Research Brief OPRE 2017-18
February 2017

Considerations in Preparing to Analyze Administrative Data to Address Child Care and Early Education Research Questions

CCEEPRA
Child Care and Early Education Policy and Research Analysis

Considerations in Preparing to Analyze Administrative Data to Address Child Care and Early Education Research Questions

OPRE Research Brief #2017-18

February 2017

Submitted by: Van-Kim Lin, MSPH, Kelly Maxwell, PhD, and Nicole Forry, PhD, Child Trends

Submitted to: Ivelisse Martinez-Beck, PhD, Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Kathleen Dwyer, PhD, Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract Number: HHSP23320095631WC
Project Director: Kelly Maxwell
Child Trends
7315 Wisconsin Avenue
Suite 1200W
Bethesda, MD 20814

Suggested citation: Lin, V., Maxwell, K., & Forry, N. (2017). Considerations in Preparing to Analyze Administrative Data to Address Child Care and Early Education Research Questions. OPRE Research Brief # 2017-18. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Acknowledgments: The authors gratefully acknowledge the guidance from Kathleen Dwyer, senior social science analyst at the Office of Planning, Research and Evaluation. We appreciate the thoughtful review and feedback from two experts, Elizabeth Davis and Julia Henly. We extend a special thank you to Leigh Bolick for her review and commentary about the issues from a state agency perspective. Finally, we appreciate the support of the Child Trends Communication team for their assistance in the preparation of this product.

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

Considerations in Preparing to Analyze Administrative Data to Address Child Care and Early Education Research Questions

The purpose of this resource is to help researchers prepare for issues that may arise when using administrative data as the primary data source for a research project. This is the third in a series of resources related to the analysis of administrative data. The first resource, *Developing Collaborative Partnerships with State Agencies to Strengthen Research Using Early Care and Education Administrative Data*, provides considerations for building a strong partnership between researchers who want to analyze administrative data and the state partners who oversee the administrative data. The second resource, *Determining the Feasibility of Using State Early Care and Education Administrative Data*, is designed to help researchers and their state partners determine whether analyzing administrative data is feasible and appropriate for addressing their child care and early education research questions. Once researchers and state agency partners have determined that it is feasible to use administrative data to address a question of shared interest, then this third resource can be helpful in preparing to analyze the data. These resources have been designed for use by researchers who are new to the analysis of administrative data as well as seasoned researchers who are expanding their research to include new types of administrative data or expanding into new states or new agencies. The information generated for each of these resources was developed through conversations with grantees and researchers who have experience analyzing state administrative data.

Administrative data is defined as information about individual children, families, and/or providers of early care and education and other family benefits that are collected and maintained as part of the operation of government programs.

This resource is organized into six sections applicable to analyses of administrative data: 1) understanding the scope and limitations of administrative data when developing an analysis plan, 2) selecting variables to analyze, 3) assessing the feasibility of the plan, 4) preparing a data request, 5) creating a dataset, and 6) developing and

maintaining adequate data documentation. For each of the sections, we have provided considerations, examples, and/or questions to ask that are specific to the use of state administrative data related to child care and early education. The purpose of each section is to provide insights to help researchers in identifying variables and problem solving issues that may arise in the analysis of administrative datasets. Although the sections are described separately, we expect the process to be iterative rather than linear, and to require continued discussions and reconsideration of decisions as new information is learned.

This resource was developed as part of the Child Care Administrative Data Analysis Center (CCADAC) through the Child Care and Early Education Policy and Research Analysis contract at Child Trends. The work is funded by the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. CCADAC works to strengthen the ability of state/territory child care administrators and their research partners to utilize administrative data to address policy-relevant early care and education research questions.

The Perspective of a State Child Care Administrator

Leigh Bolick
South Carolina State Child Care Administrator
Division of Early Care and Education
South Carolina Department of Social Services

Refining the analysis plan and preparing to analyze the data are time-consuming processes. Although researchers may know the critical importance of State Child Care Administrators, the administrators themselves may not think that they can allocate the time or might not understand how important they are to the process. Researchers should plan to thoroughly review the entire methodology with program staff early in the process and ask managers to assign to the project staff who collect and analyze the data as part of their ongoing job responsibilities. State Administrators are generally willing to do this if they know they will be updated on a regular basis.

As referenced in the article, researchers should consider co-constructing the questions to be addressed so that they are thoughtfully developed and address issues of interest to program staff. When research provides data that benefit agency staff, managers are much more willing to put in the time to ensure a sound product. Sometimes it is helpful to determine questions posed repeatedly by policymakers—questions that agency staff are often at a loss to answer. Having a respected researcher available to help provide some answers, or even to discuss the quality of the data, lends credibility to the work of program staff.

Researchers should also consider non-traditional methods of familiarizing themselves with the data and data collection methods employed by agency staff. For example, it can be enlightening to sit with direct service staff to review the data system from their point of view; similarly, sitting in on interviews between parents and subsidy eligibility staff as they ask questions and input the information is a great way to obtain a true picture of the data collection as it relates to service provision.

Finally, helping program staff design tangible and usable documents is a great strategy to promote the benefits of dedicating time to work on research design and data analysis. Providing formatting for data dictionaries that staff can complete and update on an ongoing basis, disseminating samples of systems manuals, starting a timeline of policy changes—all of these are important to program administration yet may not exist or may not be maintained and kept current. Relationships between researchers and program staff are critically important to the process; ensuring that both groups benefit from the final product will not only contribute to the validity of the research but just might also be the beginning of ongoing work with agency staff and administrative data sets related to child care and early care and education.

Understand the scope and limitations of the administrative dataset when developing an analysis plan

An analysis plan details the research questions to be answered, the research design that can best answer the research questions, the particular variables needed to answer each research question, and the type of statistical analysis used to answer the question. Although sections of the plan may be useful to share with state agency staff who are interested in understanding the details of the project, the analysis plan primarily is useful in guiding the work of the research team.

To develop an analysis plan for research using administrative data, the research team needs to understand as much as possible about the administrative dataset(s) to be used. By thoroughly understanding the administrative data and its limitations, researchers are more likely to develop a feasible analysis plan. Listed below are considerations when developing a plan for analyzing administrative data.

- **Determine the unit of analysis for each administrative dataset.** Administrative data can be collected at various levels (e.g., state, site, provider, family, or child). Administrative data can also be longitudinal, with multiple observations for a particular child, family or provider over time. Subsidy payment records, for example, are made to a specific child care provider for services on a particular date for a specific child. Researchers should thoroughly review and understand at which unit of analysis data are collected, as this has implications for the way analysis datasets are set up, the types of analyses that can be conducted, or the types of conclusions that can be reached. For example, a research team may be interested in examining family characteristics that are associated with child care choices for those families receiving child care subsidies. Some child care subsidy data are collected at the child level (e.g., child age, gender, provider); other information is at the family level (e.g., income) and provider level (type of provider, quality). To use this information in an analysis, the researcher may need to be able to link each child to his or her family or provider across data sets that are organized at different levels (child vs. family vs. provider). If the data are not available or cannot be linked, then the researcher may need to adjust her question and analytic strategy to better fit the available data.

- **Recognize the populations represented in administrative data.** Researchers should consider who is represented in the administrative data and who is not. Administrative data may be insufficient to answer certain questions if it does not represent the full population of interest. For instance, state Quality and Improvement Rating Systems (QRIS) often include only a portion of centers and family child care homes, not all, and may exclude some programs, like state-funded pre-Kindergarten in public schools. Thus, QRIS data likely cannot be used to describe the quality of early care and education statewide. However, QRIS data may be appropriate for examining changes in quality over time in rated programs receiving a particular type of quality improvement (e.g., coaching). As another example, child care subsidy administrative data cannot be used to compare those who use subsidy with those who do not because no information is available on children who did not receive a subsidy. However, child care subsidy administrative data can be used to compare subgroups of children whose care is subsidized (e.g., Latino vs. white children, children from families with varying levels of income). If the analysis plan includes questions related to describing experiences of certain subpopulations, the research team should also conduct a power analysis to determine whether the dataset has enough members of that subgroup population to support analyses.

- **Acknowledge the limitations of administrative data.** While administrative data provide a wealth of information, researchers should consider the limitations of administrative data when crafting questions and an analysis plan. Administrative data may not have the level of detail needed, for example, to understand the "why" of a behavior or change. To answer a particular question of interest, researchers may need to collect new data to supplement the administrative data.

- **Determine whether the data can be analyzed using the intended research design and statistical analyses.** The research team also needs to ensure that the data are available (and comprise a large enough sample) and in the proper format (including unit of analysis) to analyze with the intended statistical approach. For example, although a research team may be interested in conducting a longitudinal analysis of families' use of child care over a 5-year period, they may learn that this is not possible because historical data are deleted from the system every 6 months. As another example, although a researcher may be interested in factor analyzing administrative data to determine whether particular indicators of a QRIS are related to each other, it may not be possible if there are not enough cases available for the number of indicators of interest.
- **Refine policy-relevant research question(s).** As the research team becomes more knowledgeable about the administrative data to be used in the research, they may need to work with state agency staff to refine the questions to better match the available data. We suggest that researchers work closely with state agency leaders to develop and refine mutually beneficial policy-relevant research questions. If research questions are co-constructed, the agency staff may be more invested in helping the researcher access the needed data, and the research project will more likely influence policy or practice.

Selecting variables to analyze

Variables needed for a research study may be in one dataset or in multiple datasets that need to be merged together. The research team also may have to decide which variable—of several similar variables—is the most appropriate to include in the analysis. The following tips may be helpful in selecting and analyzing appropriate variables from administrative data sets.

- **Compile data documentation prior to selecting variables**. Obtaining or developing a codebook in the variable selection phase of a project can save time and minimize the chance of obtaining erroneous results from analyses. Though compiling data documentation is important for any analysis, the data contained in administrative data codebooks may contain different, or less, information than research codebooks because the data were not collected for research purposes. For example, if researchers consider using an education variable and one of the response options is "some college," they should determine the definition of what counts as "some college" and how it may be distinguished from an associate's degree or certificate program. Conversations with state agency staff and/or data managers may be necessary to learn more about the specific meaning of available variables. It might be useful to ask questions about who originally provided the information, what instructions were given to individuals who collected and recorded the information, and how staff use the data.
- **Select appropriate variables for analysis.** Variables with similar names may capture different information. For example, enrollment information for a child care program may be recorded in different ways (e.g., number of children present at the time of a licensing visit or number of children listed as enrolled in the program). Similarly, child care subsidy data may have multiple, similar variables such as "eligibility start date," "enrollment start date," and "authorization start date." These variables may reflect the first time the child participated in the program or may be updated to reflect the most recent period of participation. In these instances, researchers and state partners can discuss the variables and determine the most appropriate ones to address the question of interest. Researchers may also want to review the existing literature, specifically any reports generated from similar analyses of administrative data (and particularly the administrative datasets they are planning to use), in order to identify which variables have been used previously.
- **Evaluate whether data for variables of interest are high-quality.** Researchers using administrative data should consider both how complete and how clean data are as they consider which variables to

analyze. Below are a few questions researchers can ask in assessing whether data on specific variables are complete and clean enough to be used.

  - **How much missing data are there?** Researchers may need to talk to data managers and data collectors to ascertain whether missing data reflect a non-positive response or actual missing data. For example, if data are collected on race/ethnicity, are there dummy variables for each race/ethnicity category in which "1" indicates that an individual identifies with that race/ethnicity and blank indicates they do not, or does blank indicate that data are missing?

  - **Are there out-of-range values in the data?** If out-of-range values are found in the data, the research team may want to consult with data managers to see what these out-of-range values may represent or whether they need to be considered missing data.

- **Reconcile variables that changed in definition over time.** Administrative data are collected based on policies at a given time. Therefore, data collected over time are subject to change based on changing policies. For the variables included in the analyses, researchers should ask about policy changes that happened during the time period used for analyses. To the extent possible, the research team should know the date of the policy change, the extent of the change, and the possible ways that the change might have affected the data. For example, a state may have changed the sources of income used to determine income eligibility from one year to the next. In addition, over time, the data coding method may have changed. For example, the state might have added categories to a variable, such as "allowed activities," or changed the categories coded for race and ethnicity. When analyzing data from multiple years, researchers and state program staff can determine together how best to interpret findings from analysis using variables that changed during the time of the study or decide whether findings are not comparable across years.

- **Consider current and historical variables.** Programs may keep only the most current piece of information on children and families (e.g., a family's address or income from the most recent eligibility determination). However, researchers may be interested in understanding changes that have occurred over time. Some agencies maintain historical data, whereas other agencies overwrite previous data with current information. Researchers should consider which variables reflect current only and/or historical information, the feasibility of accessing historical data, and when particular variables have been updated (if at all). If the state has a data warehouse, they may be able to store data collected from multiple points in time (rather than just the most recent data) so that researchers can examine questions about historical trends or changes in the data.

Assessing the feasibility of the project

As researchers take preliminary steps to execute their analysis plan, they should consider the following questions unique to the analysis of administrative data to help prepare them for analyses. The resource, *Determining the Feasibility of Using State Early Care and Education Administrative Data*, provides more details about important considerations in determining the feasibility of using administrative data in research.

- **Does the research team have adequate technical resources for completing the work?** Administrative data are often collected for a large number of children, families, and programs, resulting in a very large dataset. Additionally, depending on where the data are stored, the data may be available only in specific formats. The research team should ensure that they have the technical capacity to deal with these administrative datasets. An assessment of technical resources should include a review of the software needed for the analysis. For example, will data be sent in a format that can be analyzed using software the team members already have/are able to acquire and master? Will specific software (e.g., Stat/Transfer) be needed to transfer the data from one format to another? Will researchers' current

software and computers be sufficient for analyzing the large number of administrative records found in administrative datasets (i.e., sufficient memory space)?

- **Does the research team have adequate time to complete the analysis plan?** Because administrative data are collected for program operations, rather than research, these data often require extensive cleaning and preparation before analysis. In developing an analysis plan, the researchers should consider how long it realistically will take to acquire, clean, merge, analyze, interpret, and report the data. In considering this timeline, researchers should anticipate potential delays in obtaining data (due to additional permissions that may be needed, limited availability of agency data managers to export the data, etc.). Additionally, researchers should build time into their project plan for requesting clarifications from agency staff and data managers regarding data issues that arise throughout the project.
- **Does the agency staff have the interest, capacity, and time to work on the project?** While it is important to determine the capacity of the research team to conduct a research project, it is equally important to determine the capacity of the state agency staff. Although the research team may have had preliminary discussions about the project and data request with the agency leader, the research team most likely will work with other program and data management staff on the day-to-day aspects of the research project. It is important to assess the interest and capacity of everyone in the state agency who is critical to accessing, analyzing, and understanding the data.

Preparing a data request

Unless the agency staff have experience providing data to researchers or have a research-ready dataset, researchers need to provide as much information as possible to the staff about the data they want and the format in which the data should be sent. The data request may be part of the data sharing agreement, an appendix to it, or a separate document. Much of the information needed for the data request is in the analysis plan, but the data request explains the specific information needed by the staff to extract the data for the researcher. The following are suggested items to include in a data request:

- **Summary of research questions and analysis plan.** To frame the data request, a short overview of the research project can provide context for the data request. Additionally, the summary can aid the staff member who pulls the data for the research team.
- **Data elements.** The research team should specify the variables to include in the requested dataset. Through collaborative conversations with the state agency team, the research team and the state agency staff can identify the specific variables available and the variables that best answer the research question.
- **Dataset.** A research question may best be answered by examining a whole group (e.g., all child care programs receiving subsidy payments) or a subgroup (e.g., family child care providers receiving subsidy payments). Researchers need to determine what is most appropriate for the question of interest. If a subset is adequate, then they may need to develop criteria to identify the cases and variables that should be included in the dataset, which might be more manageable to analyze than having the whole dataset. However, researchers must consider whether asking for a subset of the data causes undue burden on the state agency. If the state agency can easily provide a subset of the data, then requesting only a portion of the data may be less burdensome for the agency and more manageable for the research team. On the other hand, if requesting a specific subset of data requires substantial time and capacity for the state agency, it may be easier for the state agency staff to give the researchers the whole dataset, which the researcher can then use to extract the cases and variables of interest.

- **Data timeframe.** Researchers may be interested in having longitudinal data or cross-sectional data from only one point in time. For longitudinal data, it is important to understand the data collection frequency for the variables of interest (e.g., weekly, monthly) and discuss how often the staff will extract data for the research team. Researchers can work closely with state partners to determine a timeframe that balances the needs of the research project and the capacity of the state agency. For example, states provide child care subsidy data to the federal government in monthly data files. They may find it easier to provide these monthly data files to a researcher than to provide data on a weekly or annual basis.
- **Data structure.** The structure in which the data are sent to the research team may be dependent on the data management system as well as the capacity of the state agency staff to format the dataset in a way that would be ready for the research team. For example, the research team may want each row of the dataset to represent an individual child, whereas the state agency may store data in such a way that each row is a family. Requesting data in the structure that is easiest for the staff is likely to lead to fewer complications and a shorter time frame for delivery. The researcher may prefer to merge or link data sets rather than have the agency staff do so. The research and state agency team can determine together the most efficient and efficacious strategy.
- **Data format.** The format of the data may also be dependent on the data management system that the state agency uses. Unless the state is using a sophisticated statistical program, data are often extracted as an Excel, comma separated values (csv) file, or a text file, which can be used easily with most statistical programs. However, the research team may want to find out if the state has limited capabilities.

Creating a research-ready analysis dataset

In order to conduct analyses, the research team needs to create a dataset that can be analyzed for research purposes. Oftentimes, researchers underestimate the amount of time needed to create a research-ready dataset. The following are considerations for creating a dataset from administrative data.

- **Structuring the data for research purposes.** To determine the structure of the dataset, the research team needs to decide at what level they want to analyze the data. For instance, are they interested in examining data at the child level, family level, or program level? The data must be merged at the appropriate level. As an example, if the research team is interested in analyzing child-level outcomes, they may want to build a dataset that has a row for each child, with all the information pertaining to that child. In that case, the research team needs to determine a way to identify the same child across datasets (e.g., unique identifier) and may need to aggregate data that are at a smaller unit of analysis (e.g., row for each time the state pays a provider for care the child received) or disaggregate data at a larger unit of analysis (e.g., data that have a row for each family). Researchers may also choose to create multiple datasets at varying levels to correspond with the appropriate research question. Regardless of the level and the number of datasets needed, the research team should spend time at the beginning of the project to delineate the data structure that best answers each research question.
- **Preparing the data.** The research team may need to prepare the data in order to create an analyzable dataset. Sometimes, data are recorded differently and need to be made consistent for research purposes. Data provided at the county level, for instance, may vary from county to county based on local differences in rules or guidelines. The team may also need to identify spelling errors, determine how missing data are recorded (e.g., 999 or left blank), or determine whether numeric variables were altered (e.g., a license number with leading 0s may show up without the 0s in a spreadsheet).
- **Merging the data.** When necessary, the research team may need to merge or link multiple datasets together in order to answer the research questions. For instance, a research team may be interested

in examining how child care programs' ratings in a QRIS change over time after the system has been revised. The research team may receive the QRIS rating levels for a set of programs before the new structure is implemented, and then receive their updated ratings three years later. The research team may need to identify a way to link the two datasets (e.g., using the program's license number) and merge the two datasets into one to look at the change in rating over time. In order to merge accurately, the research team can carefully review the variables they intend to use to match individuals across the multiple datasets. If first and last names are to be used in matching, researchers may need to review the names to identify near matches that might reflect errors (e.g., Timm vs. Tim). It might also be possible to use more sophisticated probabilistic matching procedures to match individuals across multiple datasets. If the research team does not have the expertise to do this, a third party may be able to assist the team with this task.

Documentation

Analyzing administrative data is often an iterative process. Thus, researchers can document the questions/issues that arise and decisions made throughout the research process, not just at the beginning of the project. In addition to being a reference for the research team when decisions may need to be revisited, good documentation can help when researchers are writing reports/papers and responding to questions from stakeholders or reviewers.

Here are a few suggestions for information that should be documented:

- **Data definitions for each variable**, including original and revised coding, questions asked, and clarifications received from agency staff, along with the date of the question/clarification;
- **Data quality issues** that arose and documentation about how those issues were resolved, to the extent possible; and
- **Policy context**, including any documented changes to policy/administrative practices and the dates that these changes were implemented.

Summary

Once the research team has determined, in partnership with state program staff, that it is feasible to use administrative data to address a research question about early care and education, then the team needs to carefully review the administrative data with program staff and prepare to conduct the study. The team needs to understand the scope and limitations of administrative data as they develop or refine an analysis plan. Selecting which variables to analyze requires careful review of the available data and discussions with program staff to ensure that the variables are the most appropriate and to determine how to request these data. Once variables are selected, the team can prepare a data request to submit to the state agency. When the research team receives the data, they may need to process it further before preparing a dataset for analysis. Thorough documentation of the variables (including how they were recoded) and decisions about how to use the data ensure that both researchers and state leaders understand how the administrative data were used to address a particular question of interest. Even after the researchers begin to analyze the data, they may need to revisit some of the decisions in the analysis plan and continue working closely with program staff to ensure that the analysis and interpretation of findings are appropriate.

The CCADAC team has partnered with Research Connections to organize various resources about analyzing administrative data. For more information, please visit http://www.researchconnections.org/content/childcare/understand/administrative-data.html