	7
1	0	0	12.5	17	9	11
1	0	1	4.17	4	8	3
1	1	0	66.6667	69	70	69
1	1	1	33.3333	36	25	27
1	2	0	5.5556	7	5	3
1	2	1	11.1111	6	18	11
2	0	0	11.1111	12	17	12
2	0	1	5.5556	17	6	13
2	1	0	10	5	17	10
2	1	1	6.6667	8	9	9
2	2	0	50	46	54	40
2	2	1	50	37	35	56

Acknowledgements

We would like to acknowledge the support of the National Science Foundation, the USDA/National Agricultural Statistics, the Department of Education/National Center for Educational Statistics, the Social Security Administration, and the Department of Agriculture/Economic Research Service through grant NSF SES-1115618. The school-based WASH study was funded by the Bill & Melinda Gates Foundation and led by CARE.

References

1. Freeman MC, Greene LE, Dreibelbis R, Saboori S, Muga R, Brumback B, Rheingans R. Assessing the impact of a school-based water treatment, hygiene and sanitation programme on pupil absence in Nyanza Province, Kenya: a cluster-randomized trial. *Tropical Medicine and International Health* 2012; **17**:380–391.
2. He Z. Causal inference with complex sampling designs. *Doctoral Dissertation*, 2012. Permanent Link: http://ufdc.ufl.edu/UFE0044538/00001 [last accessed 22 November 2013].
3. Wooldridge JM. *Econometric Analysis of Cross Sectional and Panel Data*. The MIT Press: Cambridge, MA, 2002.
4. Johnston KM, Gustafson P, Levy AR, Grootendorst P. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine* 2008; **27**:1539–1556.
5. Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American Journal of Epidemiology* 2009; **169**:273–284.
6. Burgess S, Thompson SG. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine* 2012; **31**:1582–1600.
7. Bhattacharya J, Goldman D, McCaffrey D. Estimating probit models with self-selected treatments. *Statistics in Medicine* 2006; **25**:389–413.
8. Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* 2000; **19**:1849–1864.
9. Burgess S. CRP CHD genetics collaboration. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine* 2013; **32**:4726–4747.
10. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**:21–29.
11. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444–455.
12. Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 2000; **29**:722–729.
13. Jo B, Stuart EA. On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine* 2009; **28**:2857–2875.
14. Small DS, Ten Have TR, Joffe MM, Cheng J. Random effects logistic models for analysing efficacy of a longitudinal randomized treatment with non-adherence. *Statistics in Medicine* 2006; **25**:1981–2007.
15. Albert JM. Estimating efficacy in clinical trials with clustered binary responses. *Statistics in Medicine* 2002; **21**:649–661.

16. Jo B, Asparouhov T, Muthen BO, Ialongo NS, Brown CH. Cluster randomized trials with treatment noncompliance. *Psychological Methods* 2008; **13**:1–18.
17. Jo B, Asparouhov T, Muthen BO. Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine* 2008; **27**:5565–5577.
18. Ma Y, Roy J, Marcus B. Causal models for randomized trials with two active treatments and continuous compliance. *Statistics in Medicine* 2011; **30**:2349–2362.
19. Long Q, Little RJA, Lin X. Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework. *Applied Statistics* 2010; **59**:513–531.
20. Cheng J, Small DS. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society, Series B* 2006; **68**:815–836.
21. Cai B, Small DS, Ten Have TR. Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. *Statistics in Medicine* 30:1809–1824.
22. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics Theory and Methods* 1994; **23**:2379–2412.
23. Robins JM. Correction for non-compliance in equivalence trials. *Statistics in Medicine* 17:269–302.
24. Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**:360–372.
25. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B* 2003; **65**:817–835.
26. Korhonen P, Laird NM, Palmgren J. Correcting for non-compliance in randomized trials: an application to the ATBC study. *Statistics in Medicine* 1999; **18**:2879–2897.
27. Ten Have TR, Joffe M, Cary M. Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine* 2003; **22**:1255–1283.
28. Joffe MM, Brensinger C. Weighting in instrumental variables and G-estimation. *Statistics in Medicine* 2003; **22**: 1285–1303.
29. Vansteelandt S, Goetghebeur E. Sense and sensitivity when correcting for observed exposures in randomized trials. *Statistics in Medicine* 2005; **24**:191–210.
30. Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On instrumental variables estimation of causal odds ratios. *Statistical Science* 2011; **26**:403–422.
31. Cain LE, Cole SR, Greenland S, Brown TT, Chmiel JS, Kingsley L, Detels R. Effect of highly active antiretroviral therapy on incident AIDS using calendar period as an instrumental variable. *American Journal of Epidemiology* 2009; **169**:1124–32.
32. Skinner CJ, Holt D, Smith TMF. *Analysis of Complex Surveys*. John Wiley & Sons: Sussex, 1989.
33. Korn EL, Graubard BI. *Analysis of Health Surveys*. John Wiley & Sons: New York, 1999.
34. McCarthy PJ, Snowden CB. The bootstrap and finite population sampling. In *Vital and Health Statistics 2-95*, Public Health Service Publication 85-1369. U.S. Government Printing Office: Washington, D.C., January 1985; 1–23.
35. Shao J. Impact of the bootstrap on sample surveys. *Statistical Science* 2003; **18**(2):191–198.
36. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 1983; **51**:279–292.
37. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
38. Bang H, Davis C. On estimating effects under non-compliance in randomized clinical trials: are intent-to- treat or instrumental variables analyses perfect solutions? *Statistics in Medicine* 2007; **26**:954–964.
39. Matsuyama Y. Correcting for non-compliance of repeated binary outcomes in randomized clinical trials: randomized analysis approach. *Statistics in Medicine* 2002; **21**:675–687.
40. Sitlani CM, Heagerty PJ, Blood EA, Tosteson TD. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Statistics in Medicine* 2011; **31**:1738–1760.
41. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*, 3rd edn. Lippincott Williams & Wilkins: Philadelphia, 2008.
42. McCullagh P, Nelder JA. *Generalized Linear Models*, 2nd edn. Chapman & Hall: London, 1989.
43. Agresti A. *Categorical Data Analysis*, 2nd edn. John Wiley & Sons: Hoboken, 2002.