Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Administrative data, such as education, medical, earnings, and criminal justice records, provide an increasingly rich data source for measuring outcomes for randomized controlled trials (RCTs) of interventions, policies, and programs. Administrative data are typically cheaper to collect than survey data, and can offer larger sample sizes with lower attrition and nonresponse. These data, however, can be difficult to obtain due to data privacy concerns protected by law, although there has been some recent progress by Federal agencies, such as the U.S. Census Bureau's Center for Administrative Records Research and Applications, in improving data access for evidence building (U.S. Office of Management and Budget, 2018).

One approach for facilitating access to administrative data for RCTs is to request *aggregate* outcome data for the study sample—such as group averages—rather than individual-level data. This approach is preferable, for example, to sending computer programs to data agencies to conduct the analysis, because it provides researchers with some control over the data and allows for follow-up analyses not anticipated in initial analysis protocols (Card et al, 2010). Furthermore, the availability of grouped data may help reduce obstacles to producing public or restricted use datasets for future research (and for study replication as journals sometimes require), that cannot always be produced using individual-level records due to data destruction clauses or other restrictions in data use agreements. Thus, the use of grouped data is a viable alternative to other approaches to protect data privacy, such as masking individual-level data while preserving the statistical properties of the data (Matthews and Harel, 2011).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

This paper focuses on the following question: What is lost in terms of bias and precision if average treatment effects (ATEs) for RCTs are estimated using only group-level means on study outcomes, covariates, and weights? In addition, we examine the question: What additional aggregate data (sums of squares and cross-products) are required to fully replicate the impact findings using the individual data? The first question is of particular practical interest because requesting only group averages can substantially simplify data requests, thereby potentially improving data access. For example, replicating the individual-level analysis based on a weighted impact regression model with five covariates would require 71 statistics per group per outcome variable, compared to only 9 statistics if the analysis was conducted using group-level means only. Furthermore, the analysis based on group-level means could be easier to conduct using existing software, such as RCT-YES (www.rct-yes.com) that applies to all designs considered in this article.

We consider a full range of RCT designs, including clustered designs (where groups such as hospitals, schools, or communities are randomized) and blocked (stratified) designs, for models with and without baseline covariates and weights. We consider both full sample and baseline subgroup analyses. Our analysis uses recently developed *design-based* ATE estimators for RCTs

(Yang and Tsiatis, 2001; Freedman, 2008; Schochet, 2010, 2013, 2015/2016; Lin, 2013; Miratrix et al, 2013; Imbens and Rubin, 2015, Schochet and Kautz, 2018) that are conducive to using grouped data. Design-based methods use the building blocks of experimental designs with minimal assumptions to yield consistent, asymptotically normal estimators, and apply to both continuous and binary outcomes. These estimators have been shown to perform well in simulations (Schochet, 2015/2016; Schochet and Kautz, 2018).

Our analysis using data on group-level means draws on the large literature over many years on the statistical implications of using aggregate data to make inferences on micro-level relationships. This literature focuses on efficiency losses from using grouped data to estimate well-specified regression models (Prais and Aitchison, 1954; Feige and Watts, 1972; Dhrymes and Lleras-Muney, 2005), and ecological inference biases due to omitted model explanatory variables and non-linear micro-level relationships (Robinson, 1950; Goodman, 1953; King, 1997; Freedman et al., 1998). While some authors have discussed the value of using aggregate data for RCTs (e.g., Rieken and Boruch, 2013; Jacob, 2016), this literature has not formally examined the statistical properties of this approach. This paper helps fill this gap for a wide range of RCT designs. While our focus is on RCTs, our results apply also to quasi-experimental designs with comparison groups. We also apply the theory by reanalyzing data from two RCTs in the education area, one using a non-clustered design and one using a clustered design.

This paper fits directly with the conference theme of responding to diverse demands for evidence by developing statistical methods to help improve access to administrative records data for rigorous research. The goal is to expand the conduct of impact evaluations using administrative data, and to produce faster results at lower cost.

Statistical, Measurement, or Econometric Model:

Description of the proposed new methods or novel applications of existing methods.

The discussion below outlines how using group-level means only can be used to produce design-based impact estimators for non-clustered and clustered designs. We focus on the super-population model where inference generalizes to a broad study universe. We then summarize the empirical analysis.

Non-clustered designs. Non-clustered designs occur when individuals are randomized directly to the treatment or control groups. Under this design, the data generating process for the observed outcome for an individual (y_i) can be expressed as follows:

$$y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \qquad (1)$$

where $Y_i(1)$ is the potential outcome for individual *i* in the treatment condition, $Y_i(0)$ is the potential outcome for the same individual in the control condition, and T_i is the treatment status indicator. Design-based theory is based on a rearrangement of (1) to produce a regression model:

$$y_i = \beta_0 + \beta_1 (T_i - p) + e_i, \qquad (2)$$

where $\beta_1 = (\mu_T - \mu_C)$ is the treatment-control difference in population means, *p* is the treatment group sampling rate, and e_i is a mean zero error defined by the randomization mechanism with population variance σ_T^2 for the treatment group and σ_C^2 for the control group. It can be shown that estimating this model using OLS produces a difference-in-means estimator ($\hat{\beta}_1$) that is unbiased and asymptotically normal as the sample size N approaches infinity with variance:

$$Var(\hat{\beta}_1) = \frac{\sigma_T^2}{N_T} + \frac{\sigma_C^2}{N_C}, \qquad (3)$$

where N_T and N_C are sample sizes. The variances, σ_T^2 and σ_C^2, can be estimated using sample variances, S_T^2 and S_C^2. Hypothesis testing can be conducted using t-tests with $(N_T + N_C - 2)$ degrees of freedom.

To generate grouped data, we assume that administrative data agency staff randomly sort the data into $G_T \geq 2$ groups of size $Z_{Tg} = Z$ for treatments and $G_C \geq 2$ groups of size $Z_{Cg} = Z$ for controls, and then release $\bar{y}_g, T_g,$ and Z_g to the research team. To develop estimators in this case, note first that because of random sorting, the *same* model and error structure as in (2) holds at the group level and has the same statistical properties. Specifically, the group-level model is obtained by stacking the group-level averages:

$$\bar{y}_g = \beta_0 + \beta_1 (T_g - p) + \bar{e}_g, \qquad (4)$$

where $\bar{e}_g$ is now a mean zero error term with variance σ_T^2 / Z for the treatment group and σ_C^2 / Z for the control group. The OLS estimator using (4) is identical to the OLS estimator based on the individual data, $\hat{\beta}_1$, and is asymptotically normal with the following variance:

$$Var(\hat{\beta}_{1G}) = \frac{\sigma_T^2}{G_T Z} + \frac{\sigma_C^2}{G_C Z}. \qquad (5)$$

Note that (5) is identical to the variance in (3) because $N_T = G_T Z$ and $N_C = G_C Z$. Unbiased estimates for σ_{TI}^2 / Z and σ_{CI}^2 / Z can be obtained using sample variances for the group averages, S_{TG}^2 and S_{CG}^2.

We find then that impact estimators and variances are the *same* using the individual and grouped level data. As shown and quantified in the paper, the statistical cost of using the grouped data relative to the individual data is fewer degrees of freedom (*df*) for the t-tests: $(G_T + G_C - 2)$ compared to $(N_T + N_C - 2)$. This leads to reduced power to detect treatment effects if the number of groups is small. These effects can be mitigated by selecting more groups and fewer individuals per group to the extent possible. Another approach is to request multiple groupings (which does not compromise data privacy).

The paper extends this analysis to models with covariates, where grouped data leads not only to *df* losses but also to standard error inflation due to increased correlations between the covariates and the treatment status indicator. The paper also considers models with weights (to allow for groups of different sizes), blocked designs, finite population models, and subgroup analyses.

Clustered designs. The above theory can be extended to designs where m clusters (such as schools) are randomized instead of individuals. In this case, the data generating process for the observed mean outcome for school j, $\bar{y}_j = (\sum_{i=1}^{n_j} y_{ij} / n_j)$, can be expressed as follows:

$$y_{ij} = T_j Y_{ij}(1) + (1 - T_j) Y_{ij}(0). \tag{6}$$

Rearranging this relation yields the following regression model generated by the experiment:

$$y_{ij} = \beta_{0,Clus} + \beta_{1,Clus}(T_j - p) + (u_{j,Clus} + e_{ij,Clus}), \tag{7}$$

where $\beta_{1,Clus} = E(w_j[\mu_{Tj} - \mu_{Cj}]) / E(w_j)$, w_j are cluster-level weights, and μ_{Tj} and μ_{Cj} are cluster-level population means. This model is the usual random effects specification with mean zero between- and within-cluster error components that are uncorrelated with $(T_j - p)$, but where the error variances differ for the treatment and control groups.

Consider WLS estimation of (23) using the individual-level data and the weights, w_{ij}, where the explanatory variables include a $1xv$ vector of covariates, $\mathbf{x_{ij}}$, and where the model includes T_j rather than $(T_j - p)$ (both yield the same results). The covariates can be at the individual or cluster level. Schochet (2015/2016) and Schochet and Kautz (2018) show that as M increases to infinity, the multiple regression, $\hat{\beta}_{1,Clus}$, is asymptotically normal with asymptotic mean, $\beta_{1,Clus}$. A consistent variance estimator for the true asymptotic variance is

$$V\hat{a}r(\hat{\beta}_{1,MR,Clus}) = \frac{1}{(1 - R^2_{TX,Clus})}\left[\frac{MSE_{T,Clus}}{M_T} + \frac{MSE_{C,Clus}}{M_C}\right], \tag{8}$$

where

$$MSE_{T,Clus} = \frac{1}{(M_T - v\hat{p} - 1)\bar{w}_T^2} \sum_{j:T_j=1}^{m_T} w_j^2 (\bar{y}_{jW} - \hat{\beta}_{0,Clus} - \hat{\beta}_{1,Clus} - \bar{\mathbf{x}}_{\mathbf{jW}} \hat{\boldsymbol{\gamma}}_{\mathbf{Clus}})^2 ,$$

$$MSE_{C,Clus} = \frac{1}{(M_C - v(1 - \hat{p}) - 1)\bar{w}_C^2} \sum_{j:T_j=0}^{m_C} w_j^2 (\bar{y}_{jW} - \hat{\beta}_{0,Clus} - \bar{\mathbf{x}}_{\mathbf{jW}} \hat{\boldsymbol{\gamma}}_{\mathbf{Clus}})^2 ,$$

$\bar{w}_T = \frac{1}{M_T}\sum_{j:T_j=1}^{m_T} w_j$ and $\bar{w}_T = \frac{1}{M_C}\sum_{j:T_j=0}^{m_C} w_j$ are mean cluster-level weights, $R^2_{TX,Clus}$ is from a regression of T_j on the covariates and an intercept, and $\hat{p} = \frac{1}{\sum_{j=1}^{M} w_j}\sum_{j=1}^{M} T_j w_j$ is the weighted proportion of clusters in the treatment group. Hypothesis testing can be conducted using t-tests with $(M - v - 2)$ degrees of freedom, which is based on the *number of clusters*, not the number of individuals.

The variance estimator in (9) is based on regression residuals averaged to the cluster level. Intuitively, the model is estimated using the individual data, but standard errors are calculated using residual sums of squares based on cluster-level residuals. Note the covariates will only affect the ATE estimates and increase precision if mean covariate values vary across clusters, but $\hat{\boldsymbol{\gamma}}_{\mathbf{Clus}}$ captures both the outcome-covariate relationships between and within clusters.

These results suggest that for clustered designs, a natural grouping scheme is to *request administrative data by cluster*—for example, school- or hospital-level averages—to minimize information loss. With this scheme, (7) still holds at the group level, yielding consistent estimators. For models without covariates, analyzing data averaged to the cluster level yields the *same* impact and variance estimators with the same *df* as using the individual data, so no information is lost. The same result also holds if the model includes only cluster-level covariates.

If the model includes individual-level covariates (that vary within and between clusters), the impact and variance estimators will differ using the grouped and individual data (but both are consistent). In this case, cluster-level grouping will yield slightly larger standard errors (quantified in the paper) due to larger expected correlations between the covariates and treatment status indicator. These design effects can be reduced if additional groups are formed by sorting individuals into groups within each cluster. The paper quantifies these gains.

Empirical analysis. To examine how the theory using the grouped data applies in practice, the paper analyzes data from two RCTs in the education field. The first non-clustered RCT, the New York City School Voucher Experiment (Mayer et al. 2002), examined the effects of offering scholarships to private schools worth up to $1,400 a year for three years to children from low-income families. Eligible students who applied for scholarships were randomly selected for the treatment group using a lottery system. The second clustered RCT, the Teach for America Evaluation (Decker et al. 2004), examined the impacts of the Teach for America (TFA) Program that recruits seniors and recent graduates with strong academic records from selective colleges to teach for a minimum of two years in low-income schools. Students were randomly assigned to classrooms (clusters) taught by TFA teachers or traditional teachers in the same grade and school. The above theory can be extended to designs where m clusters (such as schools) are randomized.

The empirical analysis forms groups of various sizes and compares the impact findings using the individual-level data to those using the grouped data for models with and without covariates. The findings align closely with the theory.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

We believe that the rigorous design-based methods discussed in the paper will improve access to administrative records data for impact evaluations to help avoid data privacy issues that often plague data access. Our hope is that education researchers will consider using these methods in the future, and conduct research to improve them.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Our main finding is that using grouped data only to analyze RCT data using design-based methods can be an effective strategy to gain access to administrative records data. No biases result if group-level averages are used for the analysis rather than individual data. The key reason is that the individual-level model defined by randomization also holds at the grouped level for both non-clustered and clustered designs. For non-clustered designs, statistical power losses from using grouped data are less than five percent if the number of groups is at least 30 and the model contains only a few key covariates. If needed, obtaining multiple groupings could be a good strategy to minimize design effect losses. For clustered designs, little information is lost if administrative data are collected as cluster-level (school-level) averages. In this case, there are no degrees of freedom losses and tolerable standard error increases due to correlations between the covariates and the treatment status indicator. The empirical analysis supports these conclusions.