Using Aggregate Administrative Data in Social Policy Research

Robin Jacob, University of Michigan

DECEMBER 2016

OPRE REPORT #2016-91

INTRODUCTION

Many researchers are reluctant to use aggregate data in program evaluation and other policy relevant research. Aggregate data refers to data on individuals that have been averaged by year, by geographic area, by service agency, or in some other way. For example, data on the individual earnings of job training participants may be aggregated by year or student test scores may be aggregated by school. This brief explores when aggregate data, instead of individual-level data, can be used to effectively address social policy research questions.

The reluctance to use aggregate data has several sources. A seminal article by William Robinson (1950) demonstrated that the correlation between two variables measured at the individual level and the correlation between those same variables measured at the aggregate level are not the same. Specifically, Robinson showed that the relationship between an individual's status as an immigrant and the literacy skills of that individual is not the same as the relationship between the proportion of immigrants in a state and the average literacy levels of that state. This is often referred to as Robinson's "ecological fallacy" and is frequently cited as a reason why it is inappropriate to use aggregate data when analyzing the relationship between two variables. What is often overlooked, however, is that if one is, in fact, interested in the association between two aggregate variables (e.g., percentage of female students in a school and the average mathematics achievement of a school), or even in the relationship between an aggregate variable and an individual-level characteristic, then using the aggregate data is not only appropriate, but often necessary.

> If one is interested in the association between two aggregate variables, or even in the relationship between an aggregate variable and an individual-level characteristic, then using aggregate data is not only appropriate, but often necessary.

Dr. Robin Jacob is an associate research professor at the Institute for Social Research and the School of Education at the University of Michigan. She researches and evaluates education reform initiatives, education outcomes, and student achievement.

At the same time, the more recent focus on multilevel modeling, which emphasizes the importance of taking clustering into account when analyzing nested data—for example, data in which students are nested within classrooms that are nested within schools—to ensure that standard errors are estimated properly, has led many to believe, incorrectly, that *any* analysis of clustered data that does not use multilevel modeling will distort standard errors and lead to incorrect inferences. This is not always the case. While it is inappropriate to estimate models using individual data that do not account for the ways in which individuals are nested within groups, from a statistical perspective, it is appropriate to analyze data at the group level without accounting for the individuals that make up those groups (see for example, Van den Noortgate, Opdenakker & Onghena, 2005; and Moerbeek, 2004 for research that demonstrates this point).

While individual-level data are often preferable and provide researchers with maximum flexibility in their analyses, in many instances it is both difficult and costly to obtain individual data. With increasing concerns about individual privacy, especially in the health and education sectors, these data are getting more difficult to obtain and requests for such data place a substantial burden on the agencies that manage those requests. In this brief I argue that when individual-level data are not available or are too difficult or costly to obtain, aggregate administrative data can address many policy-relevant research questions. For the purposes of exposition, I focus on the use of aggregate data in the context of program evaluation, but the findings described here are applicable to other contexts as well.

THEORETICAL BASIS FOR RELYING ON AGGREGATE DATA

Previous research has demonstrated that in program evaluation, aggregate data will yield exactly the same results as individual-level data when two conditions hold. First, the data must be perfectly balanced. By perfectly balanced we mean that a) there is an even split between the program and control groups in an evaluation design, and b) there are exactly the same number of individuals in each group by which the data are being aggregated (e.g., the same number of doctors per hospital, clients per job training facility, students per school). Second, the aggregate data will yield exactly the same results as the individual data if no covariates are included in models used to estimate the relationships between variables (see, for example, Raudenbush & Bryk, 2002). Unfortunately, the conditions behind these proofs are unlikely to hold in practice. Schools, hospitals, and other service agencies never have exactly the same number of individuals per group, and often the number of individuals per group varies widely. In program evaluation, the numbers of program and control group participants are often different, and researchers usually also wish to include covariates in their models. However, as described in more detail below, in most cases results obtained using aggregate data will still be nearly identical to those obtained using individual data even when these strict assumptions do not hold.

AN EXAMPLE FROM EDUCATION RESEARCH

Several years ago my colleagues and I were evaluating a professional development program for school leaders (Jacob et al., 2015). We wondered if the professional development program affected the subsequent achievement of the students in those schools, and had arranged to obtain individual-level student achievement data from the state. Unfortunately, due to budget reductions, staff turnover, and concerns about Family Educational Rights and Privacy Act (FERPA), the state was unable to provide us with the data when we were ready for it. However, student achievement data aggregated to the school and grade level were easily available for download from the state department of education's website. The question was whether these data would be sufficient to conduct our analyses of the impact of the program.

To explore whether the school-level data would be sufficient, we used data that we had obtained for an earlier evaluation conducted during a time when the state was able to share individual-level student achievement data (for more information, see Jacob, Goddard, & Kim, 2014). That data set contained fourth-grade reading and math scores for 5,031 students in 78 schools across the state from 2005. We were also able to download school-level data from that same year from the state's website. We simulated a school-level treatment and compared the results from the two sets of analyses (one using the aggregate data downloaded from the state and one using the individual-level data). We found that the estimated treatment effect was 9.805 points using the individual data and 9.821 using the aggregate data. While these estimates were not identical, the differences did not change the statistical significance of the results nor their substantive interpretation. This was true although the data were not balanced. In fact, the number of students per school ranged from 17 to 224 in our data set. Similarly, even with the addition of covariates, the differences between aggregate and individual data were quite small, and did not change the statistical or substantive interpretation of results.

In most cases, results obtained using aggregate data will be nearly identical to those obtained using individual data.

We then conducted a series of simulations to assess under what conditions the aggregate data would not yield comparable results. We found that only when the data are highly imbalanced (e.g., some groups have as few as 20 individuals and others have 180 individuals or more), are substantive differences between analyses that use aggregate as opposed to individual data ever observed. Even when the spread is large, the substantive interpretation of the estimate is affected only 5 percent of the time and these differences are relatively small (around 0.03 standard deviations in the most extreme cases). For example, as already described, the number of students per school ranged from 17 to 224 in the data set we obtained from the state, yet, the estimated impacts varied only slightly despite the high degree of imbalance.

The simulations also identified few substantive or statistical differences whether covariates were included in our models. Furthermore, where the data were highly imbalanced, adding aggregate-level covariates to our models actually helped bring the results more in line with those from the individual-level models.

WHEN ARE AGGREGATE DATA ADEQUATE?

These findings are encouraging and suggest that aggregate data have wider applicability than typically thought. Several factors should be considered when assessing whether aggregate data are appropriate to use for a particular set of statistical analyses.

What are the research questions? As noted above, if you are interested in understanding the relationship between two individual-level variables, such as gender and college attendance, aggregate data are not appropriate. If, on the other hand, you are interested in understanding the relationship between a group-level characteristic (e.g., whether a school participated in an intervention or a group of individuals participated in a job training program) and an individual-level outcome such as student achievement or individual earnings, then aggregate data are generally suitable. Similarly, aggregate data are always appropriate for understanding the relationship between two group-level variables (e.g., the proportion of individuals arrested within a city and the crime rate across cities).

How are the data constructed? The second issue to consider is whether the aggregate data and the individual data are based on the same underlying data set; specifically, whether some individuals or groups are excluded from the aggregate data. For example, in education, for confidentiality reasons, states usually establish minimum reporting requirements that mandate aggregate data be withheld if the total number of students included in the aggregate falls below a certain threshold. State reporting requirements often restrict the reporting of data for groups of fewer than 10, thereby limiting the sharing of some aggregate test score data, including data about subgroups of students for which there are small numbers of students per school and about rural schools with small numbers of students per grade.

In the example described above, the first step in assessing whether the aggregate data would meet our needs was to establish that the aggregate data were based on exactly the same data that would have been included in an individual-level file. We were able to establish that all of the 78 schools in our student-level file were also included in the school-level data, and that the values for the publicly available school-level variables exactly matched the values obtained from averaging the corresponding variables in the restricted-use student-level file. However, had we selected a different sample of schools, this might not have been the case. In the state, there were 38 elementary schools with at least one but fewer than 10 third-grade students. Had one of these schools been part of our data set, it would not have been represented in the aggregate data. In some states, the minimum reporting requirements are quite high (in at least one state, aggregate data are not released unless there are a minimum of 40 students included in the calculation). This makes it more likely that the aggregate data might not include all schools of interest. In other sectors, other factors may impact whether the aggregate data are based on exactly the same data as the individual data.

Several factors should be considered when assessing whether aggregate data are appropriate to use for a particular set of statistical analyses.

What types of outcome measures are available? Researchers must also consider whether the outcome measures that are available in aggregate form are sufficient to answer the questions of interest. In education, for instance, some states only make cut-scores (e.g., the percentage of students meeting a certain proficiency threshold) available, but do not post aggregate raw or scale score results. Metrics like the percent proficient limit the sensitivity of the data to analyses (Ho, 2008). If an intervention impacts student in the lowest quartile of achievement in a state, but the only scores available indicate the number of students reaching a proficient level, the impact of the program might not be apparent. The state that was the subject of our study made both cut-scores and scaled scores available, so this was not a problem for our analyses. However, as of 2013 only 25 of 50 states reported average scale scores as part of their publicly available data. In other sectors, there may be additional factors to consider regarding the outcome measures that are available.

WHAT DO INDIVIDUAL-LEVEL COVARIATES ADD? ARE THEY NECESSARY?

One potential reason for preferring individual rather than aggregate data is that the researcher can use individual-level covariates in models that estimate program impacts. Under some circumstances this can increase the precision of the models. However, previous research has shown that aggregate-level covariates can be equally (or in some cases more) effective in improving precision in comparison to individual-level covariates (Bloom, Richburg-Hayes & Black, 2007).

In our example, adding individual-level covariates to a model that already included aggregate-level covariates did nothing to improve precision. In fact, it slightly increased the standard error of the estimated impact. Still, our data also contained some covariates that were available at the individual level that were not publically available at the school level. For example, there was information indicating each student's gender, and whether they were classified as Limited English Proficient (LEP) and/or received special education services in the individual-level file; these variables were not available in the aggregate file. When we used the individual-level data to create aggregate covariates (e.g., % LEP, % Special Education, % female) and added these variables to our models, the minimum detectable effect size was reduced from 0.20 standard deviations to 0.18 standard deviations. Thus, one benefit of obtaining individual-level data is that it may provide more variables to use in model estimation. At the same time, the increase in precision obtained in our example was relatively small and a small increase might not be worth the effort and cost to obtain the individual data. However, the inclusion of a larger set of aggregate variables in public-use files would make these data sets even more useful.

> Adding individual-level covariates to a model that already included aggregate-level covariates did nothing to improve precision.

LIMITATIONS TO USING AGGREGATE DATA

Although aggregate data have wider applicability than typically thought, there are some analyses that are difficult to conduct using aggregate data. For instance, in addition to understanding overall program impacts, researchers are also often interested in understanding how the impact of a treatment varies across different types of individuals, such as estimating impacts on only the males in the sample or only on those starting the intervention with the lowest skills or experience level. This is typically accomplished by conducting subgroup analyses. Subgroup analyses can be somewhat more difficult to undertake with aggregate data. In our example, the aggregate data could not be used to test whether the intervention was more effective for students who were low achieving at baseline compared to those who were higher achieving at baseline. For this, we would have needed information on individual student achievement.

However, in education, the No Child Left Behind (NCLB) act requires reporting of some results disaggregated by subgroup, and these disaggregated data can be employed to answer some questions regarding how impacts vary across individuals of different backgrounds. NCLB requires states to report results separately for (a) students who are ethnic minorities, (b) students who speak English as a second language, (c) students who are economically disadvantaged, and (d) students who are emotionally, physically, or mentally disabled to the extent that they need Individualized Education Plans (IEPs). Using these disaggregated results, one can conduct analyses to estimate subgroup differences. Instead of using individual-level data to select subgroups, publicly reported average school and grade-level subgroup scores can be used as outcome measures. Our analyses indicate that as long as the number of students per subgroup per school is greater than five in most schools, conducting analyses in this way will yield results that are quite comparable to those that would have been obtained using individual student-level data.

In addition, longitudinal and growth modeling are not possible with aggregate data. In our study, for instance, we could not follow individual children over time as they progressed through school to understand the longer-term impact of the program on individual children. Although we were able to use aggregate data on successive cohorts of children to explore the global impact of the program over time, because children move and change schools, these cohorts contained some children who had not been exposed to the intervention.

CONCLUSIONS

Aggregate data potentially have wide applicability, and their use can substantially decrease the burden on state and federal agencies because they are often publically available and do not require additional safeguards regarding individual confidentiality. However, the quality of aggregate administrative data can greatly impact how useful those data are and as a result how likely researchers are to use them instead of requesting information at the individual level. Aggregate data are most useful when the data are based on the same records as would be included in an individual-level file and to the extent that they minimize the individuals or units that are excluded. Policies could be put in place to help reduce the data that get excluded from aggregate data sets. In education, for example, reducing the minimum reporting requirements to 5 or 10 instead of 30 or 40 would help ensure that aggregate data are comparable to individual data, thereby increasing their usability. Work done by various federal agencies that report aggregate data and are also required to ensure the confidentiality of their respondents suggest that minimum reporting requirements of 5 or 10 are sufficient to protect respondent privacy (Klein, Proctor, Boudreault, & Turczyn, 2002; Lauger, Wisniewski, & McKenna, 2015). There are also other mechanisms that can be used to help ensure that privacy is protected without suppressing data (Lauger et al., 2015; Yang et. al., 2011).

Disaggregating data by key subgroups (for example, reporting aggregate results separately by gender or racial and ethnic subgroups) would also allow researchers to answer various questions of interest without the need to access individual-level data. In education, this type of disaggregation is already required for several individual characteristics. Other sectors might consider similar types of disaggregated reporting.

In addition, to the extent that a wide range of outcome variables are made available in aggregate form, the data will be more useful. In education, making raw test scores or scaled scores available, in addition to the percentage of students meeting proficiency benchmarks, would make the aggregate data substantially more beneficial to researchers and others wishing to use the data.

Finally, developing data portals that are easy to find, well documented, and easy to access would help facilitate the use of these data. While developing such infrastructure will require an upfront investment of time and resources from federal and state agencies and others that provide access to data, such investments can reduce the cost and administrative burden of managing data requests in the future.

Researchers should carefully consider whether aggregate data can be used before requesting individual-level data.

At the same time, researchers should carefully consider whether aggregate data can be used before requesting individual-level data. The following questions can help guide that decision:

- ❖ What is the research question to be answered? Is the question about the relationship between two group-level characteristics or between group-level characteristics (e.g., whether or not a school participated in an intervention or a group of individuals participated in a job training program) and an individual-level outcome (e.g., student achievement, earnings)?
- ❖ Is there aggregate data available on the outcome of interest?
- ❖ What is the quality of the aggregate data?
    - ○ Is it based on the same data as the individual-level data?
    - ○ Does it have a rich set of covariates?
    - ○ Are appropriate outcome metrics available?
- ❖ Do the research questions strictly require following individuals longitudinally or disaggregating the data by individual characteristics?
    - ○ This question is worth careful consideration. Even if individual data can be obtained, would there be sufficient statistical power to answer questions about variation across subgroups? How much new or useful information will following individuals longitudinally provide? Are there other ways to exploit the aggregate data to provide information about variation across individuals or over time?

Although individual-level data are extremely flexible and maximize the types of analyses that can be conducted, it may not be worth the cost (in both time and money) to justify their use. See Jacob, R., et al. (2014) for more detailed information on the analyses referenced here. Please also see the presentation, “Using Aggregate State Assessment Data to Assess the Impact of School-Based Interventions,” which is available from http://www.opremethodsmeeting.org/2015presentations.html.

REFERENCES

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351–360.

Jacob, R., Goddard, R., & Kim, E. S. (2014). Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public use data. *Educational Evaluation and Policy Analysis, 36*(1), 44–66.

Jacob, R., Goddard, R. D., Kim, M., Jung, E., Goddard, Y. L. & Miller, R. (2015). Exploring the causal impact of the McREL Balanced Leadership Program on leadership, principal efficacy, instructional climate, educator turnover, and student achievement. *Educational Evaluation and Policy Analysis, 37*(3), 314–332.

Klein, R. J., Proctor, S. E., Boudreault, M. A., & Turczyn, K. M. (2002, June). *Healthy People 2010 criteria for data suppression* (Statistical Notes No. 24). Hyattsville, MD: National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/statnt/statnt24.pdf

Lauger, A., Wisniewski, B. & McKenna, L. (2015) Disclosure avoidance techniques at the U.S. Census Bureau: Current practices and research. *JSM Proceedings, Government Statistics Section*, 3630–3642.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129–149.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357.

Van den Noortgate, W., Opdenakker, M.-C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*(3), 281–303.

Yang, Y. M., Mushtaq, A., Pramanik, S. Scheuren, F., Hiles, D., Buso, M., & Butani, S. (2011). An evaluation of BLS noise research for the Quarterly Census of Employment and Wages. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, 4216–4229). Retrieved from http://www.bls.gov/osmr/pdf/st110120.pdf

This brief was prepared for the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. It was developed under Contract Number HHSP23320095651WC. RTI International assisted with preparing and formatting the brief. The ACF project officers are Anna Solmeyer and Nicole Constance. The RTI project director is Anupa Bir.

This brief is in the public domain. Permission to reproduce is not necessary. Suggested citation: Jacob, Robin (2016). *Using aggregate administrative data in social policy research,* OPRE Report #2016-91, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This brief and other reports sponsored by the Office of Planning, Research and Evaluation are available at www.acf.hhs.gov/opre.

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.