Estimating cluster-level local average treatment effects in cluster randomised trials with non-adherence

Schadrac C. Agbla[1], Bianca De Stavola[2], and Karla DiazOrdaz*[1]

[1]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK
[2]Faculty of Population Health Sciences, UCL GOS Institute of Child Health, UK

Abstract

Non-adherence to assigned treatment is a common issue in cluster randomised trials (CRTs). In these settings, the efficacy estimand may be also of interest. Many methodological contributions in recent years have advocated using instrumental variables to identify and estimate the local average treatment effect (LATE). However, the clustered nature of randomisation in CRTs adds to the complexity of such analyses.

In this paper, we show that the LATE can be estimated via two-stage least squares (TSLS) regression using cluster-level summaries of the outcome and treatment received under certain assumptions. We propose the use of baseline variables to adjust the cluster-level summaries before performing TSLS in order to improve efficiency. Implementation needs to account for the reduced sample size, as well as the possible heteroscedasticity, to obtain valid inferences.

Simulations are used to assess the performance of TSLS of cluster-level summaries under cluster-level or individual-level non-adherence, with and without weighting and robust standard errors. The impact of adjusting for baseline covariates and of appropriate degrees of freedom correction for inference is also explored. The methods are then illustrated by re-analysing a CRT carried out in a specific UK primary care setting.

*corresponding author

TSLS estimation using cluster-level summaries provides estimates with small to negligible bias and coverage close to nominal level, provided the appropriate small sample degrees of freedom correction and robust standard errors are used for inference.

Keywords: Cluster randomised trials, non-adherence, local average treatment effect, instrument variable, cluster-level analysis.

1 Introduction

Cluster randomised trials (CRTs), which randomise groups of individuals, are common in public health and primary care. The adoption of this design is often justified given the reduction of "cross-over contamination" between the experimental arms and improved adherence with allocated treatment [1–3]. Nevertheless, treatment non-adherence is as common in CRTs as it is in individually randomised trials [4]. Dealing with non-adherence is more challenging because there are at least two levels at which deviations from protocol can occur, e.g. cluster or individual level [5]. We say that adherence is at the cluster-level if all individuals within a cluster receive the treatment the cluster was randomised to. In contrast, we say that adherence is at the individual-level, if the treatment received varies across individuals within the same cluster, so that some individuals received the treatment allocated to their cluster, while others did not.

The standard analysis of randomised clinical trials is intention-to-treat (ITT), which compares average outcomes across randomised groups. However, if the effect of treatment received is confounded, in the sense that there are measured and unmeasured common causes of receiving treatment and experiencing the outcome, the ITT provides the causal effect of being offered, rather than of receiving, the treatment. An ITT analysis with poor adherence may dilute a true treatment effect [6]. Recently, there has been an increased interest in estimating other estimands alongside the ITT, as highlighted by the International Council for Harmonisation addendum to guideline E9 (Statistical Principles for Clinical Trials). Amongst them, the causal effect in those adhering to treatment has been singled out as being of interest for patients [7].

In the presence of unmeasured confounding, instrumental variable (IV) methods can estimate consistently the causal effect of an exposure under certain assumptions [8,9]. An IV is a variable which is correlated with the exposure but is not associated

with any confounders of the exposure–outcome association, nor is there any pathway by which the IV affects the outcome, other than through the exposure.

Since randomised treatment is usually a valid instrument, IV methods have been proposed to estimate the treatment causal effect in the context of randomised clinical trials affected by non-adherence [10,11]. The population to which an IV estimate applies however, depends on the assumed behaviour of the instrument [9]. When, as it is often the case, randomised treatment influences treatment received *monotonically* [9], in the sense that the level of treatment received is greater when randomised to treatment, than when randomised to the control (the precise technical definition will be given shortly), IV methods lead to estimating the causal effect among the adherers, known as the local average treatment effect (LATE) or complier-average causal effect.

This estimand can be estimated via the ratio estimator or the two-stage least squares (TSLS) approach [12]. The latter consists of a "first stage", which regresses treatment received on randomised treatment, and a "second stage", which models the outcome on the predicted treatment received. Additional covariates can be included in each stage to control for measured confounding or increase precision. The regression coefficient for the predicted treatment received in the second stage model is a consistent estimator of the LATE, provided that the first stage model is a linear regression, containing all the variables appearing in the second stage [13, 14].

Extensions of this approach for the estimation of LATE in CRTs have been proposed, ranging from a TSLS of individual-level data with variance inflation by the design effect factor [15], to multilevel mixture models that include the latent compliance class membership as a regressor and a random effect for cluster [16, 17]. An alternative approach suggested by Schochet [18] constructs Wald-type ratio estimators using cluster-level (CL) summaries for both treatment received and outcome.

In this paper we focus on TSLS estimation applied to CL outcome summaries. Similar to [18], this approach exploits well-known methods from cluster-level analysis, which consist of calculating for each cluster a relevant summary measure of the individual-level outcomes, such as means or proportions, and then analysing these using appropriate statistical methods, such as regression. Because each cluster provides only one data point, the units of analysis can be considered to be independent, but the procedure is inefficient [19]. Estimation by weighted least squares, where the weights are defined either by the cluster size or by the so-called minimum variance weights can improve efficiency [15]. Comparing these alternative estimation strategies for the implementation of TSLS estimation using CL data is the focus of this

paper. We also demonstrate that using individual-level covariate-adjusted cluster summaries in the (weighted) TSLS regression can increase efficiency.

The rest of the paper proceeds as follows. Section 2 provides an overview of cluster level analysis methods, defines the estimand of interest, the LATE, and introduces the identification assumptions and the different cluster-level TSLS approaches. The finite-sample performance of the methods considered is evaluated using Monte Carlo simulations, presented in Section 3. In Section 4, we illustrate the methods by re-analysing the TXT4FLUJAB trial, a UK based CRT evaluating the effectiveness and efficacy of text messaging influenza vaccine reminders in increasing vaccine uptake amongst patients with chronic conditions [20]. Section 5 concludes with a discussion.

2 Methodology

Consider a two-arm CRT, with n participants, indexed by i, in J clusters, indexed by j, each of size n_j. Let Z_j denote the binary treatment randomly allocated at the cluster-level with probability 0.50. Let Y_{ij} denote the continuous or binary outcome, and $D_{ij} \in \{0, 1\}$ be the treatment received by individual i in cluster j. Let W_j and X_{ij} be baseline covariates at cluster and individual-level, respectively (which can be vectors of variables).

With a slight abuse of notation, we let Y_j denote the CL outcome (mean or proportions), *i.e.* $Y_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$, hereafter referred to as the unadjusted CL outcome. Analogously, let D_j denote the unadjusted CL treatment received, $D_j = \frac{1}{n_j} \sum_{i=1}^{n_j} D_{ij}$.

In the cluster-level adherence settings, D_{ij} is constant within clusters, and therefore D_j is binary. In contrast, when non-adherence is at the individual level, D_j is a continuous measure that varies from 0 to 1, representing the proportion of individuals receiving the active treatment in cluster j.

2.1 Cluster-level analysis

The unadjusted CL analysis, uses simple CL summary statistics as the outcomes in subsequent analyses. Let σ^2 denote the variance of Y_{ij}, which can be decomposed as $\sigma^2 = \sigma^2_\epsilon + \sigma^2_v$, where σ^2_v is the between-cluster variance and σ^2_ϵ the within-cluster

variance. The intra cluster correlation coefficient (ICC) for Y_{ij} is then $\rho_y = \frac{\sigma_v^2}{\sigma_\epsilon^2+\sigma_v^2}$. The variance of Y_j is

$$\begin{aligned} \text{Var}(Y_j) &= \sigma_v^2 + \frac{\sigma_\epsilon^2}{n_j} = \frac{\sigma_v^2\{1+(n_j-1)\}+\sigma_\epsilon^2}{n_j} \\ &= \frac{\sigma^2+\sigma_v^2(n_j-1)}{n_j} = \sigma^2\left\{\frac{1+\rho_y(n_j-1)}{n_j}\right\}, \end{aligned} \tag{1}$$

where we have used the fact that $\rho_y = \sigma_v^2/\sigma^2$ in the last equality [22].

Since CL outcomes are continuous regardless of whether the original variable was binary, they can be thought to be approximately normally distributed provided n_j is sufficiently large. Thus, a linear regression with CL outcome Y_j as dependent variable and Z_j as the explanatory variable can be fitted to estimate the ITT effects. In the simplest setting without adjustment for other covariates, we have

$$Y_j = \alpha_0 + \alpha_Z Z_j + \eta_j, \tag{2}$$

where η_j is a random error term, assumed to be independently and identically distributed (i.i.d.), with mean 0. The ITT is estimated by α_Z.

Efficiency is gained by estimating this model using generalised least squares, with the weights being either the cluster size n_j, or the so-called *minimum-variance weights* given by [21],

$$\omega_j = \frac{n_j}{1+\rho_y(n_j-1)}.$$

When $\rho_y \approx 0$, minimum-variance weights are approximately equivalent to cluster size weights, while if $\rho_y \approx 1$, minimum-variance weights are approximately 1 [22]. These equivalences can have practical implications when the variance of η_j cannot be consistently estimated, for example if the number of clusters is small, so weighting by the cluster size or even no weights, are viable alternatives. Where clusters are large, weighting by cluster size is inefficient [23].

Since the η_j can be heteroscedastic especially when cluster sizes are very imbalanced, the standard errors should be obtained using a method that takes this into account, such as the Huber-White standard errors (SE) [24] which are consistent when there is heteroscedasticity [25].

Finally, because each cluster now contributes only one observation, inference should be based on the number of clusters J. Therefore, if p is the number of parameters being estimated, hypothesis tests and confidence intervals (CIs) should be based

on appropriate distributions, for example t_{J-p} and not on normal-based approximations. We refer to this as small-sample degrees of freedom (SSDF) correction. Where J is sufficiently large (> 40) normal approximations are adequate.

Regression analyses of CL summary outcomes can only adjust for CL covariates directly, as using CL summaries for individual-level regressors is not appropriate [15]. However, where there is interest in adjusting for baseline covariates at the individual level, whether for scientific reasons or to increase statistical efficiency, this can be done through a two-step procedure [26]. First, an individual-level regression analysis of the outcome is performed incorporating all the relevant covariates into the regression model except for the treatment indicator and ignoring clustering , e.g. with only one covariate X_{ij}, we have:

$$Y_{ij} = \lambda_0 + \lambda_1 X_{ij} + e_{1_{ij}}. \tag{3}$$

In the second step, the sample mean of the fitted residuals for this model $\widehat{e}_{1_{ij}}$ is calculated for each cluster j,

$$e_j = \frac{1}{n_j} \sum_{j=1}^{n_j} \widehat{e}_{1_{ij}}.$$

These are then used as CL outcomes in any subsequent analyses. See the Appendix for the formulation for binary outcomes.

We refer to these summaries as adjusted CL outcomes (adCL). Regression models involving them can also be estimated by generalised least squares, with inference based on normal approximations or Huber-White SEs and/or small-sample degrees of freedom corrections, as before. Of note, if CL covariates are used to compute adCL outcomes, the degrees of freedom must be further reduced by the number of cluster-level regressors used to obtain the CL outcome. No such adjustment is necessary for individual level variables [26]. In this work, we only use adCL outcomes obtained by adjusting for individual level variables. In the remainder, we denote the CL summary outcomes by Y_j, whether they are unCL or adCL, will be clear from the context.

2.2 LATE for CL data

2.2.1 Notation and technical assumptions

Denote by $Y_{ij}(\mathbf{d}_j)$ the *potential outcome* that would manifest if, possibly contrary to fact, the j-th cluster to which the individual belongs receives treatment $\mathbf{d_j}$, a vector

of length n_j of 0s and 1s, where we are assuming *no interference between clusters*, i.e. the potential outcomes and potential treatment received of individuals in the j-th cluster are unrelated to the treatment status of individuals in other clusters [5]. "No interference between clusters" is a special case of partial interference, where individuals can be partitioned into groups such that interference does not occur between individuals in different groups but may occur between individuals in the same group [27]. This is commonly assumed in clustered randomised trials [16, 17]. We also assume *counterfactual consistency*: for $j = 1, \ldots, J$, if $Z_j = z$ then $D_{ij} = D_{ij}(z)$, $Y_{ij} = Y_{ij}(z, D_{ij}(z))$ and $Y_{ij} = Y_{ij}(d)$ for all $i = 1, \ldots, n_j$, and z and d.

The consistency assumption implies that the outcome realised under observation of treatment at level d will equal the potential outcome under a hypothetical intervention to set treatment to value d, regardless of the nature of this hypothetical intervention, in what is called "treatment-variation irrelevance" [28, 29]. More precisely, if we index the different ways of setting the treatment at level d by k_d, the consistency assumption says that $Y_{ij} = Y_{ij}(d, k_d)$, if $D_{ij} = d_{ij}$, no matter the value of k_d.

2.2.2 Estimand of interest and identification assumptions

Assuming no interference between clusters and consistency, allows us to define the estimand of interest, the local average treatment effect (LATE) [8].

In the setting considered here where both Z_j and D_{ij} are binary, the vector of potential treatment received under alternative random allocation, $(D_{ij}(0), D_{ij}(1))$ partitions the participants into four different *compliance classes* [30]: $C_{ij} = n$ (never-takers) if $D_{ij}(0) = D_{ij}(1) = 0$; $C_{ij} = a$ (always-takers) if $D_{ij}(0) = D_{ij}(1) = 1$; $C_{ij} = c$ (compliers) if $D_{ij}(z) = z$ for $z \in \{0, 1\}$; and $C_{ij} = d$ (defiers) if $D_{ij}(z) = 1-z$ for $z \in \{0, 1\}$.

The estimand of interest here is the so-called *population* LATE, defined as

$$\begin{aligned} \beta &= E_j E_i \left[\{ Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0)) \} | C_{ij} = c \right] \\ &= \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \{ Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0)) \} \{ I(D_{ij}(1) = 1, D_{ij}(0) = 0) \}}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} I(D_{ij}(1) = 1, D_{ij}(0) = 0)} \end{aligned} \quad (4)$$

This is said to be a "local" causal effect as it is conditional on the stratum of complier individuals.

Following [5, 31], we write the cluster version of the corresponding identification assumptions [9] as follows:

(A1) Cluster-level unconfoundedness : $Z_j \perp\!\!\!\perp D_{ij}(z), Y_{ij}(z, D_{ij}(z)), \quad z \in \{0, 1\}$.

This is also known as cluster randomisation assumption, and is often stated in terms of the cluster randomization to treatment being independent of measured and unmeasured confounders of the relationship between the treatment-received and the outcome. In the context of cluster randomised trials, we know this holds by design.
(A2) Exclusion restriction at the individual level: Conditional on the treatment received $D_{ij} = d$, the treatment assignment Z_j has no effect on the outcome. In terms of potential outcomes, we have:

$$Y_{ij}(1, d) = Y_{ij}(0, d), \quad \forall d \in \{0, 1\}.$$

(A3) Instrument relevance: Also referred to as first stage assumption:

Z_j is causally associated with treatment received D_{ij}, i.e. $Z_j \not\perp\!\!\!\perp D_{ij}$.

We remark that in the standard TSLS literature, a weaker version of the assumption A3 is made instead, namely that the instrument Z and treatment received D are only associated, but not necessarily causally. Denote this assumption by **(A3')**. In the causal inference literature an "associational" instrument is known as proxy or surrogate instrument. Now, if Z and D are associated but not causally associated, there exists a common cause V, which is the causal instrument, and may be unobserved. In order to define, identify and interpret the LATE causally, under A1-2 and A3', we further require that Z is conditionally independent from D and Y given V, and that V is binary. We refer the interested reader to [32] for further details.

For point identification of local treatment effects, **(A4) monotonicity** of the treatment mechanism is often assumed: $D_{ij}(1) \geq D_{ij}(0)$. In the case of a causal binary instrument, as randomised treatment, the monotonicity asumption implies that there are no individuals who would have received the active treatment when randomised to control ($Z = 0$) and not received it when randomised to it. This assumption is often referred informally to as "there are no defiers" [8]. We remark that the definition and interpretation of the monotonicity assumption is also more complex in non-causal instrument settings [33]. The monotonicity assumption is often justified by design, when the active treatment is not available to those in the control group. Where this is not the case, the investigators have to argue carefully why monotonicity is still plausible.

We also remark that in the case of cluster-level randomisation, but with individual level non-adherence, we need to assume that monotonicity holds at the individual level [31]. For the cluster-level non-adherence setting, where D_{ij} does not vary within clusters, then this becomes monotonicity at the cluster level , i.e. $D_j(1) = 1, \text{and } D_j(0) = 0$.

An extra assumption necessary when using adjusted CL outcomes is that the model used to derive them is correctly specified.

2.2.3 Cluster and individual-level non-adherence

The population-level LATE estimand β can be thought of as a weighted average of the cluster-specific LATE β_j for each cluster j, namely

$$\beta = \sum_{j=1}^{J} \psi_j \beta_j, \tag{5}$$

where

$$\begin{aligned} \beta_j &= E_i \left[Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0)) | C_{ij} = c \right] \\ &= \frac{1}{n_{c,j}} \sum_{i=1}^{n_j} \left[\{ Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0)) \} \{ I(D_{ij}(1) = 1, D_{ij}(0) = 0) \} \right], \end{aligned} \tag{6}$$

with $n_{c,j}$ is the number of individual-level compliers in each cluster j, assumed here to be > 0 for all clusters. The weights corresponding to each β_j are $\psi_j = \frac{n_{c,j}}{\sum_{j=1}^{J} n_{c,j}}$, i.e. the number of cluster-specific compliers divided by the total number of compliers.

This result is useful when interpreting the estimates obtained using CL summaries. We first note that the Wald ratio estimand applied to CL summaries, β_{CL} does not always correspond to the population LATE β. The former can be expressed as [5]:

$$\beta_{CL} = \frac{E[Y_j | Z_j = 1] - E[Y_j | Z_j = 0]}{E[D_j | Z_j = 1] - E[D_j | Z_j = 0]} \tag{7}$$

In the case where treatment received is at the cluster level (*i.e.* cluster-level adherence), this CL Wald estimand indeed can be interpreted as the population LATE.

In the case where non-adherence varies at the individual level, it can be shown that $\beta_{CL} = \sum_{j=1}^{J} \psi_{CL,j} \beta_j$ where the CL-weights are $\psi_{CL,j} = \frac{n_{c,j}/n_j}{\sum_j n_{c,j}/n_j}$, i.e. the normalised

proportion of individual compliers in each cluster [31]. So, CL-LATE β_{CL} identifies the population LATE, equation (5), only if (i) the cluster sizes n_j are identical for all j, or (ii) the cluster-specific LATEs β_j are the same across all clusters, i.e. $\beta_j = \beta$, for all $j \in \{1, \ldots, J\}$.

In the remainder, with individual-level non-adherence we assume that every cluster has the same cluster-specific LATE, but allow for the cluster sizes to vary. If this is not the case, the CL-LATE β_{CL} identifies a weighted average of the heterogeneous cluster-specific LATE, because of clusters with the same proportions of compliers are weighted the same, without accounting for the cluster size.

2.3 TSLS for CL data

The conditional expectations appearing in the Wald estimand (equation 7) can be estimated via standard TSLS regression of the CL summaries (referred to as CL-TSLS). CL-TSLS is most easily explained for settings without weights or covariate adjustment. The first stage fits a regression to CL treatment received D_j on treatment assigned Z_j. Then, in a second stage, a regression for the CL outcome on the predicted treatment received is fitted. This can use either unadjusted CL summaries, or adjusted CL summaries if there are baseline individual-level variables predictive of the outcome, as this can help gain efficiency. Crucially both first and second stages must be linear models for the TSLS estimator to be guaranteed to be consistent [14, 34]. We have:

$$\begin{aligned} D_j &= \gamma_0 + \gamma_Z Z_j + \omega_{1j} \\ Y_j &= \beta_0 + \beta_{IV} \widehat{D_j} + \omega_{2j} \end{aligned} \tag{8}$$

where ω_{1j} and ω_{2j} are assumed i.i.d. with mean zero and constant variance, and such that $\omega_{1j} \perp\!\!\!\perp \omega_{2j}$. The estimate of CL-LATE is then given by $\widehat{\beta}_{IV}$.

The asymptotic variance of this estimator is given by $\frac{\widehat{Cov}(Y,E[D|Z])}{\widehat{Cov}(D,E[D|Z])}$, where we assume that $Cov(D, E[D|Z]) \neq 0$. Assuming that the residuals from the second stage $\epsilon = Y - E[Y] - \beta_{IV} D - E[D|Z])$ are such that $E[\epsilon^2 | Z = z] = \sigma_Y^2$, the asymptotic variance of the IV estimator simplifies to $\sigma^2 (D^\top P_Z D)^{-1}$, where P_Z is the projection matrix $P_Z = Z(Z^\top Z)^{-1} Z^\top Z$. See [8] for further details. These variance estimators are implemented in most commonly used software implementations of TSLS.

Cluster-level covariates can be included in the regressions to increase precision. For

example, with one CL covariate W_j, we have

$$\begin{aligned} D_j &= \gamma_0 + \gamma_Z Z_j + \gamma_W W_j + \upsilon_{1j} \\ Y_j &= \beta_0 + \beta_{IV} \widehat{D}_j + \beta_W W_j + \upsilon_{2j} \end{aligned} \tag{9}$$

where the error terms are as before.

As with the cluster level estimation of ITT, where the number of clusters is small, the CIs are constructed using a t distribution with degrees of freedom equal to $J-p$, where p is the number of parameters estimated by the second stage (i.e. the small-sample degrees of freedom correction). As before, minimum-variance or cluster size weights can be used to increase efficiency. Finally, the error terms in the CL-TSLS are assumed to be homoscedastic. Where this is not a sensible assumption, Huber-White SEs should be used [35].

3 Simulation study

We now perform a simulation study comparing the finite sample performance of TSLS estimation applied to CL data. We simulate CRT individual-level data assuming that the control group does not have access to the active intervention, referred to as one-way non-compliance, at either cluster or individual level. In this setting, there are only two compliance classes: compliers and never takers. With a fixed expected total sample size equal to 1000, we vary the number of clusters J, and the average cluster size n_j. The marginal ICC of Y also takes two values. The effect of individual and cluster level variables on the outcome and the treatment received also varies, so that the strength of the confounding is either low or high, while the value of the true LATE also has two levels. Table 1 summarises the factorial design and the values taken by the different levels.

More specifically, we simulate cluster randomised treatment $Z_j \sim \text{Bern}(0.5)$ and two independent baseline covariates, a cluster-level covariate $W_j \sim N(0, \sigma^2_W)$ and individual-level covariate $X_{ij} \sim N(0, \sigma^2_X)$ with a moderate ICC $\rho_X = 0.05$, and with variance $\sigma^2_W = \sigma^2_X = 0.08$.

We then generate a binary adherence class indicator variable C_{ij}, which is considered as latent. Let $C_{ij} = 1$ for the compliers, 0 otherwise. For settings where adherence is at the cluster level, this is constant within clusters, under the following model

$$\begin{aligned} C_{ij} &= C_j \sim \text{Bern}(\pi_j) \quad \text{with} \quad \pi_j = P(C_j = 1) \\ \text{logit}(\pi_j) &= \lambda_0 + \lambda_W W_j, \end{aligned}$$

with $\lambda_W = 0.05$ equivalent to an odds ratio OR ≈ 1.05 per unit increase in W (denoted "small effect") and $\lambda_W = 0.7$ equivalent to OR ≈ 2 ("large effect").

For settings with individual-level adherence, the data generating model is

$$\begin{aligned} C_{ij} \sim \text{Bern}(\pi_{ij}) \quad \text{with} \quad \pi_{ij} = \pi = P(C_{ij} = 1) \\ \text{logit}(\pi_{ij}) = \lambda_0 + \lambda_W W_j + \lambda_X X_{ij} + \zeta_j \\ \zeta_j \sim N\left(0, \sigma_\zeta^2\right) \end{aligned}$$

with $\sigma_\zeta^2 = \pi^2/3$, so that the ICC for compliance is $\rho_C = \sigma_\zeta^2/(\sigma_\zeta^2 + \pi^2/3) = 0.50$.

We derive treatment received at the individual level as

$$D_{ij} = C_{ij} Z_j,$$

so that those individuals in clusters randomly allocated to control have always control treatment, but those in clusters randomised to the active intervention can switch to the control treatment, depending on their adherence class. We finally generate continuous outcome Y_{ij}, under the exclusion restriction assumption,

$$Y_{ij} = \beta_0 + \beta_C C_{ij} + \beta_{CZ} C_{ij} Z_j + \beta_W W_j + \beta_X X_{ij} + \upsilon_j + \epsilon_{ij} \tag{10}$$

with $\upsilon_j \sim N(0, \sigma_\upsilon^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, where the values for σ_υ^2 and σ_ϵ^2 are chosen such that the marginal ICC for Y has the corresponding value according to the simulated scenario, given that $\text{Var}(Y_{ij}) = \sigma^2 = 1$.

For simplicity but without loss of generality, we assume that there is no direct effect of complying on the outcome, and thus $\beta_C = 0$, so that the mean potential outcome in the control never-takers is equal to the mean potential outcome of the control compliers and thus complying with the control treatment has no effect on the outcome. Since $\beta_{CZ} \neq 0$, there is a non-zero effect of complying with the active arm (i.e. receiving active treatment). The choice of β_C does not affect our estimation, as the effect of β_C (and the intercept β_0) cancels out for the average treatment effect in the compliers and non-compliers respectively.

The choice of the parameters' values is reported in Table 1.

Table 1: Factorial design of the data generating processes and values taken by the parameters in the simulations.

Parameter	Label	Level	Value
CRT size			
n	Total number of individuals	Moderate	≈ 1000
J	Number of clusters and	Moderate clusters	$J = 50, n_j \sim \text{Poi}(20)$
n_j	individuals per cluster	Few large clusters	$J = 10, n_j \sim \text{Poi}(100)$
Baseline variables			
W_j	Cluster-level variable	-	$W_j \sim N(0, 0.08)$
ρ_X	ICC for X_{ij}	Moderate	0.05
X_{ij}	Individual-level variable	-	$X_{ij} = X_j + e_{ij}, X_j \sim N(0, 0.004)$, $e_{ij} \sim N(0, 0.076)$
Adherence to treatment			
π	Expected probability of adherence	Moderate	0.60 (cluster-level adherence) 0.85 (individual-level adherence)
λ_W, λ_X	W_j and X_{ij} effects on log odds of adherence	Small Large	$\lambda_W = 0.05, \lambda_X = 0.05$ $\lambda_W = 0.70, \lambda_X = 0.70$
C_j	Cluster-level adherence class	-	$\text{Bern}[\text{expit}(\lambda_0 + \lambda_W W_j)]$
C_{ij}	Individual-level adherence class	-	$\text{Bern}[\text{expit}(\lambda_0 + \lambda_W W_j + \lambda_X X_{ij} + \zeta_j)]$
ζ_j	Cluster-level random effects	-	$\zeta_j \sim N(0, \pi^2/3)$
ρ_C	ICC for C_{ij}	Moderate	0.50
Outcome			
β_0	intercept		β_0=0
β_C	effect of complying amongst controls		β_C=0
β_W, β_X	W_j and X_{ij} effects on outcome Y_{ij}	Small Large	β_W=0.1 σ, β_X=0.1 σ β_W=0.4 σ, β_X=0.4 σ
β_{CZ}	True LATE	Small, Large	0.1 σ, 0.4 σ
ρ_Y	ICC for Y_{ij}	Small, Large	0.05, 0.20

We need the data generating process to result in randomised treatment Z being a valid IV, but with this choice of parameters, some combinations may result in weak instruments, for example, cluster-level non-adherence settings, with only 5 clusters per arm, and the proportion of non-adherent clusters set at 40% (the median proportion of non-adherent clusters reported in [4] being 44.8%). Thus, after creating each dataset, we perform an unadjusted first stage regression of D_j on Z_j and reject simulated datasets where the resulting $F-$statistic is < 10, (Staiger & Stock's rule of thumb for weak instruments [36]). We continue this process until we have 2500 datasets per scenario.

Estimation in each scenario involves using unadjusted CL summary of treatment

received in the first-stage, and either unadjusted or individual-level variable adjusted CL summary outcomes, for the second stage. Each regression in the TSLS was fitted via ordinary least squares or generalised least squares, the latter with either cluster size or minimum-variance weights. We also consider TSLS where each stage model is either unadjusted or adjusted for a cluster-level variable. Finally, we obtain SEs assuming homoscedasticity or Huber-White SEs, and small-sample degrees of freedom-based or normal approximation CIs. A summary is given in Table 2. Details of the Stata code used for analysis are found in the web-Appendix.

Table 2: Overview of TSLS estimation and inference strategies used in the simulation study

Analyses strategy	**Levels**	
CL outcome	Unadjusted	Adjusted for X_{ij}
TSLS adjusted for W_j	No	Yes
Least square method	Ordinary	Weighted
Weights (if using)	CS	MV
SE estimation	Normal theory	HW SE
SSDF correction	No	Yes

CL: cluster level; HW: Huber-White; CS weights: cluster-size weights; MV: minimum variance; SE: standard error; SSDF: Small sample degrees of freedom correction

The performance criteria used are empirical bias and coverage rates of the 95% CIs over the 2500 replicate datasets per scenario. For the bias, we construct a 95% confidence interval (CI) using the Monte Carlo Errors. The coverage rate sampling error given the size of the simulation results in a valid range between 94.1% and 95.9%. See the Appendix for the formal definitions.

3.1 Results

We present the results by plotting the empirical bias with the Monte Carlo Error-based CIs. The coverage rate valid range is represented by horizontal dashed lines.

Figure 1 and Figure 2 report the empirical bias and 95% CI coverage resulting from each of the different CL-TSLS estimators, when adherence is at cluster or individual level respectively, and for scenarios where the true LATE is large. The corresponding figures for small true LATE are in the Appendix, Figure 7 and 8.

Each figure reports results where $J = 10$ (Panel A, top) or $J = 50$ (Panel B), and with the ICC for Y, ρ_Y, is either small (first three columns) or large (last three columns). In each cell, the results for alternative combinations of TSLS (unadjusted/adjusted for W_j) applied to unCL or adCL outcomes are plotted along the horizontal axis.The different data generation scenarios are identified by $*, +, \times$, and $\circ$, corresponding to varying strengths of the effects of X and W on Y.

We see that all CL-TSLS estimators show some finite sample bias in settings where the number of clusters is small ($J = 10$, Panel A), regardless of whether the non-adherence was at the cluster or individual level and whether the CL summary for Y was adjusted or unadjusted, or W_j was included or not in the TSLS regressions. However, the Monte-Carlo error CIs includes 0 in many settings. The bias is more severe when the ICC for Y is larger (right hand side of each Figure). The bias is somewhat attenuated when we adjust for W_j in the TSLS, and the non-adherence is at the cluster-level (Figures 1 and 7). In contrast, for settings with individual-level non-adherence, this adjustment instead increases the bias, especially if W has only a small confounding effect. In these scenarios, the estimates exhibit a small but statistically significant bias, which disappears when the number of clusters is larger (Figures 2 and 8). In general, the bias is not affected by the choice of weighting strategy, nor by whether ρ_Y is small or large. The bias is negligible for settings where the number of clusters is moderate or large ($J = 50$).

Comparing the results of the 2^{nd}, 3^{rd}, and 4^{th} rows in each panel (Figures 1 and 2), we see that the coverage rate is affected by the choice of SE estimation and also by whether small-sample degrees of freedom correction is used. When the number of clusters is small, a small-sample degrees of freedom correction must be used as failing to do so results in under-coverage (Panels A). The low coverage is more serious when TSLS adjusts for W (second and fourth set of results in each panel).

Overall, the results in Panel A of each figure show that coverage is closer to the nomical levels when using small-sample degrees of freedom correction when constructing CIs (3^{rd} and 5^{th} rows). Using Huber-White SE or not has little to no impact if there is no small-sample degrees of freedom correction. However, the small-sample degrees of freedom correction for the CIs resulting from a TSLS using unCL outcomes can lead to under-coverage, as shown in the specific case with cluster-level non-adherence, large ρ_Y, and large true LATE, but where only X is strongly associated with Y (Figures 1 and 7, 3^{rd} and 5^{th} rows of Panel A, right hand side columns, scenario represented by $\times$ in the plots). The use of adCL outcomes (*i.e.* where the CL outcome is the residual after adjusting for individual level variable X) recovers coverage close to nominal. This is not the case when the non-adherence is at the

individual level, and both W and X are confounders in the data generating process.

In both cluster and individual-level non-adherence settings, it can be seen that using minimum-variance weights increases the coverage by a small fraction, when compared with cluster size weights, especially for scenarios with $J = 50$ and large ρ_Y. However, since minimum-variance weights require an estimate of the cluster-level variance, and this is badly estimated when the number of clusters is small ($J = 10$), we can see that minimum-variance weights are less efficient than using either no weights or cluster size weights. This is most clearly seen when no Huber-White SE correction has been used.

We can also see that when small-sample degrees of freedom correction is used, then not using Huber-White SE can result in small over-coverage especially for cluster-level non-adherence settings, which is improved when Huber-White SE are used (Figures 1 and 7, 3^{rd} and 5^{th} rows of Panel A). When $J = 50$ (Panel B), the use of small-sample degrees of freedom-based distributions is not expected to make any material difference, and this is indeed the case. The impact of using Huber-White SE or the different weighting strategies is also minimal.

3.2 Additional simulations

Two extra additional scenarios are now considered to investigate the sensitivity of the CL-TSLS performance to number of clusters and cluster size imbalances, at both cluster and individual level adherence, but focusing on settings where confounding is strong with a large true LATE.

In the first additional simulation, we explore the impact that the outcome ICC and the number of clusters have on bias, while leaving the expected total sample size fixed (= 1000).

We consider two marginal ICC for Y_{ij} ($\rho_Y = 0.05$ and $\rho_Y = 0.80$) and three average cluster sizes ($n_j = 20$, 10 and 2.5, corresponding to whether the number of clusters varied from $J = 50, 100$ or 400), which includes one of the scenarios previously considered in the main simulations for comparison. Though CRTs rarely have ICCs above 0.10 [37], the value of $\rho_Y = 0.80$ is included to evaluate the performance of the methods in extreme settings.

In the second additional set of simulations, we explore the effect of high cluster size imbalances. While keeping the average sample size equal to 1000, and $J = 10$

or 50, we create high cluster size imbalance using a Pareto distribution to generate the cluster sizes [38]. The Pareto distribution parameters are chosen so that approximately 40% of the clusters have a size below 15, and 60% a size above 15, while the average cluster size is 20 and the minimum cluster size is 10, resulting in approximately 1.8 for the shape and 9.1 for the scale.

3.2.1 Results

Figures 3 and 4, corresponding to cluster and individual level non-adherence settings, show that for a fixed number of clusters (cells in the same row), the bias increases with increasing ICC for Y, but that as the number of clusters increases (moving down the column in the Figure), CL-TSLS results in negligible mean bias, even a very large ρ_Y. It is well known that TSLS is only asymptotically unbiased, and with CL analyses, we expect the asymptotics to depend on the number of clusters, and not the number of individuals. Nevertheless, the CL-summaries treated as outcomes for the two models involved in TSLS contain less "information" when the ICC is higher, which translates into a larger number of clusters being necessary for the bias to be negligible.

The impact of high cluster size imbalance is reported in Figures 5 and 6, where non-adherence is at the cluster and individual level respectively. We see that even with the use of Huber-White SE, failure to do small-sample degrees of freedom correction results in under-coverage, especially when is ρ_Y large. Looking at Panel B in Figure 6, we can see that using cluster-size weights results in even lower coverage. This is because cluster size weights are known to perform well when the cluster level residuals are homoscedastic, which is unlikely when cluster sizes are very imbalanced [15]. The use of small-sample degrees of freedom correction brings the coverage close to nominal levels.

4 Illustrative example

We now illustrate the methods in practice by applying each in turn to the analysis of the TXT4FLUJAB trial. This was a CRT of general practices in England aiming at estimating the effect of text messaging influenza vaccine reminders on increasing vaccine uptake in patients with chronic conditions, carried during the 2013 influenza season [20]. General practices (GPs) were stratified by the type of software used for text messaging and randomised to either standard care (control group, 79 GPs

and 51136 patients) or a text messaging campaign (active group, 77 GPs and 51121 patients). Practices were not blinded to their allocation. GPs were the unit of analysis and the outcome of interest was the proportion of influenza vaccine uptake at the GP level.

Influenza vaccination within the GPs was automatically recorded in the clinical system from which the data were extracted, so there are no missing data.

Since non-adherence was anticipated, the original statistical analysis plan specified obtaining by IV regression an efficacy estimate at the GP level [20]. The original publication reported an estimated increase in vaccine uptake from texting reminders of 14.3% (95% CI –0.59% to 29.2%) [20], after dichotomising adherence at the cluster-level as either 100% of eligible patients, compared with texting $< 100\%$.

Adherence to the intervention at the individual level could not be measured for all practices because it was recorded in a usable form only for GPs using a specific software. Therefore, for these re-analyses, we restrict the dataset to 116 GPs (58 in the intervention and 58 in the standard care arm) for which individual-level adherence data are available. Six of the 58 practices (10%) in the intervention arm, did not send any reminders. Conversely, 21 of the 58 practices (36% in the standard care arm actually sent a reminder to at least one patient. Hence non-adherence is two-sided. It also varies at the individual level. The median (range) of percentage of non-adherence at the GP level was 0% (0%-78.4%) and 21.0% (0%-83.5%) in the control and active group, respectively (Table 3).

The characteristics of the GPs and of the patients included in these analyses are comparable across trial groups (Table 3); further the marginal ICC for individual-level outcome (vaccination) and treatment received (text message reminder) was 0.03 and 0.84 on the log-odds scale, respectively.

Table 3: Baseline characteristics and percentages of non-adherence for the TXT4FLUJAB trial.

Characteristics	**Control**	**Active**
Practice-level characteristics		
Number of practices, n (%)	58 (100.0)	58 (100.0)
Open on weekends, n (%)	39 (67.2)	37 (63.8)
Patients per practice, median (range)	660 (148-1678)	684 (79-3022)
Patient-level characteristics		

Continued on next page

Table 3 Continued

Characteristics	Control	Active
Number of patients, n (%)	40633 (100)	41073 (100)
Male, n (%)	20 752 (51.1)	21 012 (51.2)
Has any disease, n (%)	39244 (96.6)	39672 (96.6)
Age, median (range)	50 (18-64)	50 (18-64)
Active treatment received		
Patients receiving text message reminders, n (%)	2628 (6.5)	11113 (27.1)
Practices sending text message reminders, n (%)	21 (36.2)	52 (80.7)
% of patients in each GP receiving reminders, median (range)	0 (0-78.4)	21.0 (0-83.5)

For our re-analysis, we begin by discussing the plausibility of the necessary assumptions.

Firstly, the consistency assumption in this setting implies that the means of sending and receiving the text message, as well as the timings are irrelevant, in the sense that all of these would lead to the same observed outcome. So whether the text was pre-programmed, or sent by a doctor or a nurse, or received in the morning or at night or weekend, it would have the same effect, of either getting the patient to get a flu vaccine, or not, irrespective of any of these factors. This seems a reasonable assumption for this intervention.

In contrast, there is a small risk of interference. The cluster defined by GP practice should minimise this, as we only need to assume no interference at the cluster level, but it could be plausible that patients interact with those outside their GP, so that the exposure to a text message reminder of one patient may indeed affect the potential outcome, in this case, influenza vaccination of another patient from a different GP. The risk is small as usually close family members belong to the same general practice.

Now, regarding the identification assumptions, we note that the unconfoundedness of the CL randomised treatment assumption is satisfied by design. To check whether cluster randomisation is a relevant instrument, we perform a test on the first stage of the CL-TSLS. The corresponding F-statistic is $F(1, 114) = 28.7 > 10$ thus passing Staiger and Stock's rule of no null first-stage [36].

The exclusion restriction at the individual level implies that there is no other mech-

anism by which the GP being randomised to sending text vaccination reminders can affect a patient's actual vaccination uptake beside via the sending of the message. This assumption needs further justification, as in principle, a GP randomised to send reminders can be more conscious of the risks the patients face during the influenza season and use other means to remind at-risk patients, either in person, by post or by putting out flyers and posters in the clinic. So, it is possible that there are patients who do not receive text reminders and yet are prompted to get vaccinated by other means, by virtue of their practice being in the active group. However, flyers, posters and postal letters already form part of regular care, so we believe they do not really vary by whether the GP is randomised to the active group.

Finally, the monotonicity assumption (that there are no defiers) also seems plausible as GPs randomised to the active group were more likely to send a text message reminder than those in the control group (see Table 3).

CL-TSLS on unadjusted CL outcomes was implemented adjusting and not adjusting for a baseline CL covariate, namely whether the clinic was open on the weekends (yes/no). Table 4 shows the CL-LATE estimates (expressed as mean risk differences), with 95% CIs and p-values obtained via different weighting strategies, and corrections.

Using cluster size weights results in different point estimates from the rest. This was expected as there is substantial cluster size imbalance (cluster size range: 148–1678 in the control group and 79–3022 in the active group (Table 3). The results obtained using no weights or minimum-variance weights leads to point estimates that are very close to those found in the original publication [20].

In terms of inference, the use of small-sample degrees of freedom correction in calculating CIs is not important, as the number of clusters is large, but the Huber-White SEs paired with minimum-variance weighting provides efficiency gains, especially for the adjusted CL-TSLS analyses. Overall however, the CIs are still very wide.

These results suggest that there is some evidence that receiving a text reminder increases the expected proportion of patients within a compliant practice that get vaccinated against influenza by 14% (95% CI: -0.5 to 29.3%, $p = 0.058$, based on the adjusted CL-TSLS using minimum-variance weights and normal-based CI with Huber-White SEs estimate).

Contrast this with the unadjusted CL-summaries mean risk difference ITT estimate, which indicates a 2.89% increase (95% CI -0.17 to 5.95, $p = 0.064$), highlighting the dilution effects deriving from the non-adherence.

One of the disadvantages of TSLS is lack of efficiency. Adjusting for individual-level baseline covariates may help obtaining narrower CIs. Since CL-TSLS cannot adjust for individual level covariates, we now perform the analyses using adCL summary outcomes, generated by adjusting for gender, age and the presence of disease. Results are reported in Table 6 in the Appendix. The results do not materially change (weak evidence of a 13% increase vaccination uptake), possibly because these individual level covariates are not strongly associated with the outcome.

The illustrative example shows the importance of choosing and pre-specifying the TSLS analysis according to the trial characteristics. In the example, the choice of weights changed the point estimate to such an extent that the small evidence in support of treatment benefit disappeared completely. So, if the trial has very imbalanced cluster sizes, Huber-White corrections can help for the SEs, but the point estimates may be biased, if large clusters are somewhat atypical.

Our application is limited by the availability of baseline cluster-level variables. Since there was only one CL-variable recorded, the impact of covariate adjustment on the CL-TSLS is negligible. Other limitations of these results include the possibility of measurement error, for if patients received their influenza vaccine outside the practice, this would not have been recorded in the system, unless the patient informed their GP.

Table 4: LATE of text message reminders to receive flu vaccination on the uptake of flu vaccine in the TXT4FLUJAB trial, using unadjusted CL-summaries

		Unadjusted		**Adjusted[a]**	
		LATE (95% CI)	**p**	**LATE (95% CI)**	**p**
No weighting	None	0.149 (-0.006,0.305)	0.060	0.148 (-0.078,0.303)	0.063
	HW	(-0.006,0.305)	0.060	(-0.005,0.301)	0.058
	SSDF	(-0.009,0.308)	0.065	(-0.012,0.308)	0.069
	SSDF + HW	(-0.009,0.308)	0.065	(-0.009,0.305)	0.064
Cluster size weights	None	0.071 (-0.065,0.207)	0.307	0.074 (-0.061,0.209)	0.284
	HW	(-0.088,0.230)	0.382	(-0.077,0.225)	0.338
	SSDF	(-0.068,0.209)	0.313	(-0.064,0.212)	0.292
	SSDF + HW	(-0.091,0.233)	0.388	(-0.081,0.228)	0.346
Minimum-variance weights	None	0.143 (-0.008,0.293)	0.064	0.142 (-0.009,0.293)	0.065
	HW	(-0.006,0.291)	0.060	(-0.005,0.289)	0.058
	SSDF	(-0.011,0.296)	0.069	(-0.012,0.297)	0.071
	SSDF + HW	(-0.009,0.294)	0.065	(-0.008,0.293)	0.064

Continued on next page

Table 4 Continued

Weighting strategy	SE & correction	Unadjusted LATE (95% CI)	p	Adjusted[a] LATE (95% CI)	p

[a] Adjusted for whether clinic is opened during weekends.

HW: Huber-White; SSDF: small sample degrees of freedom.

5 Discussion

This paper demonstrates the use of TSLS regression applied to CL summaries (CL-TSLS) as a simple and valid method for obtaining estimates of the LATE in CRTs where non-adherence occurs at either the cluster or the individual level. To improve efficiency of CL-TSLS estimates, we proposed adjusting for baseline variables; if these are cluster-level, in the TSLS regression, while if these are individual-level, by adjusting the CL-summary outcomes before performing TSLS. The performance of CL-TSLS regression of either adjusted or unadjusted CL-outcomes, and adjusting or not for CL-baseline variables was evaluated with different weighting strategies (none, cluster size, minimum variance), as well as the use of different methods for constructing CIs (alternatively using or not Huber-White SEs and/or small-sample degrees of freedom correction) in a factorial simulation study.

We have demonstrated empirically through simulations, that under the stated sufficient assumptions for identification, TSLS regression of CL summaries provides consistent estimates of the causal treatment effect in the sub-population of compliers, where non-adherence is at the cluster level. With individual-level non-adherence, the additional assumption that the cluster-specific LATE is the same across clusters is required for CL-TSLS to identify the population LATE [31]. Moreover, provided that an appropriate distribution with small-sample degrees of freedom adjustment is used when the number of clusters is small and Huber-White SEs are used if there is high cluster size imbalance, valid 95% CIs can be constructed.

Our simulation study suggests that all weighting strategies perform similarly when the number of clusters is not small. When the number of clusters is small, minimum-variance weights tend to be badly estimated, and are not recommended; furthermore when the cluster sizes are very variable, cluster size weights should not be used. Although in the simulations the choice of weights did not affect the point estimates,

these were affected in the illustrative example. Overall our results show that, unless there are very few clusters, or the outcome ICC is large, minimum-variance weighting performs well [23]. An overall summary of these findings in the format of recommendations is given in Table 5.

Although CL-TSLS is easy to implement, it suffers from being very inefficient. We can see this in the illustrative example where all CIs for the CL-TSLS LATE estimates are much wider than those for the estimated ITT. There are two reasons for this: CL-analyses are inefficient, unless the cluster sizes are (almost) equal [39], and TSLS is known to be inefficient, although adjusting for baseline covariates can ameliorate this [14]. In the context of CL-analysis it is only possible to include cluster-level baseline covariates in the regressions [15]. However, we tested the performance of CL outcomes which are adjusted for individual-level covariates [26], and showed that this indeed has the potential to improve efficiency in certain settings.

For CL-TSLS analyses, inference should be based on the number of clusters, with CIs constructed by using $t-$ distributions with degrees of freedom equal to $J - p$ [40]. The outcome ICC value is important too, with higher ICCs requiring a larger number of clusters for the asymptotical arguments to work, as well as whether the cluster-level variances are homoscedastic [15].

Other methods for estimating causal treatment effects in CRTs with non-adherence at the individual level exist, in particular Kang and Keele [31] have recently proposed a finite-sample estimator that identifies the population LATE and obtains valid inferences even when compliance is low.

We do not consider here situations where the identification assumptions are violated. There are several options to study the sensitivity to departures from these assumptions. For example, if the exclusion restriction does not hold, a Bayesian parametric model can use priors on the non-zero direct effect of randomisation on the outcome for identification [41]. Since the models are only weakly identified, the results depend strongly on the prior distributions. Alternatively, violations of the exclusion restriction can also be handled by using baseline covariates to model the probability of compliance directly, within structural equation modelling via expectation-maximisation [42, 43].

We have only focused on LATE estimands. These are often criticised because the estimates obtained apply to the "compliers" in the population, and these cannot be observed in practice, thus limiting applicability. However, LATE estimates may be used to provide information about the average causal effect in the entire population [44]. Moreover, the average treatment effect on the compliers is often of interest

Table 5: Recommendations

Adherence	**Comments**
At CL:	
If the number of clusters J is small	Use small sample DF correction to improve inference
If J is small and the outcome ICC is large	Adjust for CL variables in TSLS to reduce bias and improve efficiency
If an IL variable is a strong confounder	Use adjusted CL-outcomes in the TSLS to improve efficiency
If CS is imbalanced	Use small sample DF correction to improve inference and avoid using CS weights
At IL:	
If the number of clusters J is small	Use small sample DF correction to improve inference
If J is small and the outcome ICC is large	avoid adjusting for CL variables
If CS are imbalanced	Use small sample DF correction to improve inference and avoid using CS weights

CS: cluster sizes; CL: cluster level; DF: degrees of freedom; IL: individual level

to patients and medical decision makers, especially when they expect patients to comply with the treatment [45].

Acknowledgements
We thank the TXT4FLUJAB study team for access to the data. SCA was funded by a UK Economic and Social Research Council PhD scholarship ES/J500021/1. KDO was supported by UK Medical Research Council Career development award in Biostatistics MR/L011964/1.

Conflict of interest
The Authors declare that there is no conflict of interest.

List of acronyms used in Tables and Figures.
HW: Huber-White
CS: cluster-size
MV: minimum-variance
SSDF: small-sample degrees of freedom
unCL: Cluster-level unadjusted outcome summaries
adCL: Cluster-level outcome summaries adjusted for individual-level baseline-variable
DF: degrees of freedom
IL: individual level

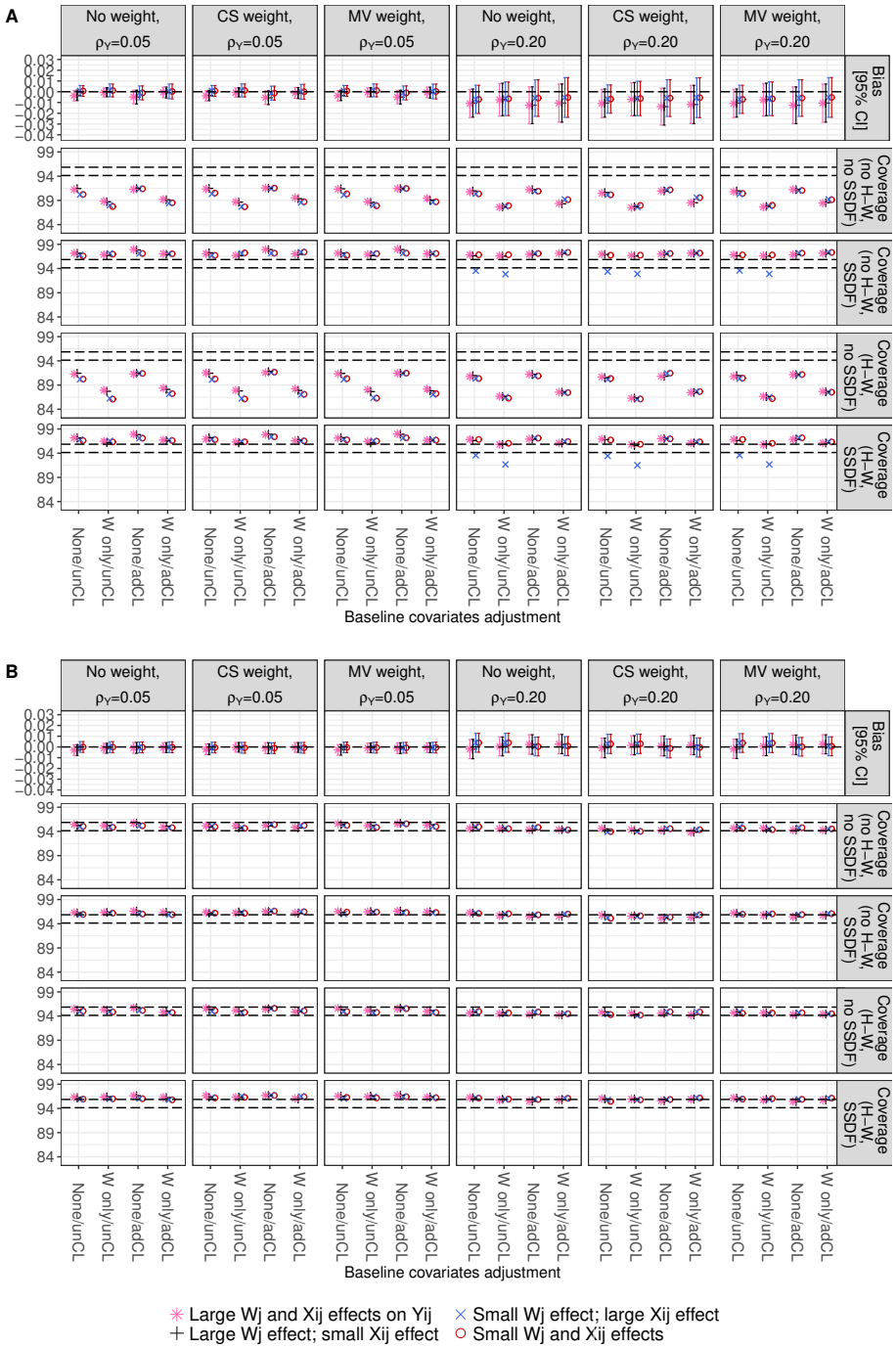

Figure 1: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with cluster-level non-adherence and large true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B .

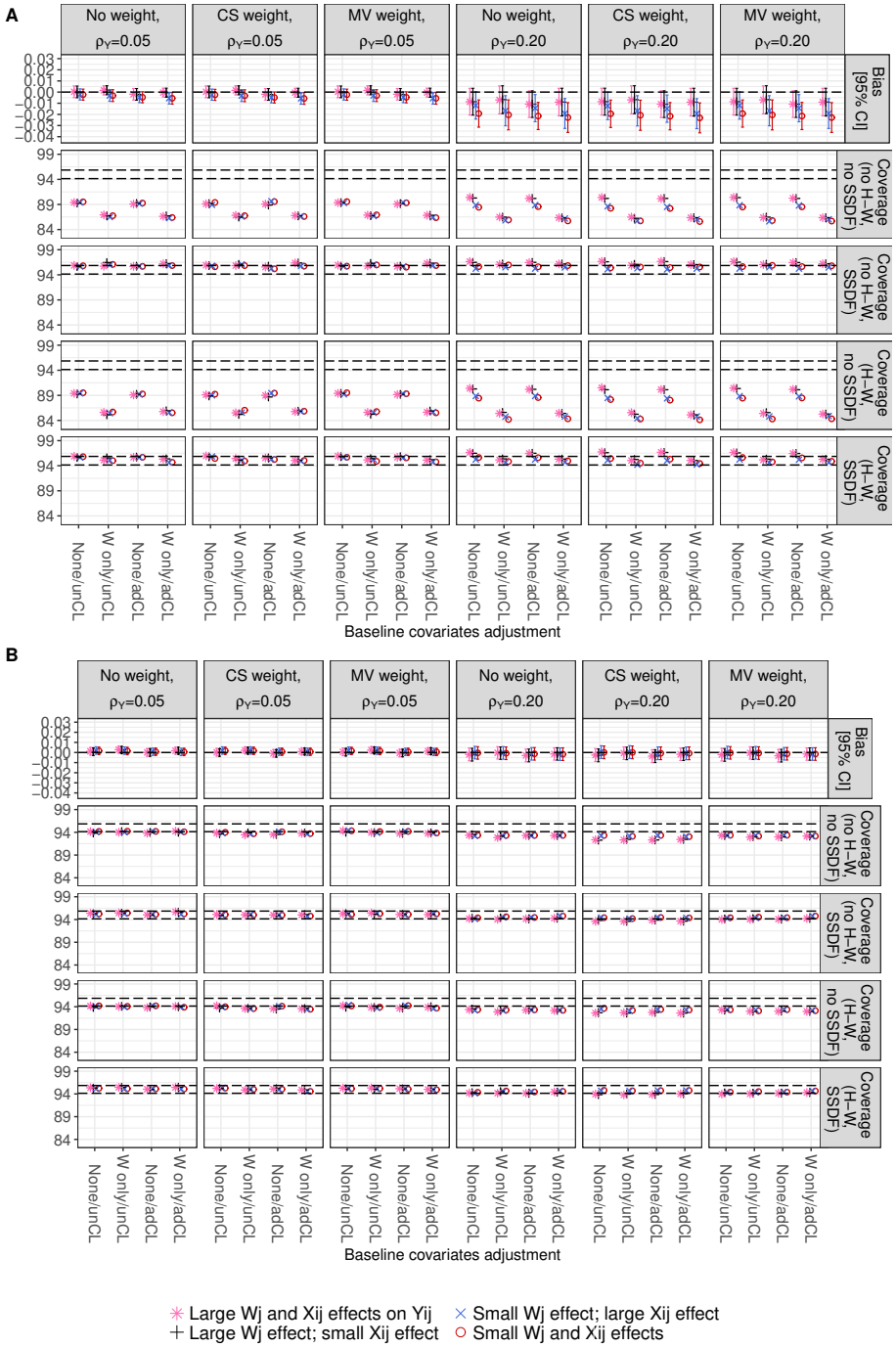

Figure 2: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with individual-level non-adherence and large true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B .

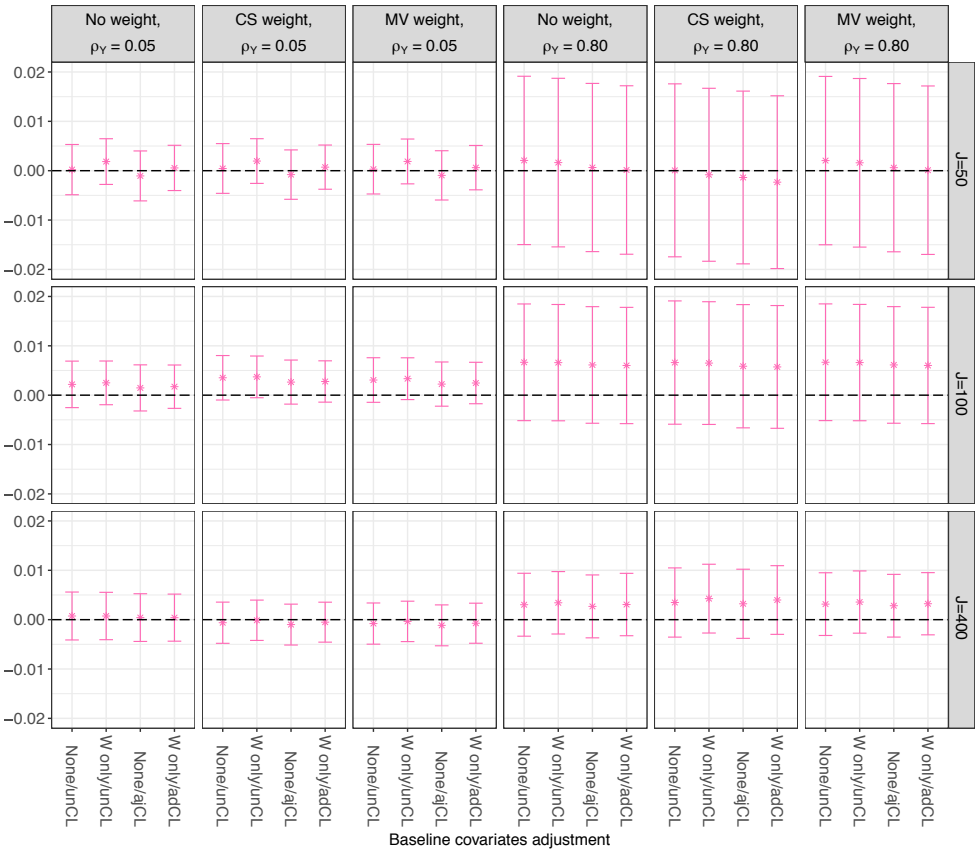

Figure 3: Bias of the CL-LATE for the extra simulation where non-adherence is at the cluster-level and a large true LATE, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Number of clusters varies by rows and ICC by column.

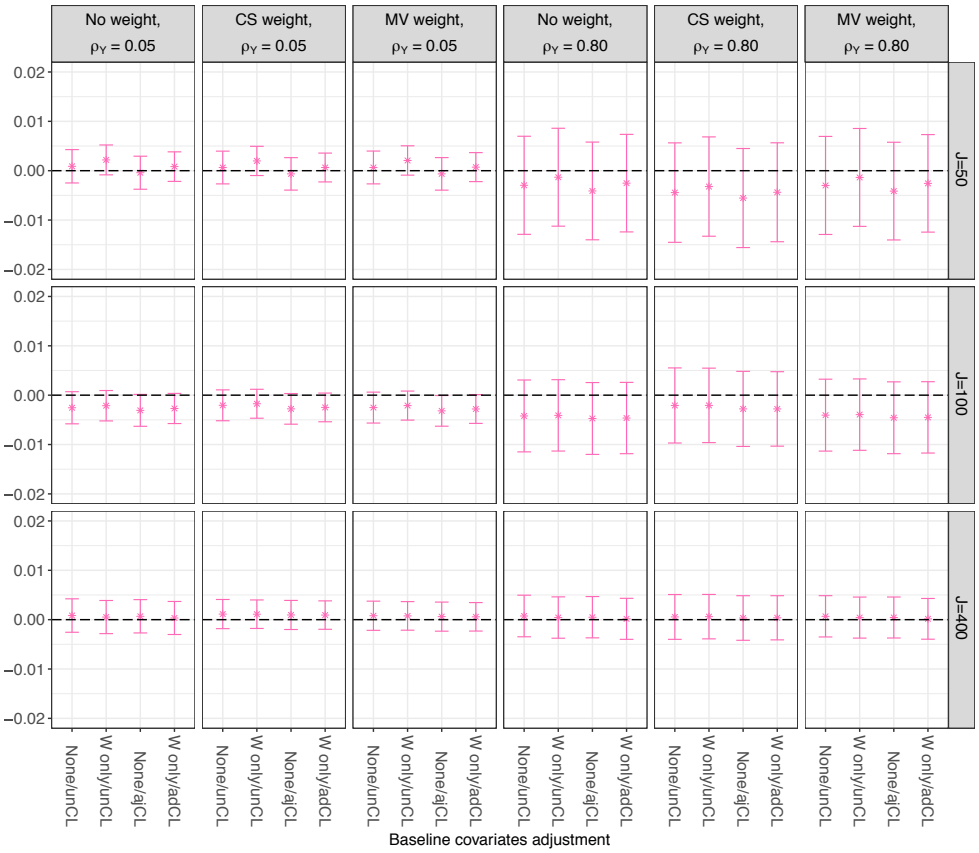

Figure 4: Bias of the CL-LATE for the extra simulation where non-adherence is at the individual-level and a large true LATE, with high ICCs and varying numbers of clusters. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Number of clusters varies by rows and ICC by column.

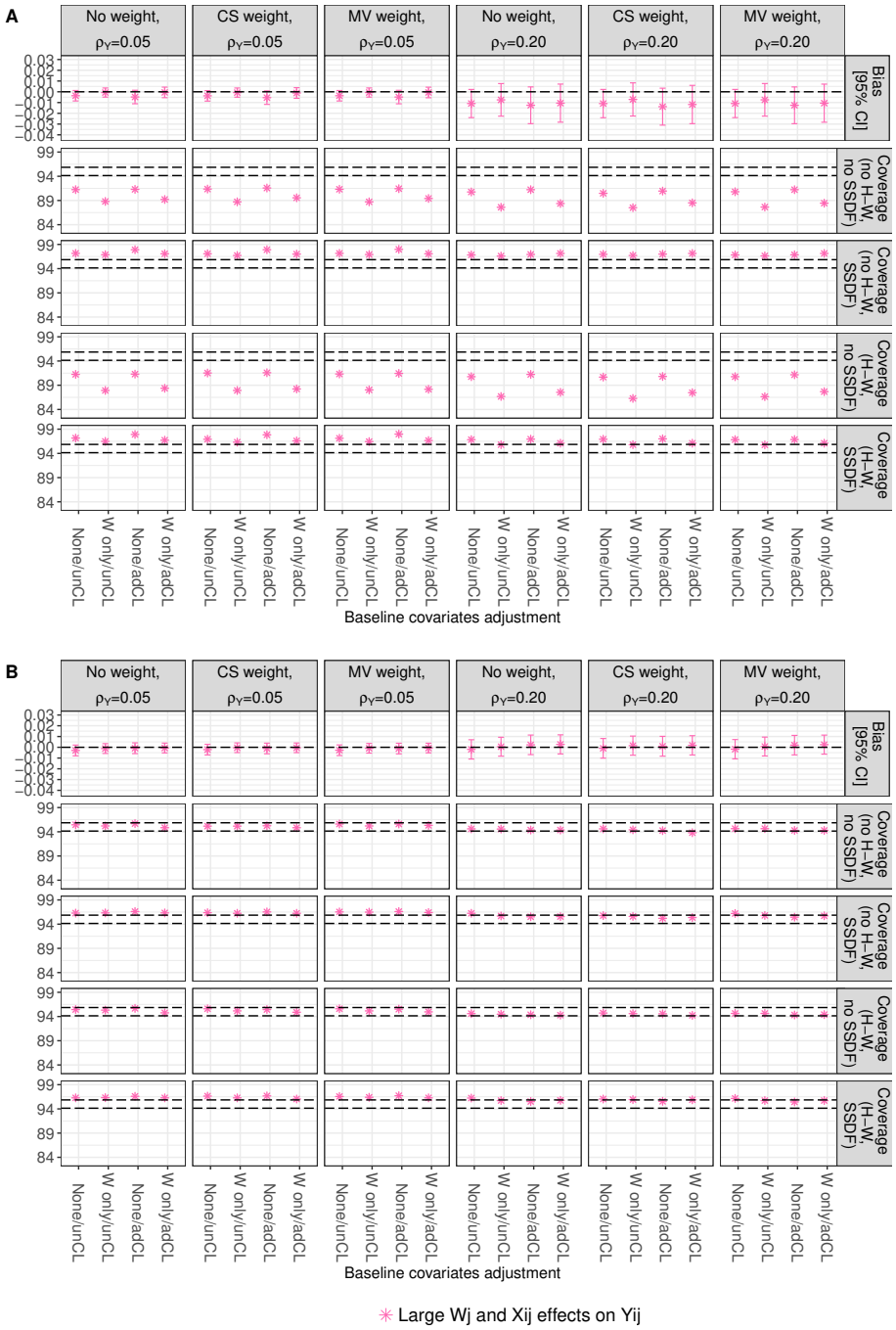

Figure 5: Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage (Huber-White SEs (or not) and SSDF corrections (or not)) of the CL-LATE where non-adherence is at the cluster-level, and a large true LATE. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Small and large number of clusters results appears in Panel A and B respectively.

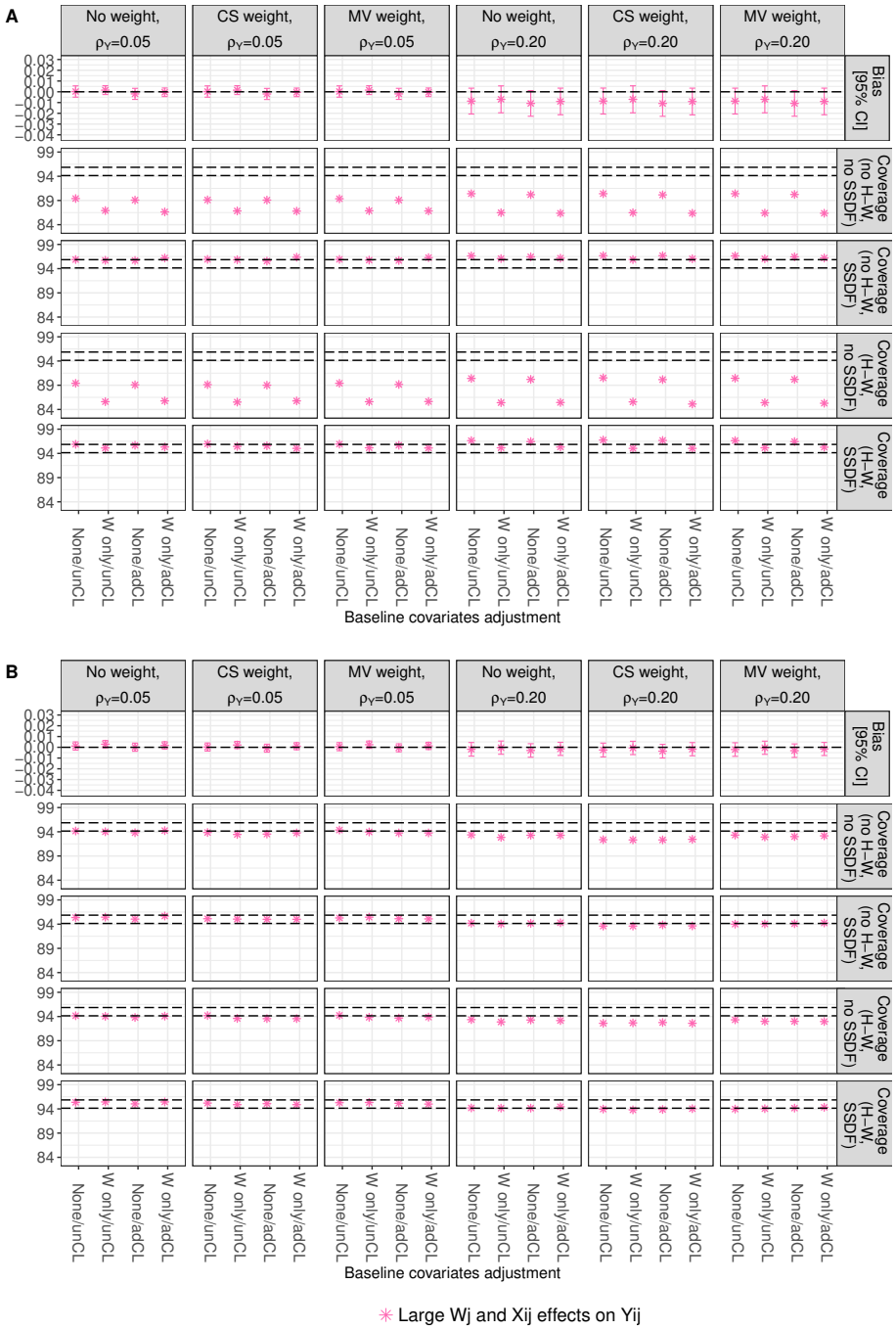

Figure 6: Extra simulation for very imbalanced cluster size settings. Bias (top row) and 95% CI coverage (Huber-White SEs (or not) and SSDF corrections (or not)) of the CL-LATE where non-adherence is at the individual-level, and a large true LATE. Estimates are obtained via unadjusted or adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)). Small and large number of clusters results appears in Panel A and B respectively.

References

[1] Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. American Journal of Public Health. 2004;94(3):416–422.

[2] Glynn RJ, Brookhart MA, Stedman M, Avorn J, Solomon DH. Design of cluster-randomized trials of quality improvement interventions aimed at medical care providers. Medical Care. 2007;45(10):S38–S43.

[3] Wright N, Ivers N, Eldridge S, Taljaard M, Bremner S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. Journal of Clinical Epidemiology. 2015;68(6):603–609.

[4] Agbla SC, DiazOrdaz K. Reporting non-adherence in cluster randomised trials: A systematic review. Clinical Trials. 2018;15(3):294–304.

[5] Schochet PZ, Chiang HS. Estimation and identification of the complier average causal effect parameter in education RCTs. Journal of Educational and Behavioral Statistics. 2011;36(3):307–345.

[6] Hernán MA, Robins JM. per-protocol Analyses of Pragmatic Trials. The New England Journal of Medicine. 2017;377(14):1391–1398.

[7] Akacha M, Bretz F, Ruberg S. Estimands in clinical trials: broadening the perspective. Statistics in Medicine. 2017;36(1):5–19.

[8] Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. Econometrica. 1994;62(2):467–475.

[9] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American statistical Association. 1996;91(434):444–455.

[10] White IR. Uses and limitations of randomization-based efficacy estimators. Statistical Methods in Medical Research. 2005;14(4):327–347.

[11] Bellamy SL, Lin JY, Ten Have TR. An introduction to causal modeling in clinical trials. Clinical Trials. 2007;4(1):58–73.

[12] Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American statistical Association. 1995;90(430):431–442.

[13] Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: Proceedings of the American Statistical Association, Section on Bayesian Statistical Science; 2000. p. pp. 6–10.

[14] Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. MIT Press; 2010.

[15] Angrist JD, Pischke JS. Mostly harmless econometrics: An empiricist's companion. Princeton university press; 2008.

[16] Frangakis CE, Rubin DB, Zhou XH. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. Biostatistics. 2002;3(2):147–164.

[17] Jo B, Asparouhov T, Muthén BO, Ialongo NS, Brown CH. Cluster randomized trials with treatment noncompliance. Psychological methods. 2008;13(1):1.

[18] Schochet PZ. Estimators for clustered education RCTs using the Neyman model for causal inference. Journal of Educational and Behavioral Statistics. 2013;38(3):219–238.

[19] Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Family practice. 2000;17(2):192–196.

[20] Herrett E, Williamson E, van Staa T, Ranopa M, Free C, Chadborn T, et al. Text messaging reminders for influenza vaccine in primary care: a cluster randomised controlled trial (TXT4FLUJAB). BMJ open. 2016;6(2):e010069.

[21] Prais SJ, Aitchison J. The grouping of observations in regression analysis. Revue de l'Institut International de Statistique. 1954;p. 1–22.

[22] Campbell MJ, Walters SJ. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Statistics in Practice. Wiley; 2014.

[23] Kerry SM, Bland MJ. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Statistics in medicine. 2001;20(3):377–390.

[24] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. 1980;48(4):817–838.

[25] Cameron AC, Trivedi PK. Microeconometrics: methods and applications. Cambridge university press; 2005.

[26] Hayes R, Moulton L. Cluster randomised trials. Taylor & Francis; 2009.

[27] Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. Journal of the American Statistical Association. 2006;101(476):1398–1407.

[28] VanderWeele, Tyler J, Concerning the consistency assumption in causal inference, Epidemiology, 2009, **20**(6), 880–883.

[29] Cole, S. R and Frangakis, C. E., The consistency statement in causal inference: a definition or an assumption?, Epidemiology, 2009, **20**, (1), 3–5.

[30] Imbens, GW. and Rubin, DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction Cambridge University Press,Cambridge. 2015, DOI: 10.1017/CBO9781139025751.

[31] Kang H, Keele L. Estimation Methods for Cluster Randomized Trials with Noncompliance: A Study of A Biometric Smartcard Payment System in India. arXiv preprint arXiv:1805.03744. 2018.

[32] Hernán, M. A. and Robins, J. M. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006 **17,** 360–372.

[33] Swanson, S. A. and Hernan, M. A. The challenging interpretation of instrumental variable estimates under monotonicity. International journal of epidemiology. 2017 **47**(4):1289–97.

[34] Vansteelandt S, Didelez V. Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators. Scandinavian Journal of Statistics. 2018.

[35] White H. Instrumental variables regression with independent observations. Econometrica: Journal of the Econometric Society. 1982;p. 483–499.

[36] Staiger DO, Stock JH. Instrumental variables regression with weak instruments. Econometrica. 1997;65(3):557–586.

[37] Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Evaluation Review. 2003;27(1):79–103.

[38] Guittet L, Ravaud P, Giraudeau B. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. BMC medical research methodology. 2006;6(1):17.

[39] Donner A. Some aspects of the design and analysis of cluster randomization trials. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1998;47(1):95–113.

[40] Donald SG, Lang K. Inference with difference-in-differences and other panel data. The review of Economics and Statistics. 2007;89(2):221–233.

[41] Conley TG, Hansen CB, Rossi PE. Plausibly exogenous. Review of Economics and Statistics. 2012;94(1):260–272.

[42] Jo B. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. Statistics in Medicine. 2002;21(21):3161–3181.

[43] Jo B. Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications. Journal of Educational and Behavioral Statistics. 2002;27(4):385–409.

[44] Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Statistics in medicine. 2014;33(13):2297–2340.

[45] Murray EJ, Caniglia EC, Swanson SA, Hernández-Díaz S, Hernán MA. Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. Journal of Clinical Epidemiology. 2018;103:10 – 21.

Appendices

A Adjusted cluster-level summaries for binary data

For binary a standard logistic regression model is usually fitted for binary outcomes, which assumes that

$$\text{logit}\left(\pi_{ij}\right) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \lambda_1 + \lambda_2 X_{ij} \tag{11}$$

Let M_j and $\hat{M}_j$ be the observed and predicted number of successes in the jth cluster, respectively. After fitting model (11), $\hat{M}_j$ is calculated as

$$\hat{M}_j = \sum_{l=1}^{m} \hat{\pi}_{ij} = \sum_{i=1}^{n_j} \text{expit}\left(\hat{\lambda}_1 + \hat{\lambda}_2 X_{ij}\right).$$

Then the observed and predicted numbers of success are compared by computing a residual for each cluster. If we want to estimate the adjusted RD, the residual, known as difference-residual, for each cluster is calculated as

$$e_j = (M_j - \hat{M}_j)/n_j,$$

and treated as a continuous outcome in any subsequent analyses.

B Performance criteria

Let the mean of the estimated LATE across the replicate datasets in each scenario, indexed by $l = 1, ..., L$, with $L = 2500$ be $\bar{\hat{\beta}}_{IV} = \frac{1}{L}\sum_{l=1}^{L} \hat{\beta}_{IV_l}$. The following criteria were used to assess the performance of the methods investigated

(a) **Empirical bias:** estimated by $\bar{\hat{\beta}}_{IV} - \beta_{CZ}$.

(b) **Monte Carlo error of empirical bias** $= \sqrt{\sum_{l=1}^{L}\left(\hat{\beta}_{IV_l} - \bar{\hat{\beta}}_{IV}\right)^2/[L(L-1)]}$.

(c) **Coverage rate of the nominal of 95% CIs** $\frac{1}{L}\sum_{l=1}^{L} I\left(|\hat{\beta}_{IV_l} - \beta_{CZ}| < 1.96 s_i\right)$, where s_i denotes the model-based standard error for $\hat{\beta}_{IV_l}$. The Monte Carlo Error of coverage is $\sqrt{\sum_{l=1}^{L}(0.95)(0.05)/L}$.

C Results for adjusted CL summaries CL-TSLS for TEXT4FLUJAB

Table 6: TSLS estimation of practice-level LATE of reminder text messaging to receive flu vaccine on the percentage uptake of flu vaccine in the TXT4FLUJAB trial using adjusted CL outcomes, adjusting for individual-level covariates gender, age and presence of disease.

		Unadjusted		**Adjusted[a]**	
		LATE (95% CI)	**p**	**LATE (95% CI)**	**p**
No weighting	None	0.133 (-0.016,0.282)	0.081	0.133 (-0.017,0.282)	0.082
	HW	(-0.016,0.282)	0.081	(-0.014,0.280)	0.077
	SSDF	(-0.019,0.285)	0.086	(-0.021,0.286)	0.089
	SSDF + HW	(-0.019,0.285)	0.086	(-0.018,0.283)	0.083
Cluster	None	0.068 (-0.063,0.198)	0.310	0.071 (-0.058,0.200)	0.280
size weighting	Huber-White	(-0.081,0.216)	0.372	(-0.069,0.212)	0.320
	SSDF	(-0.065,0.201)	0.316	(-0.061,0.203)	0.288
	SSDF + HW	(-0.084,0.219)	0.378	(-0.073,0.215)	0.328
Minimum-	None	0.128 (-0.017,0.273)	0.084	0.128 (-0.017,0.273)	0.084
variance weighting	HW	(-0.015,0.271)	0.080	(-0.014,0.269)	0.077
	SSDF	(-0.020,0.275)	0.090	(-0.021,0.277)	0.091
	SSDF + HW	(-0.018,0.273)	0.086	(-0.017,0.273)	0.083

[a] TSLS estimation was adjusted for weekend clinics (yes/no).

D Results for small true LATE

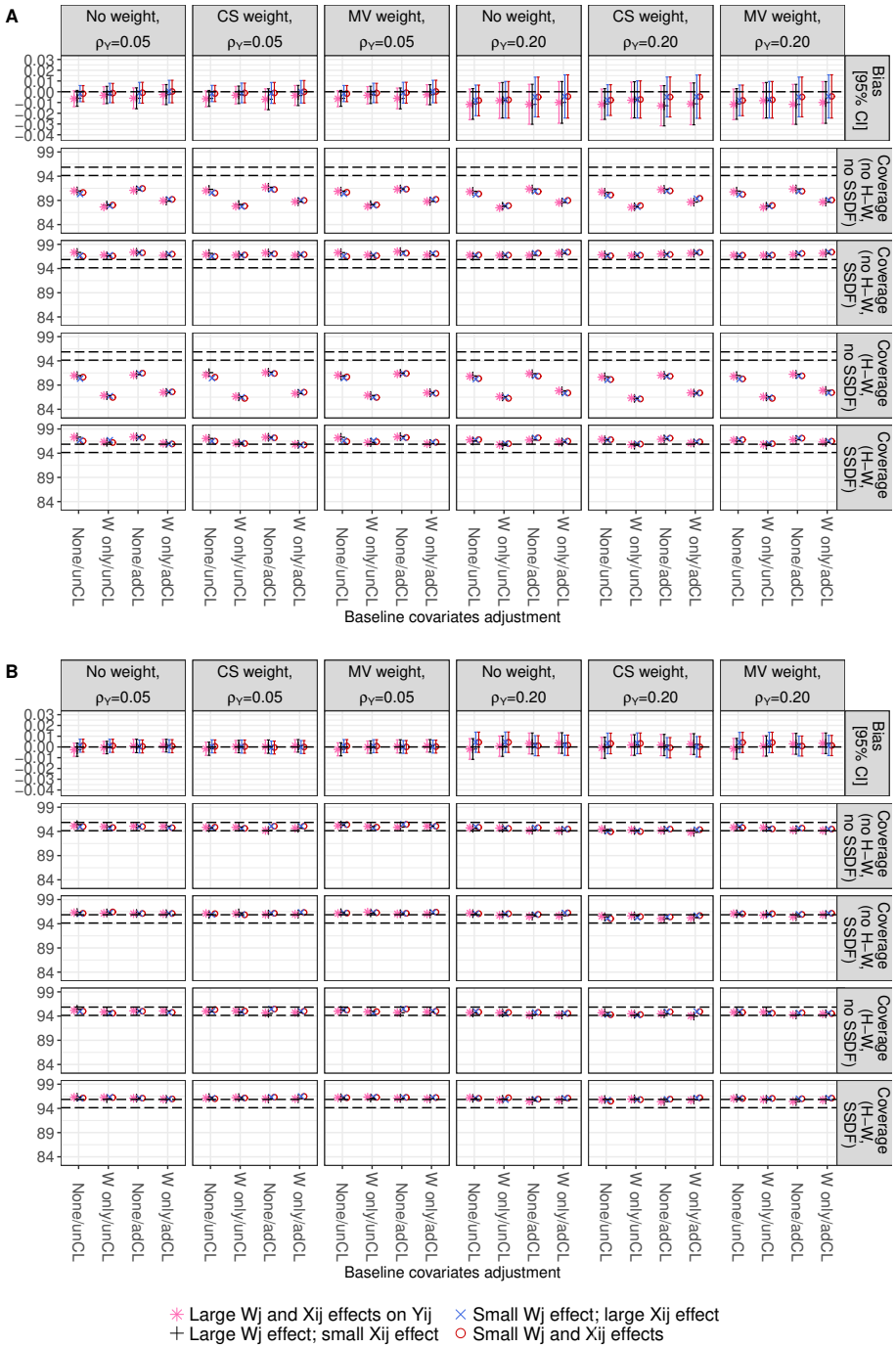

Figure 7: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with cluster-level non-adherence and small true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B .

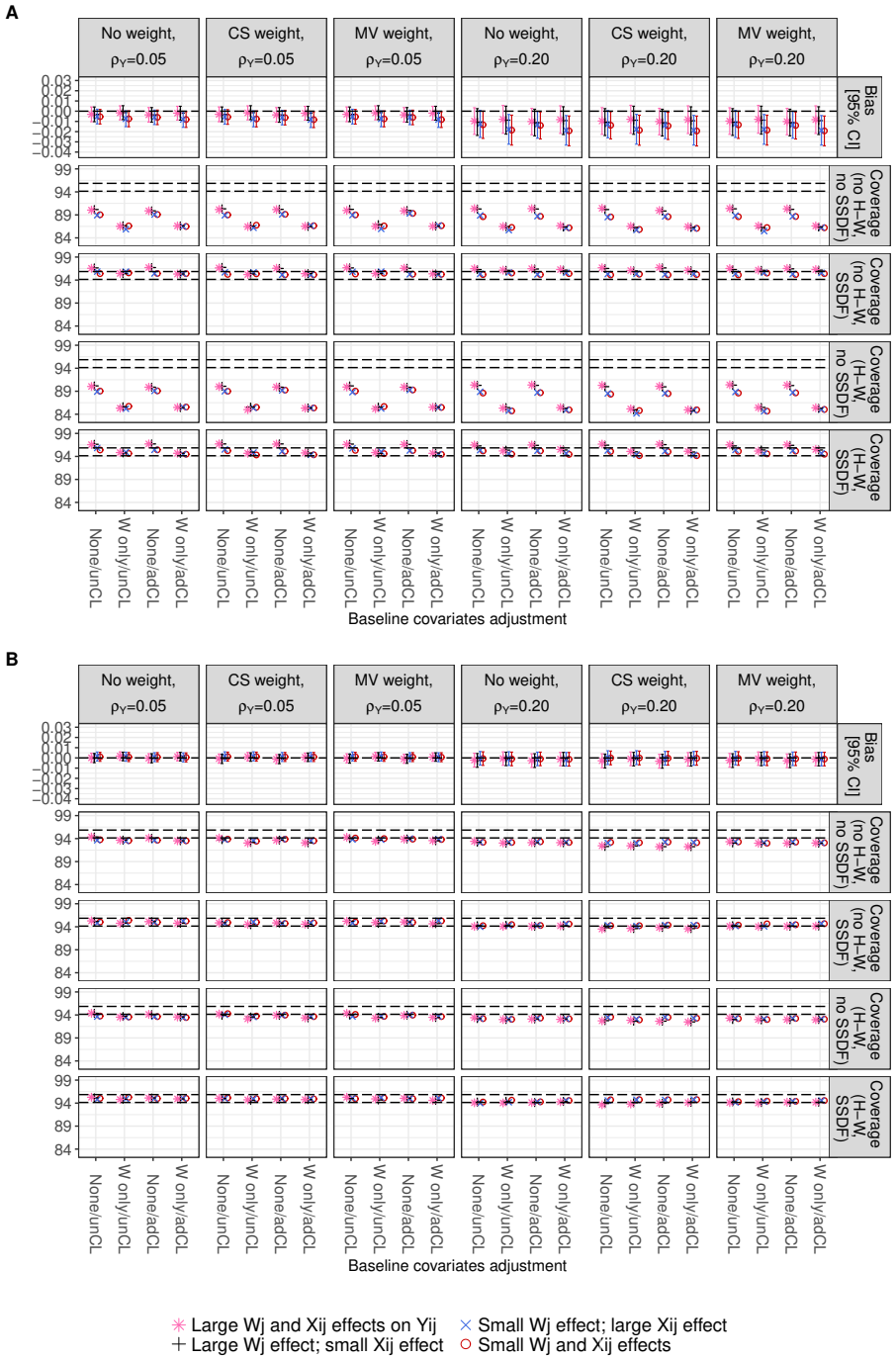

Figure 8: Bias (top row) and 95% CI coverage (rows 2–5) of CL-LATE with individual-level non-adherence and small true LATE. Data generation scenarios represented by $*, +, \times$, and $\circ$. Estimates are obtained via unadjusted or W-adjusted TSLS with different weights (none, cluster size (CS) and minimum-variance (MV)) (by column) using CL unadjusted or adjusted for X outcomes ("unCL" or "adCL"). Small ($J = 10$) and large ($J = 50$) number of clusters results are shown in Panel A and B .