THE CONSORTIUM
stimulating self-sufficiency & stability scholarship

July 2017

By Emily Wiegand[1], Robert Goerge[2], and Leah Gjertson[3]

Family Self-Sufficiency Data Center Creating a Data Model to Analyze TANF Caseloads

The Family Self-Sufficiency Data Center (FSSDC) is funded by the U.S. Department of Health and Human Services, Office of Planning, Research and Evaluation, to facilitate use of administrative data by researchers and administrators to improve understanding of, and identify methods for increasing, family well-being. The FSSDC has worked with several state Temporary Assistance for Needy Families (TANF) offices to improve the accessibility and usefulness of TANF administrative data. This brief summarizes lessons learned from that work. It also describes a model for using TANF data to understand caseload dynamics and the data holdings necessary to make those analyses possible.

A data model is a sample data structure that can be created from common data elements and facilitates easy data analysis; it can provide a guiding structure for agencies that wish to extract and transform their data to answer common policy questions. As described in *Family Self-Sufficiency Data Center: Needs Assessment Report* (Weigensberg et al., 2014), the most common hurdles agencies face in making data accessible for research and program analysis include unwieldy data systems, insufficient staff capacity to conduct or facilitate analysis, and incomplete or poor-quality data. This brief presents a data model for Temporary Assistance for Needy Families (TANF) program data. The data model is intended to make limited data and capacity as useful as possible by streamlining the process of connecting data with policy questions. The data model can be useful in addressing the gap between information technology (IT) staff or consultants and program or policy staff by translating policy and program management questions into the kind of technical specifications IT staff need to extract and reformat data.

This brief demonstrates the value of the data model and provides detailed information on how it can be implemented. The brief begins with a short discussion of the research questions and key ideas that guided development of the data model. Next, it describes the basic assumptions of the data model—what data elements agencies need to use this structure. The third section defines the data model in detail and describes the transformations necessary to create files in this structure, including specific examples of how data sets from two states were transformed using the data model. The final section describes possible avenues for growth and expansion for the model.

In addition to this theoretical discussion, the FSSDC team has made sample data in the data model format and sample scripts for doing some of the more complex transformations available on GitHub. These materials can be accessed at http://www.github.com/chapinhall/fssdc.

The **FSSDC** is a partnership between the University of Chicago Harris School of Public Policy, Chapin Hall at the University of Chicago, and NORC at the University of Chicago.

FSSDC
Family Self-Sufficiency Data Center

[1] Chapin Hall at the University of Chicago.
[2] Director of the FSSDC, University of Chicago Harris School of Public Policy, Chapin Hall at the University of Chicago.
[3] Chapin Hall at the University of Chicago.

INTRODUCTION TO THE TANF DATA MODEL

The design of the TANF data model was driven by research questions that states seeking to better understand their TANF programs frequently asked.[1] In addition, simplicity, flexibility, and variability were a focus of the design, to make the potential audience for this work as large as possible.

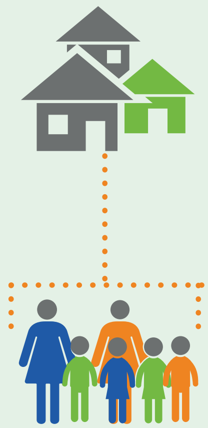
The basic data model is designed for analysis of caseload dynamics, including changes in caseload size and characteristics, and questions of administrative process, duration, and recidivism. Most agencies have some routine reports about the number of active cases and members, but their ability to view these data in detail is limited. The data model is ideal for answering such questions as the following:

- How has the overall size of the caseload and its subgroups (certain types of cases or members) changed over time? Have new entries, or a variable rate of exits, driven this change, or has the caseload been static?
- How many cases, what types of cases, and how many members (by demographic characteristics) are on the TANF caseload in a given geographic region (for example, state or county) at a given time?
- At what rates are individuals or cases that have been leaving the caseload later returning to the caseload? Is there any pattern to this recidivism, perhaps over time or related to why the cases were closed in the first place?

We model our work on the agile approach to software development: start with a simple, small area; build success, knowledge, and buy-in in that area; and then grow and develop based on user needs.

The caseload dynamics data model contains two key units of analysis: (1) case, and (2) member (see Figure 1). Data points tracked at the case level include case types, household size, location, benefits received at the case level, and head of household characteristics (if the administrative data identify a head of household). Data points tracked at the member level include demographics such as birth date, race, education or work status information, marital status, and the person's eligibility for case-level benefits (because often there are people—such as the parent(s) or guardian(s) in child-only TANF cases—who are part of a household but who do not qualify for all case benefits received).

Figure 1. Case and member data



Case-level Data:
Case type, household size, location, benefit receipt, head of household (HOH) characteristics

Member-level Data:
Birth date, race/ethnicity, education and work status, marital status, individual eligibility

Depending on the content and quality of the data collected, data on member relationships also may be available. These data, which relate members to each other, do not store easily at either the member or case level and must be transformed to those levels to facilitate analysis, based on the analytical questions of interest. One transformation, for example, would be to include a field for each member indicating that member's relationship to the head of household.

Table 1 provides examples of core data elements for the data models and additional data elements that can supplement the core data, to the extent they are available.

The data model was developed based on the following key ideas:

- **Simpler is better.** Administrative data systems often are large and complex, tracking a wide array of information; however, the quality and depth of knowledge of these fields vary greatly. We model our work on the agile approach to software development: start with a simple, small area; build success, knowledge, and buy-in in that area; and then grow and develop based on user needs. This also reduces the learning curve for analysts and researchers, who need only learn a few key fields to become comfortable asking and answering analytical questions. "Simple" data minimize what analysts and researchers need to know to use the information intelligently and increase the potential for state-to-state comparisons.

on this page

[1] These questions came from direct conversations between pilot state partners and the FSSDC team, as well as from *Family Self-Sufficiency Data Center: Needs Assessment Report* (Weigensberg et al., 2014).

Table 1. TANF Data Model Inputs		
Core Data Elements	**Purpose**	**Examples**
Individual identifier	Identify members and link to cases	client ID, member ID
Case identifier	Identify cases and link to members	case ID
Case type	Create categories meaningful to program administrators	basic cash assistance, TANF, child only, two-parent family
Geographic indicator	Classify by jurisdiction	county name, administrative region
Basic client demographics	Characterize client population	date of birth, gender
Additional Data Elements	**Purpose**	**Examples**
Case status indicators	Analyze specific issues; adjust for abnormalities in the data	paid late, timed out, recipient of specific intervention
Other benefit receipt	Observe other program participation	SNAP receipt, Medicaid receipt
Head of household identifier	Identify case head of household	HOH client ID
Additional client demographics	Better describe client population	marital status, race/ethnicity, education attainment, citizenship, refugee, employment status

SNAP = Supplemental Nutrition Assistance Program

Our data model focuses on understanding caseload dynamics, because, in our experience, the data necessary to understand these questions are often some of the best-known and best-maintained information. This is in part because these data are used for federal reporting, as well as for providing financial benefits to individuals and households. We discuss areas of growth and expansion for this model (particularly in work and work activities) later in this brief. For many states, however, even this simple structure represents a significant increase in analytical capacity.

Although the data model is intended to be simple and versatile, the data must meet some basic requirements to fit in this structure.

- **Flexibility of use is essential.** Analysts want data they can use to answer an array of questions, rather than needing to seek a new report for every new question that arises. Although our data model is created for ease of use and understanding, it is also designed to maximize versatility. Like any data framework, it is designed to answer questions on a limited set of topics, but it allows for exploring those topics in depth.
- **The model needs to allow for variation in inputs.** The implementation of TANF programs varies greatly across states. Variables highly important to one state may be meaningless to another. Rather than focusing on specific fields, our model focuses on types of information (for example, case types or member demographics), showing agencies where to incorporate the information most meaningful for them without relying on data points that may be unwieldy or impractical for a particular state.

The data model described in this brief seeks to target a range of policy and practice questions related to caseload dynamics, while balancing simplicity in construction and use with the ability to customize to the needs of a particular agency.

ASSUMPTIONS OF THE DATA MODEL

Although the data model is intended to be simple and versatile, the data must meet some basic requirements to fit in this structure. In particular, cases and members must be identified consistently over time, and historical data must be maintained in some form. In addition, there are some nuances to how historical data are preserved and how changes or corrections are made to those data that can affect the interpretation of analyses on those data. Table 2 outlines the basic data model requirements and lists sample questions to assess data quality and the capacity to meet model requirements.

The case and the member are the units of analysis for the data model; however, the data do not need to be structured in this way in the agency's data system if the relevant information is tracked in a consistent format. One major goal of data transformation is to create files by these units of analysis. For example, raw data may include information on members and on

Table 2. Requirements for Using the Data Model	
Data Model Requirement	**Sample Questions to Assess Data Quality**
Data identify cases and members clearly and consistently across files.	• What are the fields that identify unique cases and members? • Are any of the identifier fields entered into the data system manually and therefore subject to error (for example, SSN, name, or birth date)? • Are the fields that identify cases and members available in all the files needed to construct the data model? • Are there any regional or other variations in data entry practices that may influence the consistency of identifiers?
Data identify cases and members clearly and consistently over time.	• Are members or cases ever issued new identifiers when they come back into the system after a case closing or other event? If so, is it possible to recognize them in the data? • How are changes to time-variant identifiers (for example, name or address) or corrections (for example, fixing a misspelled name or incorrect birth date) captured in the data? • Have changes in system field names or data entry practices influenced case or member identifiers? When did these changes occur?
Historical data are available, and the format(s) of historical data are documented.	• What time span do the available data cover? Are there differences by file type (for example, case files are available beginning in 1996 but payments files are available only after 2002)? • What is the format of the historical data (for example, repeated point in time files, spells files)? • Have there been changes in case management or database systems that have resulted in different file formats, field names, etc.? When did these changes occur? How feasible is it to combine data across these system changes?

payments; in this case, primary data processing focuses on transforming payments (which indicate case activity) into cases.

Although extracts pulled from states may not naturally default to these units of analysis, the administrative data usually must include some kind of case identifier and some kind of member identifier. These IDs are necessary to track cases or individuals. Social Security numbers (SSNs) alone are not an adequate replacement for member identifier, for two reasons. First, a certain quantity of data entry gaps, duplicates, or other errors is to be expected, a problem not associated with a sequence field assigned by the database system. Second, depending on state policy, there may be instances where individuals receiving benefits do not have valid SSNs (such as infants and the undocumented). A data set where individuals or cases are identified only by personal information fields (such as name, SSN, or address) and not by a system identifier will require a deduplication/data integration process to identify the same individuals or cases before additional data transformation.

The other major requirement for analysis of caseload dynamic questions is the presence of historical data (that is, data over a span of time). These data can be stored in several ways, but most are variations on two basic options: (1) point in time, or (2) spells. Point in time data represent a snapshot of the caseload and of case and member characteristics at specific times. Longitudinal changes can be identified in point in time data by looking at changes in certain characteristics across data sets. Spells identify start and stop dates for a given characteristic (for example, case address or benefit receipt).

Although the data model requires that data on eligibility status and benefit receipt be tracked over time, all data elements may change over time, and the level to which systems track these changes affects what longitudinal analyses are possible and how they will look. Data such as addresses are expected to change regularly. If cases or members are regularly aggregated by geographic region, every case- or member-month should reflect the region of that case or member during that month. A system that does not retain a longitudinal history of addresses may still have enough information to look at changes in caseloads over time, but as the data go further back in time, the accuracy of geographic analyses will worsen.

The necessity of data transformation is one of the details most frequently overlooked by leadership wanting to use administrative data to understand policy questions.

There are cases in which even historical analyses may be better served by new data. Even data elements that appear to be static by definition—such as member birth date—may change, not because the underlying fact changes, but because of data entry or other errors. Presuming that a change in a field such as birth date reflects a correction, the more recent data become the most likely reflection of reality at *all points in time*. Where states do have historical information for fields such as birth date, using the most recent information at all time points is reasonable because these data always have the possibility of being corrected.[2] One drawback to this approach, however, is that reports for a given point in time may look slightly different, depending on when they are created. For example, a count of members by age for July 1, 2008, created in August 2008 and the same report created in March 2012 will have slightly different numbers if there have been any changes to the birth dates of the included members.

This last example illustrates one of the perennial challenges in working with state administrative data: matching existing numbers and reports. There is a certain amount of "drift"—corrections and other retrospective data changes—that can change "final" numbers. These changes usually represent very small percentages of the overall totals, but unexplained shifts can be very uncomfortable to administrators who want the certainty of accounting in their reports. Whether this kind of drift can be pinpointed and perfectly explained depends, in general, on how thoroughly a state understands and tracks historical change in the supposedly "static" elements of the report.[3] Because drift reflects how much data change, it can in itself be a valuable point to understand as one makes decisions and assesses how to improve data quality.

Agencies that cannot consistently identify individuals or cases and that do not preserve historical information in any format will be unable to address the kind of caseload dynamics questions that motivated the development of the data model. The first step toward using data to better understand programs in these agencies is to improve data collection and storage in these two areas. In addition, before beginning analyses, it is important to understand the richness of available historical data, what data points may or may not change, and how those changes could affect reports.

DESIGN OF ANALYTICAL FILES

If a state collects data on TANF participants, including case- and individual-level identifiers, and maintains a longitudinal history of eligibility and member receipt, the only thing standing between that state and a functioning analytical data set is transforming the data. The core purpose of the data model is to provide guidance for that restructuring.

The necessity of data transformation is one of the details most frequently overlooked by leadership wanting to use administrative data to understand policy questions. In general, this transformation requires time from a programmer and, depending on the starting data structure, the necessary transformations can be quite complex. Although states may have all the information outlined in the data model description and assumptions above, they often still find themselves unable to answer policy questions like those that inspired the data model because the data are not structured in a way to make them conducive to these analyses.

Usually the biggest challenge in figuring out the appropriate structure is deciding how to track longitudinal information. As discussed above, variation exists both in which data points are available longitudinally and in when and where it is appropriate to include longitudinal information in analysis.

For analytical purposes, we have found it useful to create analytical files that are a hybrid: they combine some static characteristics and information derived from spells into a point in time data structure: usually case-months and member-months. Static case and member characteristics are the same for all observations for a given case or member; characteristics that we wish to look at over time (for example, case type, location, benefit receipt, age, household size) are snapshots for the given case-month or member-month combination. Figure 2 displays a simple overview of the data transformation from raw administrative data files held by the state to analytical files in the TANF model format.

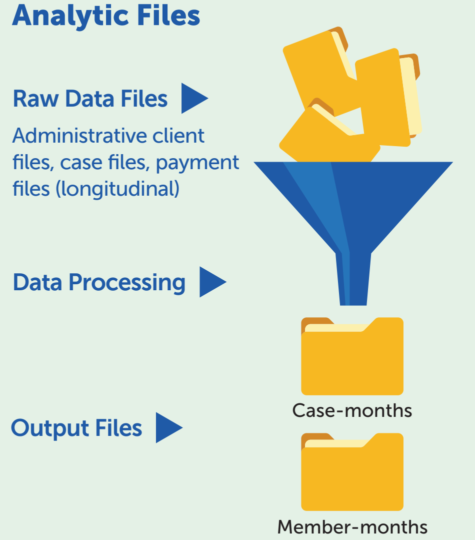
Point in time data are more straightforward, and an inexperienced analyst usually can make sense of them. These data can be easily limited to a certain date or dates for snapshot analyses or comparison between two time points, and

on this page

[2] In rare cases, there may be database-specific reasons to believe that changes in these fields do not necessarily capture more accurate data. Data from these systems should be treated accordingly.

[3] Corrections that happen after the month is over—benefits paid retroactively or refunds, for example—can also cause differences between totals pulled immediately after a month is ended and those calculated months or years after the fact. TANF data rarely have the exacting bookkeeping of accounting, and months are not "closed" after they go "on the books."

Figure 2. Transformation from Raw Data to TANF Model Analytic Files



We use data derived from spells to make it easier to answer questions of duration, entries, exits, recidivism, and churn; spell formats make start and stop dates transparent and easy to compare.

it is easy to visualize to look at things such as changes in the size of the active caseload. Point in time data also are often closest to how analysts are used to seeing and working with data, because operations data are simply point in time data where the time is the present.

We use data derived from spells to make it easier to answer questions of duration, entries, exits, recidivism, and churn; spell formats make start and stop dates transparent and easy to compare. Spell data are used to create indicators for the first and last month in a spell of benefit receipt and the first and last month the case ever received a given benefit. They also are used to create counters indicating how many months, including the current month, the case has ever received benefits and has received benefits in the current spell. After these indicators are added to the case-month level, they facilitate additional point in time analyses such as number of entries, exits, or new cases in a given month or plotted by month over time. These indicators mean that, rather than simply understanding whether a caseload is growing or declining, an analyst can easily question whether this change is driven by new cases, old cases returning, a portion of static cases, or consistent turnover (or a combination of these).

A major decision in creating a service receipt spell is when to set endpoints. Often, there may be a short break in the receipt of benefits because of administrative churn or some change in the household or its membership. In general, one does not have the information to know whether the family (or individual) was actually ineligible or whether some administrative action had been taken. Therefore, a decision has to be made about when a break in the receipt of a benefit actually constitutes an end of a spell. For example, a 2014 report on administrative churn from the U.S. Department of Agriculture defined a break of four months or less as churn (Mills et al., 2014). By this definition, multiple periods of service receipt would only be considered separate spells if the case did not receive benefits for at least five months between spells.

The specific steps and challenges in transforming raw data into the data model vary by state. However, here we provide two examples to help readers conceptualize what this process might look like in practice.

Example 1: State Alpha

Our first example comes from state Alpha. Alpha's database has a case table and a member table. Both tables contain basic information (such as county for cases or birth date for members), as well as status fields that indicate what benefits the case and member are receiving. Alpha's database includes cash assistance programs other than TANF, as well as Supplemental Nutrition Assistance Program (SNAP) and medical assistance, and status fields can indicate that a case or member is eligible for any one or any combination of these programs at a given time. There is no longitudinal history tracked as part of the database—when a case or member characteristic or eligibility determination changes, that database field is updated. However, the state saves extracts of the database at regular intervals that include a record of these changes, and state analysts regularly pull aggregated reports. Alpha's primary approach to looking at change over time is to compare counts of active cases or members, and active cases or members of specific types, across different points in time.

Figure 3. State Alpha—Analysis capacity before data transformations

- ✓ Characterize current population (individuals)
- ✓ Characterize current population (households)
- ✓ Track receipt of other benefits (SNAP, medical assistance, other cash assistance)
- ● Compare caseloads at different points in time
- X Examine changes in individual characteristics or benefit receipt over time
- X Examine changes in household characteristics or benefit receipt over time

Alpha's data are fairly rich, but they are also relatively inaccessible. Although Alpha can ask relatively robust questions about current program participants—where they live, what characteristics they have, what other programs they are enrolled in—the ability to look at case trajectories or other changes over time is limited, as shown in Figure 3. For example, it would be difficult for state Alpha to answer how long the average active case has been active, answer whether that duration has changed over time, characterize cases that used to receive benefits but no longer do, or explore whether cases often go inactive and then are reactivated. These are the kinds of questions that are often of most interest to policymakers and program administrators. Although this information does not address causal relationships or program impacts, these are the basic descriptive elements that administrators need to understand how a program is working and that evaluators need to develop studies of program impacts.

Alpha has the data to answer all these questions. When someone in senior leadership asks one of these questions, some combination of analysts, programmers, external researchers, and consultants can answer the question. However, these efforts are time intensive, costly, and slow, and after the first question has been addressed, there is often little infrastructure to address follow-up or related questions or to track the answer over time. Effective data use means the ability to relatively easily ask and answer basic questions and to explore trends and patterns in the system. Alpha's data structure consistently hinders the ability of state staff to make the most effective use of data.

Converting Alpha's data into the data model requires effort similar to that needed to answer any of these individual questions, but it creates a data set that can be routinely updated and queried to address many more questions and to facilitate data exploration. The steps to create a version of Alpha's data in the data model are the following:

1. Extract the data model inputs (including elements shown in Table 1 and any other elements that are key to Alpha's reporting and program management) from each historical extract. Add a column to each of these extracts indicating the source month and year (or quarter and year, or year, depending on the granularity of the historical extracts). There should be one extract of case-level data points and one of individual-level data points.
2. Concatenate these extracts to create two basic point in time files, one at the case-month level and one at the member-month level.
3. Create aggregated values at the case-month level as appropriate from the member-month data and append these to the case-month data. Examples include number of individuals in the household, number of individuals eligible at a time, number of children in the household, number of seniors in the household, and head of household characteristics. These values can be calculated using a SQL "GROUP BY" or similar logic.
4. Simplify case and member status variables as necessary to create indicators for TANF, SNAP, Medicaid, and other program participation at the case-month and member-month levels.
5. Create TANF receipt spells at the individual and case levels. Creating spells begins with extracting three elements from each file: (1) the unique identifier (case or individual), (2) the month/year value, and (3) the indicator of TANF receipt. A series of joins creates indicators showing whether the individual or case received TANF in the previous month and the following month. If the case did not receive TANF in the previous month, this month starts a spell that continues until the last month before a gap in benefit receipt. Iterative logic identifies these spells.
6. Use the spell data to identify (1) spell start months, (2) spell end months, (3) overall start and end months for each member or case (using the MAX and MIN functions with a GROUP BY or similar), (4) number of months of TANF receipt in the current spell, and (5) number of months of TANF receipt in all spells (cumulative). The last two values are calculated using the distance between the current month and the start month of the current spell.

Steps 5 and 6 are the most complex parts of this process. Luckily, they are also steps that look largely the same for data that begin in a variety of formats. The FSSDC has prepared (and thoroughly documented) sample code for creating and calculating spells on our GitHub repository: https://github.com/chapinhall/FSSDC. We suggest this code as a starting point for anyone who wants to do these calculations.

Figure 4. State Beta—Analysis capacity before data transformations

- ✓ Characterize current population (individuals)
- ✓ Examine changes in individual characteristics or benefit receipt over time
- X Characterize current population (households)
- X Track receipt of other benefits (SNAP, medical assistance, other cash assistance)
- X Compare caseloads at different points in time
- X Examine changes in household characteristics or benefit receipt over time

Example 2: State Beta

In state Beta, data are tracked at the level of the payment, rather than the case (although there is a case identifier that indicates payments affiliated with the same household over time). Payments have dates, amounts, and types. Individuals in a given household are affiliated with each other by the case identifier, but they are also connected to payments through benefit applications: individuals are listed together on an application, each individual receives an eligibility status based on that application, and a payment is issued for a given approved application.

In the case of Beta, it is easier for analysts to routinely look at longitudinal questions, because the full payment history is available at any given time. In particular, changes in the population of individuals receiving benefits over time are fairly easy to assess through the members/payments link. However, because there is no concept of a "case" in the routine database, it is difficult to quickly look at case size, composition, and changes at the caseload level over time (see Figure 4). (In particular, a given set of individuals can be affiliated with multiple payments of different types in a given month, so payments cannot simply be used as proxies for cases.) Furthermore, some of the same trajectory questions that were difficult for Alpha to consider (for example, how long have current recipients been receiving benefits? Do individuals or households that stop receiving TANF later come back?) are still difficult for a state with Beta's data format.

Again, Beta has all the data needed to answer questions about churn, recidivism, and caseload dynamics, but these data are not structured in a way conducive to research. In particular, the introduction of more granular units of analysis (for example, payments, applications) increases the complexity of analyses. It also introduces opportunities for accidental duplication or data discrepancies that were not a consideration in Alpha (for example, the same individuals affiliated with payments in two counties in a given month).

In practice, transforming Beta's data into the data model is similar to the process outlined for Alpha. However, for Beta, there are additional transformation steps to create the aggregated case-month and member-month files outlined above:

1. Group the payment information at the case level and simplify the individual data so that there is only one record for each member each month (rather than one record for each member/application combination). This is best done iteratively and in close conjunction with staff who oversee data collection and who understand the data and their connections to policy. "Aggregating" these data to the case and member levels involves addressing inconsistencies between records. A member may have different birth dates listed on different applications; what is the likely driver of that change (a correction?), and what should be the decision rule for selecting the appropriate birth date? How should payment types be summarized as case types? Are there types of payment information (such as refunds) that should not be interpreted as indicating a "TANF payment" was made in the month in question?
2. After data have been simplified to the case-month and member-month levels, repeat steps 3 through 6 from the state Alpha processing as detailed above.

The important challenge for a programmer transforming data such as Beta's into the data model is to consistently and repeatedly check that the data are unique at the intended level and to test any assumption that a given field is consistent across rows within the same month. Although agency policies often dictate that data should look a certain way, the nature of administrative data, especially data stretching back across any length of time, is such that files are certain to contain conflicts and inconsistencies. Transforming data into the data model can facilitate a one-time, intensive review and documentation of these inconsistencies. Although this process is time and labor intensive, after the decision rules have been implemented in code, it becomes much easier to generate high-quality, accurate analyses more quickly with the analytical data sets.

POTENTIAL GROWTH AREAS FOR TANF DATA MODEL

In addition to the caseload dynamics questions this TANF data model is suited to address, a number of additional research areas are of interest to agencies and researchers. Table 3 outlines the component information that would be necessary to expand the data model to address questions in these areas.

Employment is the strongest growth area for cross-state comparison. Other areas of inquiry are more complex, because state and county policies become much more extensively involved in preparing and interpreting the data for these topics. Therefore, it is more difficult to develop "one size fits all" research questions or data sets.

Table 3. Potential TANF Model Growth Areas

Research Area	Research Questions	Data Requirements
Employment	What are the work participation rates for cases, and how do these rates vary by caseload characteristics?	Administrative data on how cases are fulfilling work participation requirements and which have exemptions, as well as state/county policy information necessary to interpret these data points.
	What are the long-term employment outcomes of TANF participants?	Integration with an external source of employment and wage data (unemployment insurance, National Directory of New Hires).
	Questions looking at the intersection of work participation rates and employment outcomes, or at the intersection of these spaces with caseload dynamics.	Administrative data on work requirements and integration with external employment data.
Time Limits	Which cases on the caseload are approaching their time limits for benefit receipt?	Significant policy detail, in addition to caseload dynamics tracked in the data model. In particular, data provided by state must be transparent and well-defined around funding used to provide benefits (federal versus state or county diversion), around timing of exemptions, etc.
Income/Grant Amounts	What size benefits are being received by cases or individuals with a given characteristic? How do grant amounts change as a result of certain events?	Detailed accounting information on payments received, with special attention to data issues concerning aggregating this information, including risk of duplicates, corrected payments/refunds, timing of payments, and desired timing of reporting.
	What size benefits are being received by cases or individuals with a given household income? How do grant amounts change as a result of events that influence household income?	Aggregated wage data answer some basic questions. Accounting for various benefits received, EITC, federal subsidies, etc. requires much more comprehensive information, focusing on the timing of payments, potential corrections, and how this income can be appropriately aggregated.
Sanctions	Are there patterns in when sanctions occur? How do these relate to overall caseload dynamics and case characteristics?	Information on when cases are receiving sanctions. This area does not require a lot of additional data, but the research questions here tend to be state specific, because they relate to how sanctions are being used and are intended to work on a state-by-state basis.
Other Benefit Receipt	Where is the overlap between the TANF caseload and the caseloads receiving other forms of government assistance (that is, medical assistance, housing assistance, child care subsidies, EITC, etc.)? Are TANF participants taking advantage of other benefits for which they are eligible?	Rich integrated data across programs and extensive involvement from state personnel needed to identify policy realities. A simple form of these analyses is available within the current data model if the state's data already include clearly defined variables indicating overlap (for example, an indicator of SNAP receipt), which can be used to look at questions of duration, churn, and recidivism.

EITC = earned income tax credit

CONCLUSION

This brief makes an initial attempt to describe how state agencies and researchers with access to rich administrative data sets about TANF programs can restructure those data sets to make it easier to answer common research questions and to use data better for program management and evaluation. The TANF data model is presented as a way to help users understand what data are needed for these analyses and what challenges exist in preparing those data, while also presenting a roadmap for addressing those challenges.

REFERENCES

Mills, G., Vericker, T., Koball, H., Lippold, K., Wheaton, L., & Elkin, S. (2014). Understanding the rates, causes, and costs of churning in the Supplemental Nutrition Assistance Program (SNAP) –Final report. Prepared by the Urban Institute for the U.S. Department of Agriculture, Food and Nutrition Service. Available at https://www.fns.usda.gov/sites/default/files/ops/SNAPChurning.pdf.

Weigensberg, E., Schlecht, C., Wiegand, E., Farris, S., Hafford, C., Goerge, R., & Allard, S. (2014). Family Self-Sufficiency Data Center: Needs assessment report. Chicago, IL: Chapin Hall at the University of Chicago. Available at http://www.chapinhall.org/research/report/family-self-sufficiency-data-center.

This research was supported by the Family Self-Sufficiency Research Consortium, Grant Number #90PD0272, funded by the Office of Planning, Research and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services. The authors thank Scott Allard, University of Washington and Co-Director of the FSSDC, and Rupa Datta, NORC at the University of Chicago, for their valuable comments and contributions to this report. We are also grateful to our state agency partners who have been essential to the development of this work, as well as the participants in the Family Self-Sufficiency Data Center Workshop. Molly Rossow and Greg Sanders, Chapin Hall at the University of Chicago, played a crucial role in developing the data processing strategy. Mathematica Policy Research provided excellent assistance preparing this report for publication. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.