Institute for Research on Poverty
UNIVERSITY OF WISCONSIN-MADISON

Technical Report

on

Lessons Learned in the Development of the Institute for Research on Poverty's Wisconsin Administrative Data Core

by

Patricia R. Brown and Katie Thornton

with

Dan Ross, Jane A. Smith, and Lynn Wimer

Last Revised and Updated: February 2020 by Katie Thornton

Original Publication Date: December 2011

This report was originally titled, "Technical Report on Lessons Learned in the Development of the Institute for Research on Poverty's Multi-Sample Person File (MSPF) Data System." Special thanks go to several staff members of the Urban Institute—Laura Wheaton, Harvey Meyerson, Jim Kaminski, and Molly Scott—for a literature review of linkage software, a review of various name standardization techniques, and for preliminary discussions of the content and organization of this report. And thanks to Jennifer Noyes, Eunhee Han, and Maria Cancian for reading a preliminary version of this report and offering valuable comments and suggestions, as well as to Steve Cook and Hilary Shager for providing the same in revised versions of the report.

Technical Report on Lessons Learned in the Development of the Institute for Research on Poverty's Wisconsin Administrative Data Core

INTRODUCTION

In 2008, the Institute for Research on Poverty (IRP) at the University of Wisconsin–Madison and the State of Wisconsin agencies now included in the Department of Children and Families (DCF) proposed a set of integrated data development, analysis, and evaluation activities designed to improve our capacity to analyze Temporary Assistance for Needy Families (TANF) and related administrative data. This effort was facilitated by a number of existing strengths: (1) a new administrative structure that brought TANF, child welfare, child care, and child support administration within a single department (DCF) as of July 2008; (2) substantial prior experience using administrative data for research, program monitoring, and management improvement, and high-level commitment to expanding these efforts; and (3) a long-term collaborative relationship between Wisconsin State agencies and researchers at IRP.

The project—"Building an Integrated Data System to Support the Management and Evaluation of Integrated Services for TANF-Eligible Families"—was funded under the Office of Planning, Research, and Evaluation (OPRE), Administration for Children and Families (ACF), U.S. Department of Health and Human Services' funding opportunity: *Federal-State Partnerships to Build Capacity in the Use of TANF and Related Administrative Data.* The project's ultimate goal was to create a data resource to support the integrated analysis of the earnings, income, and multiple program participation trajectories of Wisconsin families participating in TANF and other income and work support programs. The expectation was that such a resource would provide the basis for important contributions to program evaluation and administration, as well as basic research.

In this report, we provide detailed information about our experience in creating this integrated data system, which has come to be known as IRP's Wisconsin Administrative Data Core (WADC). The data core is created by drawing on information extracts from the full universe of clients or participants in the State of Wisconsin's electronically available administrative data on public assistance, child support, child welfare, unemployment benefits, K–12 public education, and incarceration, and merging them to create a single file of unique individuals, called the Multi-Sample Person File, or MSPF. The 2018 MSPF totals about

7,724,000 individuals.[1] The Wisconsin Administrative Data Core refers to this master MSPF one-record-per-individual file, along with yearly program participation data, which are linkable aggregation files (parent/child, and case-level data) and participation files (monthly benefits, eligibility, payments/receipts, or spells).

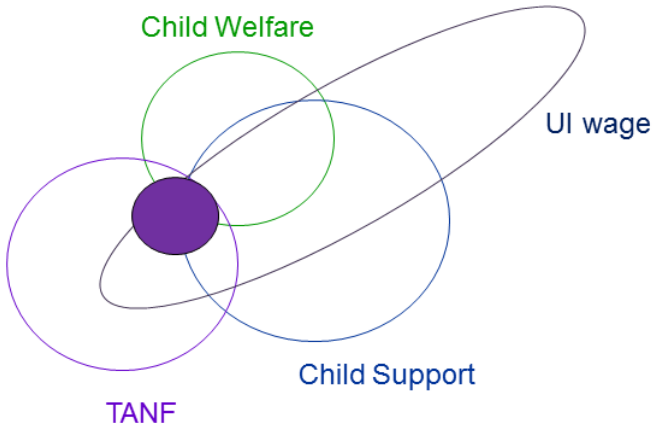
This report is organized as follows. In this section, we outline our overall approach to the creation of the WADC, and IRP's unique ability to do this work. We then review the WADC creation process, including: the data sources employed, a list of key challenges and considerations in developing the integrated data system, the staff resources needed, technical steps in the preparation of the unique individual-level master file and program participation data, and some results of the data linkage. We conclude this section with a discussion of plans for maintaining and expansion of the WADC. We then give a brief comment on considerations in the use of the data core by social science researchers.

Overall Approach

Prior to the creation of the data core, research and evaluation projects conducted at IRP using administrative data from the State of Wisconsin typically relied on creation of a specialized data extract created for the particular project. These extracts were created by selecting a sample of cases from one of the administrative data systems, defined by a particular set of characteristics. Data on the individuals in this original sample were then merged with information from other administrative data to measure participation in those systems. This was a well-used model of research for many decades, and is illustrated in a stylized way by Figure 1. Note that the research started with a project-specific **sample** generally drawn from one data source, and additional administrative data were gathered only for that sample.

[1] The total individuals in the 2018 MSPF includes individuals with limited personal identifiers who may or may not be duplicates of other individuals in the MSPF. It is not, nor is it intended to be, a comprehensive list of all living residents of the state of Wisconsin. The MSPF includes deceased individuals who were once participants in specific programs, as well as individuals in programs who have since moved out-of-state.

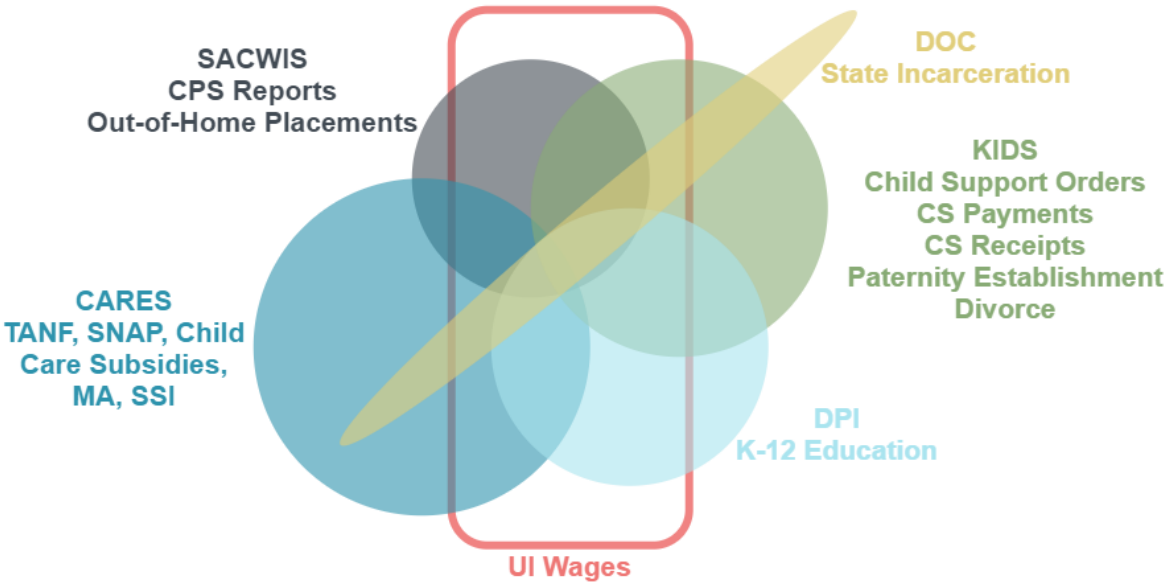
Figure 1: Old Model for Research with Administrative Data



In building the current data core, the broad goal is to move away from the above model in order to more flexibly measure the use of multiple State programs by single individuals or family members, concurrently, and over time. In order to achieve that goal, the three operational goals in building the data core are to: (1) create a file (the MSPF) with one observation per individual, with no individual knowingly appearing twice; (2) enable researchers to group individuals by case and/or by various definitions or constellations of family; and (3) provide easy-to-use program-participation data files, linkable to the MSPF. Therefore, the main task of the computer programming effort to create the MSPF is to clean and standardize identifiers and other demographic variables in each data source, and to match-merge or link individuals across data systems, creating a final research data file that contains only one observation per individual.

The creation of the MSPF has shifted the research data system model employed by IRP, which is illustrated in a simplified version by Figure 2. Note that the full **universe** of cases or individuals from one source of administrative data can be analyzed, including both those who participate in other systems and those who do not. Researchers can also easily focus on subsets of individuals who participate in certain constellations of programs or services from multiple administrative sources. The full merging of multiple sources of administrative data independent of the formulation of specific research questions has significantly broadened the set of questions that can be addressed with the constructed administrative data system.

Figure 2: New "MSPF" Model for Research with Administrative Data[2]



IRP's Unique Ability to Build the Wisconsin Administrative Data Core

IRP has a long-standing history of conducting applied public policy research, much of which has utilized administrative data from the State of Wisconsin. In a series of ongoing research agreements from 1983 to the present, IRP worked with the Bureau of Child Support to evaluate a series of child support reforms in Wisconsin. That research involved collecting and recording data on a sample of divorce and paternity court cases (known as IRP's court record data, or CRD), and merging these records with other administrative data sources. In the early years of this project (the 1980s), administrative data merged with the CRD included: Department of Revenue tax records, Unemployment Insurance (UI) wage record data, Aid to Families with Dependent Children (AFDC) cash benefits, Food Stamp benefits, and Medical Assistance (MA) eligibility information. In the late 1990s, a subsequent project with the State of Wisconsin (the Child Support Demonstration Evaluation [CSDE] project), expanded the scope of linked administrative data for evaluation research by including administrative data on all TANF recipients during a five-year time period, and linking those data with administrative data on Food Stamps (changed to Supplemental Nutrition Assistance Program [SNAP] in 2008), MA, child

[2] In this figure, the solid colors represent data sources for which the full universe is included in the WADC. The outlined box represents UI wages, which are matched to all the other data sources in the figure. UI wage records that match an individual in another data source are included in the WADC, while unmatched UI wage records are excluded.

support payments and receipts, and UI wage records. More recently, IRP entered into several research agreements with the State of Wisconsin that rely on child welfare data merged with TANF, SNAP, and MA data, as well as child support payment and receipt information, education outcomes, and incarceration spells.

The long history of cooperation between IRP researchers and State of Wisconsin policymakers has laid the groundwork and infrastructure for the development of the WADC. Extensive resident expertise has been developed and maintained at IRP for overseeing data-sharing agreements, as well as developing and monitoring data security and appropriate human subject protocols. IRP also has the computer programming capabilities (in addition to the necessary hardware and software) in a programming staff whose members have extensive experience in the details of specific State-administered programs—experience necessary to understand and properly handle each of the separate administrative data sources used in the WADC. IRP research staff and affiliated faculty contribute additional expertise in how programs are administered, which also facilitates the appropriate use of the data system.

As noted above, development of the WADC in Wisconsin was initially supported by "Building an Integrated Data System to Support the Management and Evaluation of Integrated Services for TANF-Eligible Families," funded by the U.S. Department of Health and Human Services.[3] This grant allowed IRP to develop expertise in child welfare data, and to merge those data with the universe of TANF/SNAP/MA cases and the universe of child support cases from the State of Wisconsin. The first version of the MSPF (MSPF 2008) also included incarceration cases from the State prison system and the Milwaukee Jail system, and a match with Unemployment Insurance administrative data to include wage record data for all employed individuals in the WADC. In subsequent versions of the MSPF (2010–2018), IRP improved the unduplication and linkage process, added new participants and current participation data, extended the data coverage back in time to the beginning of electronic record-keeping (mid-1990s) for most data sources, and added administrative data on unemployment spells and Unemployment Insurance benefits. Funding for all versions of the WADC has been

[3] *Federal-State Partnerships to Build Capacity in the Use of TANF and Related Administrative Data*–a grant administered through the Office of Planning, Research, and Evaluation (OPRE), Administration for Children and Families (ACF).

supplemented from multiple sources (mostly faculty-run grants for projects that have made use of this resource), and the WADC now provides the basis for a number of public policy research and evaluation projects.

A fundamental question that could be raised is: "Why don't States, themselves, automatically merge their administrative data systems, by unique individual identifiers, and thereby create such multi-system data resources, available for both administration and research?" The case has been made for this by the authors of an excellent publication by the Substance Abuse and Mental Health Services Administration (Heil, Leeper, Nalty, & Campbell, 2007). They write: ". . . data integration refers to the practice of linking (i.e., matching) diverse, routinely maintained administrative data sets at the client level to obtain a rich picture of client encounters across State agencies" (p. 1). That report provides good technical detail on the issues of interagency data-sharing agreements, unduplication and linkage software, practical guidance about a common list of high-quality identifying variables, the creation of unique client identifiers, and staffing needs for a data integration effort.

The State of Wisconsin did, in fact, begin such an effort with a "Master Client Index," which was designed to gather individual identifiers from multiple administrative sources, and add a unique master identifier to these individuals. The primary use of this index was to help differentiate "new" and "established" individuals in State administrative data systems. As far as we know, however, it has not been used for constructing research data sets spanning data systems. With such an index, a number of problems may arise, including the fact that most administrative data systems were not initially designed with a strict requirement to maintain only a single record per individual. Since State administrative data must maintain a high degree of "legal" accuracy for case management purposes, the under-matching of individual records (due to limited demographic information) can result in a proliferation of records (multiple identities) for some individuals. To correct this in hindsight is difficult to impossible. Second, since the data systems were not designed to be linkable, there is no commonly agreed upon set of high-quality identifying variables available in all systems for matching specific individuals with individuals already in the master client database. Third, in cases of conflicting information on individuals between data sources, agencies with distinct missions face challenges in determining which data source should take precedence. Fourth, related complications make it difficult to determine which agency should be responsible for the resolution of conflicting or erroneous information, or

for unduplication of multiple records involving a single individual. Fifth and finally, it is not clear how an integrated data set changes over time, and how historical records of such changes can be maintained.

The basic challenge to any State attempting to build such an integrated data system is how to preserve legally accurate records, suitable for real-time administrative use or case management, while simultaneously allowing for only one record per unique individual. By building the data core in a university research setting, this basic challenge is more easily resolved by compromising somewhat on legal accuracy, and emphasizing "one record per unique individual." The technical issues can then be resolved. For instance: we can compare variables from among the various administrative data systems to determine variations in data quality, depending upon the purpose of the administrative agency, and select accordingly. For example: unemployment benefit data may have the most accurate set of Social Security numbers [SSNs]; the child support data system may have the most accurate information on legal parenthood; and the public assistance agency may have the most accurate information on household membership. We can balance over-matching and under-matching of individuals through our choice and use of various programming techniques. Leaning too far in one direction or another would present error or bias in aggregated research results. We can tolerate some degree of individual-level error since statistically valid conclusions can be drawn from less than 100 percent accurate record linkages. In addition, we can build the data system as of a point in time, which freezes and stabilizes the data for research, which can take many months to complete, and allows the analyses to be duplicated in future years.

There are, certainly, limitations to this approach. First, the data aggregation process takes some time, and becomes dated over time, and therefore cannot be used for the most current, up-to-the-minute analysis. Second, to provide fairly up-to-date information for research (say, on a yearly basis), the work has to be repeated, or a system developed for creating new sets of updated files. Third, the programming work required for the development and updating of this research data system depends upon a high level of expertise by staff programmers in each of the State's administrative data systems, which can take months and years to develop, maintain, and fund.

CREATION OF THE WISCONSIN ADMINISTRATIVE DATA CORE

As noted above, the broad goal of the WADC is to allow researchers to measure use of and outcomes from multiple State programs by single individuals or family members, concurrently, and over time. In order to achieve this, the three operational goals in building the WADC are: (1) to create a file (the MSPF) with one observation per individual, with no individual appearing twice; (2) to enable researchers to group individuals by case and/or by various definitions or constellations of family; and (3) to provide easy-to-use program participation data files, linkable to the MSPF.

The fundamental tasks of creating a multi-sample person file are to clean and standardize variables used for linking, and to match-merge individuals from all data systems with one another, unduplicating and linking observations to the extent that only one observation per individual remains in the final version of the data. The information needed for match-merging the data systems includes individual identifying characteristics and demographics. This information normally can be gathered from one or two tables or data files extracted directly from each of the administrative data sources. Most data sources have files listing all individuals and their demographic information. Additional files are often available that record the family relationship of all individuals in the household and/or in the case. Therefore, the two basic types of identifying information are: (1) individual characteristics or demographics, and (2) the identity of a person's parents or their children.

Three desirable characteristics of identifying information are: that the information is commonly recorded, relatively uniquely identifying, and immutable (i.e., personal characteristics that do not change over time). The longer the list of identifying variables that are available for each individual, the more complete and accurate the match-merge between individuals will be. The variables available from our data sources, and used in match-merging to build the MSPF are: Social Security Number (SSN) and SSN verification code; personal identification numbers (PINs)[4] cross-loaded from one data system to another; and names, sex, dates of birth and death, place of birth, and parent identifiers (first name, date of birth, and SSN of both mother and

[4] PINs (Personal Identification Numbers) are often assumed to generally identify individual persons. Most data systems, however, contain multiple PINs for some individuals, with PINs functioning to identify a person with respect to specific family units, or cases, for purposes of case management. Also, some data sources use PINs to identify specific spells for an individual, with the understanding that a person could have multiple spells, and therefore multiple PINs.

father). Race and ethnicity are used indirectly to refine our name standardization process. Indicators of twin or triplet birth, and of adoption, are used to refine match-merging procedures for these individuals.

Once the work of creating a multi-sample person file with only one observation per individual has been completed, sets of research files are constructed for research purposes. These research files fall into three basic categories: (1) a reduced set of demographic variables in the MSPF (removing uniquely identifying personal information for purposes of individual anonymity), with the addition of sample and administrative data source indicators, and a constructed (i.e., masked) unique PIN; (2) a set of files that allow aggregation of individuals into administrative cases, or into family units; and (3) sets of participation files that provide information on program participation, over a specified time period, for each individual. All aggregation and participation files are linkable to the MSPF with the constructed unique PIN.

Administrative Data Sources

For purposes of building the MSPF and related files, we use State of Wisconsin administrative data sources, all of which differ in content, structure, and purpose. Research use of administrative data is normally considered secondary, in contrast to data that are collected specifically for research purposes. As such, these data files are more complex in structure than would be ideal for research use, generally with a multitude of inter-linked tables. Searching through available data fields for information of interest to policy researchers, evaluating different levels of data quality or completeness, and recombining data into a structure usable for particular types of analyses require a depth of understanding of both the research goals and the administrative data purpose, structure, and quality. Some data sources are more conducive to policy research than others, and a large amount of detailed knowledge about each particular data source must be gained by individual members of the programming staff who are doing the data handling. This expertise has to be kept current over time as policies and administration of programs change. (See Goerge & Lee, 2002, for a good discussion of the advantages and pitfalls of using administrative data for research purposes.)

The working goal of this project has been to extract and use the universe, or full population, of cases and individuals from each of the administrative data sources, in order to accurately measure program participation overlap. For the most current (2018) MSPF, we use the

universe of individuals and cases from public assistance (obsolete CRN and current CARES), child support (KIDS), child welfare (eWiSACWIS), State prison incarceration (DOC), Milwaukee Jail data, public schools (DPI), and unemployment benefits data (UI benefits). We also include individuals from IRP's child support-related court record data (CRD). All of the individuals in the above-listed data sources are then also matched with UI wage record data. In the current version of the MSPF, the data universes have a complete accounting of individuals and cases from the beginning of their operation (generally in the late 1990s)[5] through calendar year 2018 (with the exceptions of UI benefit data, which are currently available only back to 2006, child welfare back to 2004, and DPI data, back to the 2005–2006 school year). More details on the administrative data sources used in the WADC can be found in Appendix A.

Challenges, Considerations, and Solutions

The WADC was created in order to provide a measure of State administrative program participation and overlap. In the course of creating these files, there were a number of issues or challenges to be considered, including: (1) whether to use samples or the universe of individuals; (2) how to structure the data (by individual, by case, by family unit, or by child/adult status); (3) how to determine useful identifying variables; (4) how to pre-process or "clean" administrative data; (5) how to address considerations of legally accurate data, "fuzzy" matching, over-matching, under-matching, and cross-matching; (6) how to determine which matching technique to use (probabilistic, deterministic, or a mix of the two); and (7) how and when to update the integrated data system.

(1) Whether to use samples or the universe of individuals

IRP researchers have a long and well-established history of using **samples** extracted from administrative data to investigate questions of public policy. With the almost universal move to electronically stored administrative data in the last two decades, however, it has become increasingly apparent that there are advantages in considering the **universe**, or the full population, of cases, families, or individuals, when making administrative data available for research. Some of the issues that we considered in moving from samples to the universe are:

[5] Electronically available administrative data systems began operation at different points in time, and each developed specific policies regarding the uploading of historical (conversion) data.

- In decades prior to electronically available data, with the handling of manually gathered, hand-coded data, researchers could not afford the money, staff, electronic data storage costs, or the Central Processing Unit (CPU) time to work with large, all-inclusive data sets. When data are stored and handled electronically, however, particularly with recent reductions in CPU and electronic storage costs and the availability of inexpensive linkage software, the technical reasons for using samples rather than the universe are greatly reduced or eliminated.
- When using the universe of data, proportionally small subgroups of cases or individuals can be observed in sufficiently large numbers to discover patterns and trends for those proportionally small, but often important, subgroups. Also, when merging administrative data, case types that occur infrequently in one data system will not be found in other data systems unless the universe of both systems are available to be merged. Some examples are: small ethnic groups, extremely young or extremely old age groups, and substantiated child abuse cases.
- A request for the universe of cases from a State data system is usually more straightforward and easier to handle, from the State agency computer programmer's point of view, compared to a request for a subset of cases to be matched on a set of supplied identifiers, for example, or a 1-in-10 random sample, or individuals participating in programs from a discrete set of months or years.
- It is sometimes the case in State administrative data systems that the history of updates to information is not systematically retained. These data sources are designed to administer a program, and not for the purposes of historical research; therefore, data fields may be overwritten with more-current information, and history is thereby lost. For research purposes, we often want to preserve this history. If we request and receive the universe at a point in time, or regular points in time, then we can preserve some of that history that might otherwise be overwritten in the State's ongoing and current data.
- Observations applicable to a particular time period are often added to State administrative data months or even years later than they actually occur. These

additions are easier to observe and measure when the universe from one time period is compared to the universe from a later time period.[6]

Again, one overarching purpose for building the WADC is to record and measure the overlap of all individuals from these administrative data sources. In order to do this measurement, we need to have the universe of all individuals from the data sources we incorporate into our data files. With data extractions that are confined to a sample, we are limited in the types of analysis that can be done.

(2) How to structure the data—by individual, by case, by family unit, or by child/adult status?

In prior work of merging administrative data, we constructed our data files around a specific unit of analysis; for example, cases, or child/adult status with adults being classified as custodial or noncustodial. In working with these data files, and attempting to update them over time, we found a number of problems: custodial parents became noncustodial, and vice versa; children aged into adults; individuals (as well as the same constellation of individuals) appeared in multiple "cases." We have learned from experience that a simple structure of one observation per individual is a preferable way to structure merged administrative data.

If any kind of history of benefits or participation is developed at the onset, or planned for the future, then alternative data structures (parent/child, custodial/noncustodial, case) will rapidly deteriorate as individuals change roles, or age from child status to adult status. We believe that structuring the master file simply as an individual file, without regard to role in the case or the family, or the age of the individual, is the easiest type of file to build and to maintain. Once a one-observation-per-individual file is created, then additional parent-child and case files can easily be constructed, as well as program participation files, and made linkable with a constructed master personal identifier.

(3) What are useful identifying variables?

Commonly used individual identifiers include: name, date of birth, sex, and SSN. Names are very important identifiers, and usually have to be used to some extent. They suffer, however,

[6] An example of a data system that we have found to "age" over time is the UI wage record data. In fact, we generally do not provide UI wage record data to researchers until it can be extracted after two full calendar year quarters have passed beyond the end date of the last quarter of interest. That is, wage record data for the last quarter of 2018 would not be provided to researchers until an extract could be done after June 30, 2019.

from a multitude of data handling issues, including: misspellings; nicknames; transposition of first name and middle name, or of first and last names; changes in last name due to marriage and divorce; changes in name due to adoption; hyphenated names (particularly last names); cultural differences in the use of last names; cultural differences in name preferences; and degree of commonness/rarity of names. All of these issues must be addressed when using names for identification purposes, and require extensive programming code to handle.

Race/ethnicity is another variable, or set of variables, that are often used for record matching. These codes also suffer from a number of issues: Who is doing the race/ethnic classification (a caseworker or the individual in question)? Are multiple race and ethnic choices available for coding? Even if classified by the individual in question, and with as many codes as desired, has an individual's self-identification changed over time, as well as their willingness to indicate their race/ethnicity? In our work on the MSPF, we have not found race/ethnicity codes to be very useful in matching, and we do not rely heavily on these variables.

With the qualities of relative uniqueness and immutability in mind, we have developed a set of additional variables that are useful for matching in certain cases. One of these is the PIN number (Personal Identification Number)[7] supplied in one data source from automated data linkages with another data system. Additional variables are: birthplace, parent identification (first name, SSN, date of birth, of both mother and father), and death date.

Birthplace is recorded in some data sources, although not with great completeness or accuracy. Birthplace information often suffers from either misspellings (of full place name), or of miscoding (based on "memory" of codes) by caseworkers. For example, it would appear in some of our data sources that there are as many individuals in Wisconsin from Mississippi as from the neighboring state of Minnesota, since many caseworkers code Minnesota as "MS," instead of "MN." In terms of data handling, then, should "MS" be considered as missing, changed to "MN," or left as "MS"? Such are the questions and decisions to be made on something as simple as a state code for birthplace.

[7] A personal identification number, or PIN, is an individual-level identification number used within a particular administrative system. Different data sources generally create their own numbering systems, and therefore have a unique set of PINs. Administrative data systems that have a need for frequent exchanges of data often record the PINs of one another, along with the exchange of information.

Parent identification is the most interesting piece of identifying information that we have developed for match-merging of individuals. In five of our most important data sources, we have parent-child information, and thus can often identify the parents, particularly the mother. Sometimes, with very limited information on an individual, if we know the mother's first name or date of birth, we can be fairly sure of a positive match.

Death dates are important identifiers for specific individuals in our administrative data sources; however, completeness and accuracy of information differs by source. If an individual has left a certain program, then capturing and recording the death date for that individual will be rare. For other systems, knowing the fact of death is an important piece of information, but the exact date is not of great importance. Often we find deaths recorded as the first day of the year or the first day of the month in which the individual died—not the exact date of death.

(4) How to pre-process or "clean" administrative data?

When extracting administrative data files for unspecified future research purposes, it is advisable to retain as much information from the raw source of the data as possible. However, when attempting to merge data sources on common variables, it is necessary to recode and standardize the various formats of the identifying variables into common formats. Pre-processing (or "cleaning") of administrative data, therefore, needs to be done at an initial step prior to match-merging with other data systems. Pre-processing includes making decisions on which data fields can be used in the unduplication or match-merging effort (given data-quality characteristics), and on the set of observations brought to the match. Cleaning prior to matching with other data will ensure the highest possible rate of successful matches. Trying to match on "dirty" data will result in lower quality, less successfully matched data files. Experienced staff members, with a history of handling data from specific sources, should select the required variables and observations, and handle the cleaning and pre-processing, as each data source has its own, and often undocumented, peculiarities.

Several of the broad categories of data cleaning include creating standardized versions of certain data fields (such as name and place of birth) and eliminating unusable observations. Standardizing information includes: eliminating illegal characters from data fields; changing mixed-case character data to uppercase; changing character data to numeric data whenever possible; resolving inconsistent or conflicting information; parsing text into separate fields (for

example, parsing “Jones Jr” into “Jones” and “Jr”); and identifying or collapsing missing data codes. Eliminating unusable observations might include eliminating observations with no identifying information (for example, an individual named: “Unknown Father”); eliminating observations in the data that were used as test cases for training new personnel (for example, an individual named “Mickey Mouse”); or eliminating extraneous case members (for example, removing from a child maltreatment case the name of a teacher who reported a suspected child abuse situation).

Variation in data quality between administrative data sources is another issue in data cleaning and handling. For example, in our work we have found that the coding of SSN is very accurate in the UI unemployment data, which relies on records of employment by SSN in order to distribute unemployment benefits. Public assistance programs record and validate SSNs for the recipients of benefits; but others in the same household who are ineligible are often included in the records, but with unverified SSNs. And the recording of SSN in the child welfare system has been found to be often inaccurate or missing. This is understandable and completely reasonable, since getting this particular piece of information, and recording it correctly, would not be expected to be a high priority in the work of child welfare caseworkers.

Some data fields are electronically loaded into an administrative system from some other source, and other fields are manually data-entered. Both sources of data can be prone to error. Electronically loaded data are, of course, only as good as their source. And manually coded and data-entered information has its own list of possible problems; for example, transposition of numbers (in SSN, or dates of birth), and misspellings of names or places of birth. Manually coded and data-entered information also suffers from the issue of the source of the information. Is the coding of race/ethnicity based on caseworker observation or self-reported by the individual? Are the SSN data entered from viewing a document, or recited by an individual from memory for themselves or for another family member?

During the early months of a new electronic administrative database, information on individuals and cases is often automatically loaded from an obsolete predecessor data system, with many data fields left blank, or given generic codes. Automated conversion codes are often not documented, but rather deduced over time by individual programmers who work extensively with specific data sources. There are also administrative procedural changes, over time, that

affect the data—in terms of extent of use, quality, available codes, and meaning of codes. These administrative changes in procedure are not always documented, or may be documented in a place not readily accessible. Changes in data fields can also occur with a change in personnel of administrative system designers. These are even more rarely documented.

For a number of State-administered programs, the day-to-day handling of the program and the recording of data is based at the county level. There are often differences between counties in extent of data entry, meaning of codes, and use of special codes to indicate specific issues (for example, we have found indicators of missing date of birth codes to differ by county, which include: 1/1/1900, 2/2/1922, 2/22/1922, 3/3/1933, or 1/1/1950).

The following publications are useful resources for discussion of data quality, pre-processing, data cleaning, standardization of individual identifiers, as well as data linkage techniques: Goerge & Lee (2002); Herzog, Scheuren, & Winkler (2007); and Grannis, Overhage, & McDonald (2002).

(5) How to address considerations of legally accurate data, "fuzzy" matching, over-matching, under-matching, and cross-matching?

When working with State administrative data, one must remember that these individual data sources were developed for the purpose of administering and maintaining legally accurate records for case management in a variety of State programs. They are data sources not primarily designed for research or for linking with other data systems, and they often do not intend or expect to limit the identity of specific individuals to one observation per system. For example, in the KIDS data system, designed to record and manage child support orders, it is often the case that a father has multiple child support cases for children with multiple mothers. Sometimes the identifiers on that father are largely missing or incorrectly reported. If that father becomes identified as two different individuals, in two different child support cases, this does not affect the recording of those child support orders. And it may not be a top priority for the child support data system to be strict in reconciling the multiple identities of this father into a single record, particularly when the father's identifiers are conflicting.

For our research purposes, however, we want to reconcile these multiple identities and unduplicate this individual's multiple appearances in the data. To do this, we must sometimes accept a match between two records that is not exact, is based on somewhat incomplete or

approximate information, or is probabilistically matched (based on weighting for the rarity of certain pieces of identifying information). Inexact matching of any sort, whether through deterministic or probabilistic matching, is often referred to as "fuzzy matching." Since our research conclusions are based on large-scale patterns, and individual data are always aggregated, a rare mismatching of two observations into one is not a serious problem in terms of overall research conclusions. In contrast, even rare mismatches are unacceptable in administrative systems used to monitor program compliance and eligibility. Our fuzzily matched data are constructed for research purposes and therefore never can be considered legally accurate. Moreover, individual identifying information cannot be observed (except by those involved in building and testing the merged data files) or reported.

A significant difference, therefore, between the administrative data sources that we use and the result of our matching efforts (the MSPF) is the tightness of our matching. Since our data core files have a lower requirement for legal accuracy compared to the records kept by State agencies, we can afford to be more aggressive in our unduplication and matching efforts. In the process of unduplicating records internal to particular data systems, or matching or linking records across different administrative data sources, there are four possible outcomes of our matching efforts: true positive (correct link); true negative (no link possible); false positive (erroneous match); and false negative (failed link). While trying to maximize true positive matches, and correctly recognize and accept true negatives, we work to minimize and balance false positive and false negative matches. In fact, we consider a false negative (the failure to match duplicate records) as nearly (though not quite) as serious an error in data handling as a false positive (erroneously linking two records from two distinct individuals). After creating several versions of the MSPF, we have found that over-matching (false positives) causes more problems in the analysis of data, so we have strengthened our programming code to reduce over-matching.

After the MSPF data file is built, unique masked IDs for individuals are assigned to each individual, and participation records are created. We can then identify some of our false positive matches. This may become obvious, say, when we find a "single individual" who appears to be concurrently serving two different prison terms (assuming the dates for prison sentences are accurate). We use these impossible or unlikely overlapping participation records to identify false

matches, and, in an iterative process, prevent the erroneous match of those individuals in the next round of the MSPF.

The primary programming task in building an MSPF-type data file is to eliminate duplicate observations (unduplication, or de-duplication) for the same individual, both within each data source and between data sources. Many of our administrative data sources, however, have limited identifying information for many individuals. For example, it may not be particularly important to record detailed information about children in a child support case, so a child who has a record in the child support data system may appear some years later as an adult who is now the mother of a child. She may, therefore, appear in this role as mother in a second observation, since the two observations may not be clearly matchable. For example:

Data Source	**Type**	**First Name**	**Last Name**	**Date of Birth**	**SSN**	**Mother's First Name**
KIDS - 1	Child	Pat	Brown	7/11/1970		Wanda
KIDS - 2	Adult	Patricia	Brown		123-00-4567	

In the above example, the child record has a date of birth and a mother's first name, but is missing the SSN. A second, adult record, with same (common) last name, has a different first name, missing date of birth, missing mother's first name, but has an SSN. The State's child support administrative data system would not, and should not, consider these two individuals as the same person. In reality, however, they are duplicates of the same individual, and for purposes of the MSPF file, we do want to make the determination that these two records are for one person (unduplicate). We can do this by including more individual identifiers (perhaps place of birth or middle name) from the two observations we have for this individual, or by considering additional records from additional data sources. For example:

Data Source	**First Name**	**Last Name**	**Date of Birth**	**SSN**	**Mother's First Name**
KIDS - 1	Pat	Brown	7/11/1970		Wanda
KIDS - 2	Patricia	Brown		123-00-4567	
CARES	Patti	Brown	11/7/1970	123-00-4566	Wanda

We have now added information from a different source, the CARES public assistance data system. When we consider a third record from another source, even though nearly all of the identifying variables do not exactly match (there may be alternative spellings of the first name, or a transposition of month and day of birth, or a mistake in the data entry of the SSN), we can determine that these three records are, in fact, the same individual, and in our MSPF we collapse these three records into one. The third observation has enough identifying information on this individual that the other two poorly identified observations can be determined to match one another, as well as both matching the third observation. This is an example of "cross-matching" (or "chaining"), and illustrates how linking multiple data sources can significantly increase the number of matches, and eliminate the duplication of individuals.

(6) How to determine which matching technique to use (deterministic, probabilistic, or a mix of the two)?

We use two basic match-merging techniques in our building of the MSPF data file: deterministic and probabilistic. Deterministic matching links records that match exactly on a set of given variables. Probabilistic matching is based upon algorithms that determine the probability that two records represent the same person. Matching on a rare personal characteristic is given a higher matching score than matching on a common characteristic. Variations in recorded identifying data (such as spelling variations of names) are considered in probabilistic matching.

For electronic data handling and implementing these techniques, we use SAS software, and do most of our deterministic matching with SAS programming code. Probabilistic matching is particularly useful for minimizing the problem of false negative matching (failure to match) (see Goerge & Lee, 2002). We currently use some limited probabilistic matching in our programming code, and are looking to expand it in the future.[8]

[8] We formerly turned to probabilistic matching techniques available through a public domain AF/SAS application, The Link King software. We no longer use The Link King because we found that it caused too many false positive matches within large datasets by relying too heavily on fuzzy matches of SSNs. Since we have a large percentage of the population of Wisconsin in our data, and since SSN ranges were assigned according to State and date issued until 2011, it is not hard to find two distinct individuals with the same or similar birthdate and the same or similar name. Family members, particularly twins, were often given consecutive SSNs. We also stopped using it because it did not give us the flexibility to use parental information and place of birth in the matching process. We

Deterministic logic is our most common method of match-merging, and goes something like this:

"If KIDSfirstnameX equals CARESfirstnameY, and

KIDSssnX equals CARESssnY, and

KIDSdobX equals CARESdobY, then

personX matches personY."

These kinds of statements can have hundreds of variations, based on the set of identifying variables available for both of the individual records being matched. Some of the variations in programming statements, based on variations of these variables, can look like this:

"If KIDSfirstnameX equals CARESfirstnameY, and

KIDSssnX equals CARESssnY, and

KIDSyearOFbirthX equals CARESyearOFbirthY, and

KIDS**month**OFbirthX equals CARES**day**OFbirthY, and

KIDS**day**OFbirthX equals CARES**month**OFbirthY, then

personX matches personY."

(In this example, we accept a match on the transposition of month and day of birth in one of the observations, if first names, SSNs, and year of birth also match.)

We generally do not match two records based on a single identifying variable, including SSN. Even with exact SSN matches, we normally require some other identifying variable to confirm the match, such as first name or date of birth. Since we have access to an uncommonly long list of identifying variables, including place of birth and identifying information on parents, we rely on our own deterministic programming as a primary method for identifying matches. For

still believe The Link King is quite useful for smaller populations if twins can be excluded or for those who were issued an SSN after the numbers were randomized in 2011.

records of individuals with fairly complete identifying data, these kinds of deterministic statements yield many accurate matches.

We also do some mix of deterministic and probabilistic programming, where we use probabilities of commonness/rarity of first and last names, along with deterministic statements about other variables such as sex, date of birth, etc. Other attempts to link records across data systems have also found that using both deterministic and probabilistic, as well as hybrid approaches can be very useful (see Gomatam, Carter, Ariet, & Mitchell, 2002).

We have no guarantee that all our resulting matches are correct. One of the creators of The Link King software succinctly described record linkage quality (Campbell, 2005): "Deterministic and probabilistic algorithms classify linkages along a continuum. At one end are 'definite' matches. At the other end are 'definite' non-matches. The remaining linkages contain discrepancies that lead to varying degrees of uncertainty about the appropriateness of the linkage."

Through use of the resulting administrative program participation data, we compile lists of those linkages of individuals that follow-up evidence proves to be incorrectly matched; and we use these lists to prevent these specific false matches in future versions of the MSPF.[9]

(7) How and when to update the Wisconsin Administrative Data Core?

Unlike the original sources of administrative data, most of which are updated on a daily basis, the data core research files that we build are static as of a given point in time. With the passage of time, the question becomes: How do we add participation data for individuals in the data core? How do we add individuals who are new entrants to the administrative data sources in subsequent months and years? How do we add individuals from new data sources?

It is a fairly easy task to extract participation data for subsequent months, quarters, or years, and make this data linkable to the data core for individuals already found in the master file (MSPF). If, however, we want to add individuals to the MSPF—individuals who are new

[9] The lists of erroneous matches are in the form of: lower-value pinsource/PIN, higher-value pinsource/PIN (one observation). If one PIN has a number of erroneously-linked PINs associated with it, we sort these observations by the lowest-value pinsource/PIN first, and then by the other associated higher-value pinsource/PINs. We maintain these lists from one version of the MSPF to another, and add observations as we discover erroneous links.

entrants to one of the administrative data sources or individuals from a new data source—then the task becomes more difficult, and due to the need for these kinds of additions, we have chosen to rebuild the MSPF on an annual basis. We do this rebuilding for several reasons. First, with new demographic information on individuals currently in our data core, or new observations from new data sources for some of these same individuals, we can take advantage of new opportunities for "chaining" or cross-linking individuals across data systems (described above), and, thus, improve upon our unduplication efforts.

Secondly, since we are still in the process of refining our unduplication and linkage programming, each successive version of our MSPF is improved. An older MSPF version may have a master ID mistakenly assigned to a match of two individuals (with the assumption that this is a single individual), whereas a new version of the MSPF may accurately record these as two different individuals. To attempt to update the older version of the MSPF with new information would require an untangling of the two identities, and the creation of a new ID for one of those identities. Equally difficult would be to collapse two MSPF IDs into a single identity (given new information linking these two observations) in an "update" of an already-built master file. For these reasons, it is very difficult to "simply update" the MSPF. More to the point, it would be very difficult for users of the Wisconsin Administrative Data Core to handle frequent shifting (collapsing or expanding) of individual identities, even if those shifts were minimal. It is, therefore, our current practice to recreate the entire Wisconsin Administrative Data Core on an annual basis.

Infrastructure: Staff and Computing Capability

Staff

An important consideration in beginning work on a match-merging of the universe of individuals in multiple State administrative data systems is the programming expertise available to work with the data from each data source. Few programmers can be expected to have expertise in all data systems. At IRP, we rely on a staff of programmers who work as a team, each contributing time, programming code, and data extracts for the administrative data sources within their area of expertise. Each programmer has knowledge of several administrative data

systems, and at least two programmers share expertise on each particular data source. Since these administrative data systems were not built for research purposes, they are often complex, and require much restructuring to meet the needs of social science researchers. We have found that it takes many months for individual programmers to become skilled users, and years to become experts, in each of these data systems. Therefore, we encourage longevity in the job to enhance and retain high levels of expertise. In other words, our programming staff members are not just masters of programming logic and technique, but have substantive knowledge of the structure, source and quality of information, meaning and history of coding schemes, as well as the administrative rules governing each of the unique, state-specific administrative data sources. Expertise is also developed in collaboration with our partners in State agencies. IRP programmers typically have close contact with colleagues in State agencies, may attend data systems-related staff meetings, and keep abreast of changes in administrative data systems.

IRP has additional staff members with experience in understanding the issues and concerns in the development of data-sharing agreements with State agencies, expertise in the development of agency research agendas and projects, and who independently keep abreast of data security issues, and monitor data security at all levels of data access, handling, and usage.

Computing Resources

IRP's computing resources come from its membership in the Social Science Computing Cooperative (SSCC) at the University of Wisconsin–Madison. The SSCC provides secure servers running Linux and Windows, managed by a professional support and system administration staff. Most of the MSPF work is completed on a secure Linux server: a restricted computing environment appropriate for data that is classified as "Protected Health Information" under HIPPA and other data with similar requirements. As of this writing, the server has a capacity of 12 terabytes (twelve thousand gigabytes) of project disk space. The building's gigabit Ethernet network is protected by a firewall, but allows data transfers to and from State administrative systems. SAS is the primary Linux-based software used to do the MSPF data handling and statistical work.

Process of Creating the Wisconsin Administrative Data Core

We have experimented with different ways to process data from the multiple data sources, including demographic data, group membership, and participation data. We have tried

various methods of organizing the unduplication of individuals internal to particular data systems, and of linking individuals across data systems. The following ordered steps outline our best current methods for efficiently organizing, unduplicating, linking, and structuring the various types and sources of administrative data currently available in the 2018 data core. In the first section, we discuss the order of programming steps necessary to build the MSPF or master file for the data core—a data file of one record per individual. This is the most difficult task in building the data core. The second section reviews other auxiliary files that we build subsequent, and linkable, to the MSPF.

Building the Multi-Sample Person File: Steps 1–20

This section outlines the step-by-step process that we use in building the MSPF master data file from multiple administrative data sources.

Step 1. We begin the MSPF-building process by gathering the raw data from various sources. Each data source is prepared for matching in no particular order.

Step 2. We extract information on individuals from one administrative data source, including all identifying variables that can be used for match-merging. Most administrative data systems provide demographic information on one or two tables, separated from participation data. For those data sources that combine demographic and participation data, we split off the demographic information and handle it separately. The participation data will be re-linked to this individual at a later step.

Step 3. We pre-process or clean these data by removing illegal characters and by adding standardized name fields to each record. Appendix B describes in detail some of the data cleaning issues we have encountered.

Step 4. We develop common formatting, coding schemes, and naming conventions so that variables from different data sources to be match-merged will be homogeneous, and, thus, linkable. For example, from our various data sources we found the following different coding schemes for indicating gender: 1 or 2, M or F, 0 or 1. We choose one of these schemes (numeric, if possible), and make all other data sources conform to this common format. We convert all date fields, such as birthdate and death date, to an SAS date format. We also retain all versions of an individual's birthdate, death date, and name, including all various spellings, and all name

changes, such as maiden name. We retain all versions of putative SSNs, and any SSN verification codes available in some data sources, which generally indicate the correct SSN out of multiple possibilities (although some individuals can legitimately have more than one SSN).

Step 5. Optional: Observations with no identifying information might be dropped at this point, as the possibility for match-merging with other data sources is precluded. However, this should be done with caution, since some individuals (such as unborn children or a deceased person's estate) may be part of a case or a family, and the record of this individual should perhaps be preserved.

Step 6. Optional: Split off certain fields into separate data sets for ease of handling (reducing the length of the record), or for protection of anonymity (not storing names, dates of birth, and SSN in the same data files). For example, the list of variables that we maintain separately are: multiple versions of name (including standardized names), parent–child linkages, and SSNs.

Step 7. We identify individuals within cases, and then make a record of parent and child pairs from within those cases. We also record potential twin pairs from within cases—children who have the same set of parents, the same dates of birth, and the same last names.

Step 8. Steps 2 through 7 are repeated for all additional administrative data files.

Step 9. Append the demographic data together from all sources, assigning a "pin source" code to each row. The pin source code has two purposes: (1) to identify the originating data source, and (2) to uniquely identify a row in combination with a PIN number. PINs are unique within a data source, but not across all sources. We maintain separate files for SSNs, names, and all other demographics.

Step 10. Link individuals with non-missing SSNs, birthdates, and names using a matching algorithm that will garner the highest quality true positive matches. Individuals for whom the last four digits of the SSN are non-missing are also included. Some fuzziness is allowed in the matching algorithm at this step, but there are no known false positive links in this first round of matching. The results of this match are pairs of pin source codes / PINs, along with a "match code," which is a numeric code that indicates which part of the algorithm created a match.

Step 11. All of the pin source / PINs found in the matched pairs from Step 11 are organized so that all pin source / PINs that are linked either directly or indirectly are clustered together and given a temporary ID. We compare the resulting file against a file of known erroneous links.[10] If any of these erroneous links are associated with the same temporary ID, then a manual check on the demographics related to the linked PINs is required. If the match is confirmed to be erroneous, then Step 11 is repeated with an adjustment to the matching algorithm that will leave out the erroneous link.

Step 12. Append together the parent-child files from all sources. Attach temporary IDs to the PINs in the parent-child data, if they were assigned in Step 12. Attach a pin source code to each row, as described in Step 10.

Step 13. Use parent-child data along with names and dates of birth to link individuals with missing or questionable SSNs. If the name and date of birth are a close or exact match, then a link might be made only with the first name of the mother. Matches with a higher degree of fuzziness or matches on individuals with very common names might require a match on a temporary ID of a parent or a match on both the name and DOB of a parent. This step matches children based on information about the parents, as well as mothers and fathers based on information about children and co-parents. Again, the results of this match are pairs of pin source codes / PINs, along with a match code.

Step 14. Match all the pin source / PINs found in the matched pairs from Step 14 with the PINs that were assigned temporary IDs in Step 12. If one pin from a matched pair was already assigned a temporary ID, then the other pin will be assigned the same temporary ID. If both PINs were assigned a different temporary ID, then the higher-numbered temporary ID is dropped and all pins associated with both temporary IDs will become associated with the lower-numbered temporary ID. For some match codes, the links between two different temporary IDs are ignored. This prevents some false-positive links. If neither PIN in the matched pair was previously assigned a temporary ID, then we assign one now.

[10] We maintain a list of known erroneous links that have occurred in the past, due to similar demographics between two distinct individuals, or due to errors in the raw data. These links have been confirmed to be erroneous through a manual review of related demographic and case data.

Step 15. Perform matches using cross-system PINs. (For example, use the CARES ID from KIDS, along with other demographics, to create links that have not already been made in previous steps.)

Step 16. Find links for people with missing SSNs that were not already matched in previous steps. In some cases, we use place of birth, sex, and race in the matching process. At this stage, we use names of parents, if available, to rule out some fuzzier matches if there are conflicts.

Step 17. Combine the matches from Steps 16 and 17, and then repeat Step 15 with these new matches. Any pin source / PINs not assigned a temporary ID up until this point are given one now.

Step 18. Before finalizing the MSPF, we create a set of parent-child files that help identify further matches of individuals between data sources—when a child appears to have two biological mothers or two biological fathers, or a parent appears to have two children with the same first name and the same date of birth. This final set of newly-identified matches requires resolution, and temporary IDs are reassigned accordingly.

Step 19. The temporary IDs are assigned a final, randomized number that is 10 digits long and begins with the last two digits of the version year of the MSPF.[11] This randomized number is called the IRPID.

Step 20. We then create a subset of variables from the full version of the MSPF for researcher use. Many of the uniquely identifying variables, such as SSN, name, and exact birthdate, are not included in the researcher file. Some details of other identifying variables are removed, smoothed, aggregated, collapsed, or masked, such as the "day" in the date of birth (which is removed), or in general codes for place of birth (which are aggregated and collapsed). These identifying variables or details of variables are not required by researchers, and are removed to preserve the anonymity of individuals in the data, preventing the possibility of deductive re-identification of individuals. See Sweeney (2000) for a clear discussion of this problem.

[11] We simply assign a random number as an identifier. We do not construct a unique identifier out of the components of identifying variables as described in Karmel et al. (2010) and Heil et al. (2007).

Additionally, when we have alternative versions of some variables available from different data sources, the version of that variable that we consider the most accurate (usually the value that appears most frequently across the different data sources) is the one provided to researchers.[12]

Creation of Associated Aggregation and Participation Files

Building the MSPF is the most critical and most time-consuming aspect of this project, and the MSPF is the centerpiece of the resulting data system. Once this phase of the project is completed, additional data files are created, linkable to the MSPF by the assigned individual identification number. We create three types of additional data files: (1) group aggregation files such as case and parent–child files; (2) program participation files, by specific time-period; and (3) cross-walk files, which provide the masked MSPF individual identifiers that match masked individual identifiers of other samples used by researchers, including prior year versions of the MSPF.[13]

Auxiliary group aggregation files are necessary, as it is often desirable to analyze the individual within the context of a case, or within the context of a family. Case identification is often necessary, as some data systems record information by case, and not by individual. In these situations, a file showing all the individuals in a case, along with their role in the case, and the beginning and ending dates of that role, are necessary for appropriate measurement of program use. Parent–child links have also been found useful for purposes of defining families in various ways (some examples: a mother and all her children, irrespective of father; a mother–father pair

[12] Different data systems have different strengths in terms of accuracy of specific data elements. For example, we would consider the data system charged with recording paternity adjudication to have more accurate information on the identity of the biological father, compared to other data systems. If two data systems are in conflict over the identity of a child's father, we choose the information from the most reliable source. Training and knowledge about the purpose and history of the data systems, and experience in handling the data are critical in making the programming judgments about such conflicting information.

[13] We do a cross-walk with the prior year MSPF (rather than updating prior year versions), since under- or over-matching of individuals is discovered with each new version, using additional information on individuals, and improvements in programming the match. In some instances where a single prior MSPF ID has been assigned to several observations (making the assumption that this is one individual), this assumption is later proved to have been incorrect, and in the new MSPF version this "one individual/one ID" now has two IDs. Alternatively, in some instances where two observations were given two MSPF IDs (making the assumption that these were two people), this assumption is later proved to have been incorrect, and in the new MSPF these "two individuals" now have one ID. For these reasons, it is very difficult simply to update prior versions of the MSPF. Our current policy is to recreate the entire MSPF, and to provide a cross-walk to those researchers who want to use both current and prior versions of the MSPF in their research.

and all their children; a mother, her children, and her children's children; a child and their full or half-siblings and all the fathers of those siblings). These case and parent-child group aggregation files are intermediate files that can be linked with the MSPF and program participation data for all members of a defined group.

Program participation files are built to indicate the eligibility or the benefits provided to individuals and/or cases, during specific time periods. These time periods are usually calendar months within a year, but can also be reported as quarters within a year (for example, UI wage records), or, in the case of rare events, reported by specific date (for example, CPS reports). The depth of detail provided in the participation files depends upon the general usefulness for research purposes. For example, monthly amounts of child support paid are provided in the participation data files, but detail on source of those payments is not provided. A researcher interested in this greater level of detail must specifically request additional data handling for subsets of samples or time periods, in order to gain access to that more detailed data.

We generally give the history of data detail provided in the participation files back to the earliest date of the data system. Since some administrative data systems loaded some aspects of historical information, it is sometimes possible to provide participation data that precede the start date of those electronic data systems. A list of the participation files available with the 2018 Wisconsin Administrative Data Core is shown in Appendix C.

Some Results of the Data Integration Process

In this section we present some figures showing the number of observations that have been added or linked together to form the 2018 MSPF, as well as some numbers comparing the resulting demographics of individuals in the 2018 MSPF with the 2018 U.S. Census Bureau's population estimates for Wisconsin. We linked or added observations from these data sources:

Administrative Data Source	**Number of PINs for Individuals**	**Number of PINs for "Spells"**
CARES	4,547,000	
KIDS	3,332,000	
eWiSACWIS	1,947,000	
DOC	135,000	
Milwaukee Jail		1,148,806
UI benefits	1,270,000	
CRD	129,000	
CRN	1,503,000	
DPI	1,979,000	

The 2018 MSPF totaled 7,209,000 presumed unique living individuals.[14] As the data sources for the individuals in the WADC are not point-in-time, this number is significantly higher than the total population of Wisconsin in 2018 (5,814,000). Some unknown number of individuals would have moved out of the state in the years since they were recorded in one or more of the administrative data sources (some dating back to the 1980s). And some individuals would have died, but there may be no record of this death in any of our data sources. Almost 14 percent of all individual observations in the 2018 MSPF have no recorded date of birth, and 4 percent have no recorded gender code. It is very likely that these observations are duplicates of individuals already within the data core, but we are unable to link them. Since most of our data sources involve only individuals who have had a period of low income, have been divorced with children, or have been children in a public school, sometime in the last 20 years, we do not intend or expect to have the full universe of Wisconsin residents in the data core. We recognize that we will never be able to fully link individuals for whom we have little identifying information, but we do expect to further improve our unduplication and matching results in future versions of the data core.

The following table shows the comparison of individuals in the 2018 MSPF with the 2018 U.S. Census figures for Wisconsin. We have subtracted from the MSPF column all individuals who are recorded as deceased, or who have no date of birth, when calculating the percentages on the table below. Compared to the population of the state as a whole, the 2018 MSPF contains a lower percentage of children and the elderly, and a higher proportion of adults aged 18 to 64.

	2018 Census Data[15]	2018 MSPF
Population under age 5	5.8%	3.6%
Population under age 18	22.0%	21.7%
Population aged 18–64	61.0%	64.7%
Population aged 65 and over	17.0%	13.6%
Population female	50.2%	49.3%

In the next table, we show the results of a cross-match between the IRP's court record data (CRD) on a sample of divorce and paternity cases from 1996 through 2013, and four other administrative data systems: CARES, eWiSACWIS, DOC, and UI benefits from the 2018

[14] We removed 515,000 individuals with a recorded death date from an earlier total of 7,724,000 MSPF individuals.
[15] From: https://www.census.gov/quickfacts/WI.

version of the MSPF. The results, comparing fathers and mothers from divorce and paternity adjudication cases, are as expected. Parents in divorce cases are much less disadvantaged, and fewer appear in the administrative data sources on public assistance, child welfare, incarceration, and unemployment. Mothers are generally more disadvantaged than fathers, as more of them appear in CARES and eWiSACWIS data.

CRD Parents	CARES*	eWiSACWIS*	DOC	Unemployment Benefits	None	All
Fathers:						
Divorce	47%	24%	5%	34%	37%	1%
Paternity adjud.	76%	54%	25%	42%	15%	7%
Mothers:						
Divorce	57%	33%	< 1%	29%	32%	< 1%
Paternity adjud.	89%	69%	3%	43%	7%	1%
*Without reference to adult/child role within the data system.						

In examining the final column of this table, which shows the percentage of CRD individuals that appear in all four of the administrative data systems, we see that a full 4 percent of paternity adjudication fathers appear in all of these cross-linked data sources, compared to one percent or less of mothers and divorced parents.

Major Changes to the Creation of the MSPF

As of August 2019, we have created 10 annual versions of the MSPF. We have learned many lessons as a result of building this system every year over the last decade. We have made significant changes to the process of creating the MSPF in order to improve efficiency and create a more sustainable process. We also have made several changes to decrease the number of false-positive matches and increase the number of true-positive matches.

Original Process of Creating the MSPF

This section gives a brief outline of the process that we used for building MSPF 2015 and prior versions of the MSPF:

Step 1. We began by focusing on one data source. After cleaning and extracting the demographic data from this data source, we unduplicated the data by matching it with itself and

reorganizing the demographic data to one row per person. We then gave each row a unique ID, creating a preliminary version of the MSPF.

Step 2. We selected a second data source to extract, clean, and unduplicate. Then we matched this second data source to the preliminary version of the MSPF (containing only one data source at this point). Any unmatched individuals from this second data source were given their own personal identifier, and once again all demographic data were placed into one row to create a new preliminary version of the MSPF.

Step 3. We matched a third extracted, cleaned, and unduplicated data source to the preliminary version of the MSPF (containing two sources of data). We ran separate programs to match the demographic data from the third data source to that of the first and second data sources to create a third preliminary version of the MSPF.

Step 4. We repeated this process for all other data sources. Each new data source required separate programming to match-merge with all of the data sources already folded into the MSPF.

Step 5. We used family ("parent-child") data from various sources to find even more matches. The final version of the MSPF was created, and data were combined to one row per person.

Why did we make major changes to the process of creating the MSPF?

This section explains our reasons for making major changes to the process of building the MSPF, and why those changes were worthwhile.

- We had 28 separate matching programs to incorporate seven different data sources into MSPF 2015. Seven of those programs were used to find internal matches within each data source, and the remaining 21 programs were used to find matching pairs between each combination of data sources. This was not a sustainable process for building the MSPF for a couple of reasons:
    - While there were some differences to the matching algorithm between those 28 programs, much of the logic was the same. Making improvements to the matching algorithm was an inefficient process, because changes needed to be applied to so

many different programs. Making improvements to some programs and not others would lead to inconsistency in the way in which individuals were put together across data systems.

  - Using our old process, each new data source would require the addition of several more matching programs (eight additional programs for an eighth data source, nine additional programs for a ninth data source, etc…), making the MSPF creation process significantly more time-consuming with each additional data source.

- Beginning with MSPF 2016, all data sources are matched together simultaneously. Specific improvements to the matching algorithm require changes to only one program, making the process significantly more efficient. When we added DPI data to MSPF 2018, we didn't need to create new programming to match the individuals from DPI to other data sources. The existing programming to build the MSPF can be applied to match data from any new source, as long as the data are cleaned and standardized prior to matching.
- In MSPF 2015, each data source was unduplicated prior to matching with other data sources, and that unduplicated version of the data was used for matching. One positive aspect of using this method was that demographic data related to one person on more than one row was combined, setting things up for a better match, assuming that the unduplication was a result of true-positive matching internal to the data source. However, if the unduplication was the result of a false-positive match, then it was not possible to separate the information from the erroneous match, even upon learning new information from other data systems.
- Beginning with MSPF 2016, higher quality matches are generally performed before lesser quality matches regardless of data source. We are now less likely to miss high-quality matches or favor a low-quality match over a high-quality one.

USE OF THE WISCONSIN ADMINISTRATIVE DATA CORE

For all researchers who use the Wisconsin Administrative Data Core, we advise caution in relation to several limitations. Since none of the administrative data that we link was collected for research purposes, but rather, was collected for purposes of administrative use and case management, data sources differ in areas of data quality and completeness. In situations where

individuals are linked to multiple data sources, the resulting demographic information may be of greater completeness and accuracy, compared to individuals with no link to additional data sources. We also have the issue of historical extent of the data: we have differing lengths of database history; and when individuals leave the state, or are no longer present in one or another data system, then some information on individuals becomes dated, or is not captured (such as dates of death).

Some data sources may have undetected or unexpected gaps. For example, we extract SSI benefit data for all individuals receiving public assistance, and the assumption is often made that this is the full population of SSI beneficiaries. Although we include a very high proportion of SSI beneficiaries, this is not the full population, and the benefit information we use is not from the administrative source of those benefits, but rather a reporting of those benefits for eligibility purposes in other public assistance programs. Another example is that although we may assume our data contain all individuals eligible for MA benefits, one small subgroup that is not included in these administrative records comprises children in foster care.

Another issue for researchers is that some data sources are more concerned with attempting to maintain only one record and one identifier per individual. Other data sources do not have that goal, and therefore may be much more prone to problems of internal unduplication in our matching and linkage work.

In general, researchers must themselves become knowledgeable about the sources of data, the reasons for data collection, the differing target populations included, the history of available data, and the history of administrative changes over time, in each of these administrative programs. Kohler and Thomsen (2009) present a good discussion of these types of data considerations when using administrative data for social science research.

Another set of issues for research when using the WADC has to do with appropriate topics for research. This data system was built with State of Wisconsin administrative data, to support research that has the potential to inform the evaluation and administration of public policies. All research projects using this wealth of data therefore must have core policy issues and questions as a basis for the research, and the data may not be used for purely academic projects or without consultation with State partners. All researchers expecting to use the WADC

must receive permission from the State of Wisconsin agencies that provide this data, by submitting a proposal to the IRP Data Access Committee.

We do not present a review of completed, ongoing, and planned research using the WADC here, but a few recent examples of projects supported by the data core and the integrated data project include:

- *National Child Support Noncustodial Parent Employment Demonstration (CSPED)*: A randomized controlled trial combining administrative data from eight states to test the effectiveness of child support-led employment programs for noncustodial parents (Cancian, Meyer, & Wood, 2019).
- *Educational outcomes for children in foster care*: A project resulting from an agency request for technical assistance that resulted in the first opportunity to link child welfare and educational data in Wisconsin, and suggested that permanency alone is insufficient to promote foster youths' educational and economic attainment (Font, Berger, Cancian, & Noyes, 2018).
- *Understanding the interaction between child welfare and child support collection*: A project resulting from a WI Department of Children & Families staff question that illuminated unintended negative consequences of policy requiring parents to pay child support to offset the costs of their children's stay in foster care (Cancian, Cook, Seki, & Wimer, 2012).
- *Multiple program participation of TANF and other program participants:* A project funded by the U.S. Department of Health and Human Services' (UHHS) Administration for Children and Families (ACF) in which the authors measure multiple means-tested program participation and disconnection among TANF participants (Cancian, Han, & Noyes, 2011).

FUTURE PLANS FOR THE WI ADMINISTRATIVE DATA CORE

Future plans for new or updated versions of the WADC depend upon the utility of the data core to researchers, funding to pay for computing resources and staff time, the continued availability of current data sources, and the inclusion of additional data sources that would enhance the research potential for researchers.

Since we're still improving the matching algorithm and other programming for the WADC, and since we have identified several new data sources that we hope to incorporate into the system, we plan during this time to continue our policy of recreating the data core, rather than simply updating it, on an annual basis.

Additional data sources that we are considering adding to the data core fall into several categories. For example, we want to develop a greater historical depth to the information we have already linked, and fill in some gaps that we have identified in terms of data quality or data coverage. We also want to broaden our ability to provide information of interest to our State and Federal partners and IRP researchers, particularly in the area of child well-being. We present some data sources that we hope to add in future versions of the data core in Appendix D.

The WADC builds on a long history of collaboration between the State of Wisconsin and the University of Wisconsin–Madison's Institute for Research on Poverty (IRP). It reflects a history of investments in developing technical expertise in data system design and data security, as well as a nuanced understanding of the administrative data systems and the programs they are designed to support. The data core serves as a key resource for both research and program evaluation and administration (see Cancian, Noyes, & Han, 2011). Maintaining and updating the data core will require sustained investment and continued collaboration between IRP staff and affiliates, and our partners at State and Federal agencies.

References

Campbell, Kevin M. (2005). Rule your data with the Link King (a SAS/AF application for record linkage and unduplication). SUGI 30.

Cancian, Maria, Han, Eunhee, & Noyes, Jennifer. (2011, May). Patterns of and outcomes associated with disconnection from employment and public assistance: The Wisconsin experience. Report to the U.S. Department of Health and Human Services, Administration for Children and Families. Madison, WI: Institute for Research on Poverty, University of Wisconsin–Madison. Available at https://www.irp.wisc.edu/wp/wp-content/uploads/2018/06/Task13A_CS_09-11_Final.pdf.

Cancian, Maria, Meyer, Daniel R., & Robert G. Wood. (2019, March). Final impact findings from the Child Support Noncustodial Parent Employment Demonstration (CSPED). Madison, WI: Institute for Research on Poverty, University of Wisconsin–Madison. Available at https://www.irp.wisc.edu/wp/wp-content/uploads/2019/07/CSPED-Final-Impact-Report-2019-Compliant.pdf.

Cancian, Maria, Noyes, Jennifer, & Han, Eunhee. (2011, December). Using administrative data for program evaluation and administration: Progress report from Wisconsin. Report prepared for U.S. Department of Health and Human Services, Administration for Children and Families. Madison, WI: Institute for Research on Poverty, University of Wisconsin–Madison.

Cancian, Maria, Cook, Steven, Seki, Mai, & Wimer, Lynn. (2012). Interactions of the child support and child welfare systems: Child support referral for families served by the child welfare system. Report to the Wisconsin Department of Children and Families. Madison, WI: Institute for Research on Poverty, University of Wisconsin–Madison. Available at https://www.irp.wisc.edu/wp/wp-content/uploads/2018/06/Task13A_CS_09-11_Final.pdf.

Chalabi, Mona, & Flowers, Andrew (2014). "Dear Mona, what's the most common name in America?" Available at https://fivethirtyeight.com/features/whats-the-most-common-name-in-america/.

Font, Sarah, Berger, Lawrence, Cancian, Maria, & Noyes, Jennifer. (2018). Permanency and the educational and economic attainment of former foster children in early adulthood. *American Sociological Review*, *83(4)*, 716–743.

Goerge, Robert M., & Lee, Bong Joo. (2002). Matching and cleaning administrative data. In Michele Ver Ploeg, Robert A. Moffitt, and Constance F. Citro (Eds.), *Studies of welfare populations: Data collection and research issues*. Washington, D.C.: National Academy Press.

Gomatam, Shanti, Carter, Randy, Ariet, Mario, & Mitchell, Glenn. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine, 21*, 1485–1496.

Grannis, Shaun J., Overhage, J. Marc, & McDonald, Clement J. (2002). Analysis of identifier performance using a deterministic linkage algorithm. American Medical Informatics Association Annual Symposium Proceedings, 305–309.

Heil, Susan, Leeper, Tracay, Nalty, Dennis, & Campbell, Kevin. (2007). Integrating state administrative records to manage substance abuse treatment system performance (DHHS Publication No. [SMA] 07-4268). Technical Assistance Publication (TAP) Series 29. Rockville, MD: Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration. Available at https://www.air.org/sites/default/files/downloads/report/TAP29_06-07_0.pdf.

Herzog, Thomas N., Scheuren, Fritz J., & Winkler, William E. (2007). Data quality and record linkage techniques. New York, NY: Springer Science and Business Media.

Karmel, Rosemary, Anderson, Phil, Gibson, Diane, Peut, Ann, Duckett, Stephen, & Wells, Yvonne. (2010). Empirical aspects of record linkage across multiple data sets using statistical linkage keys: The experience of the PIAC cohort study. *BMC Health Services Research, 10*, 41.

Kohler, Markus, & Thomsen, Ulrich. (2009). Data integration and consolidation of administrative data from various sources: The case of Germans' employment histories. *Historical Social Research, 34*(3), 215–229.

Sweeney, Latanya. (2000). Simple demographics often identify people uniquely. Data Privacy Working Paper 3. Pittsburgh, PA: Carnegie Mellon University.

Appendix A

State of Wisconsin Administrative Data Files Used in the Wisconsin Administrative Data Core

CRN data system. The CRN (acronym for Computer Reporting Network) data system began in 1974, and was designed to record Aid to Families with Dependent Children (AFDC) payments, Food Stamp benefits, and Medicaid eligibility. CRN was phased out in 1994 and 1995, and replaced by the CARES data system. CRN includes demographics of participants, and their relationship to the household head. IRP has access to the CRN data for the month of December 1984, and for most months from August 1988 through the phase-out in early 1995.

CARES data system. The CARES (acronym for Client Assistance for Reemployment and Economic Support) data system began in 1994, designed to replace the CRN data system. Counties in Wisconsin began using the CARES system in different months over the course of about 12 months in 1994 and 1995. The CARES data system records the following: payments of AFDC and later of TANF and Food Stamp (now SNAP) benefits; Medicaid and BadgerCare eligibility; Child Care subsidies to parents and payments to providers; cash Caretaker Supplements for disabled parents with minor children; demographics of participants; a history of address and income changes; and a history of household members and their relationships to each other.

KIDS data system. The KIDS (acronym for Kids Information Data System) data system began in 1995, and all counties were fully online by the last quarter of 1996. The KIDS data system is a financial accounting system for the payment and disbursement of child support, and is also used for recording the details of paternity adjudication. A history of child support orders and the balance of arrearages are kept for each case, as well as the participants in the case (the mother, the legal father, the child, the payor, and the payee of child support). A historical record of address change is maintained, as well as a record of child support case type and some information on marital status and demographics. Information from cases dated prior to 1995 was converted to KIDS, although much of this information is missing, or is recorded as of the conversion date rather than specific case history event date.

eWiSACWIS data system. The eWiSACWIS (acronym for State Automated Child Welfare Information System) data system began in 2001, but was not fully implemented by all Wisconsin counties until mid-2004. Counties had different policies about loading inactive or conversion cases into the system, and much of the conversion data has missing information. This data system records all Child Protective Services (CPS) reports; all out-of-home placements of children; and the participants in all cases (the children, parents, caretakers, maltreators, and reporters of maltreatment); along with individual demographics; and the relationships of family members to the case reference person.

Department of Corrections (DOC) data system. The DOC data system began in 1990, and is a record of all individuals incarcerated in the State prison system at that time, and since. It

records the incarceration history, reasons for incarceration, and demographics of incarcerated individuals.

Milwaukee County Jail data system. The Milwaukee Jail data system began in 1993, and is a record of all individuals incarcerated in the Milwaukee County Jail or the Milwaukee County Correctional Facility-South (previously known as the House of Corrections). It records the incarceration history, reasons for incarceration, and demographics of incarcerated individuals.

Court Record Database (CRD). The court record database is not a data system maintained by the State of Wisconsin, but rather is a sample of child support-related court record information gathered from paper records and recorded electronically and maintained by the Institute for Research on Poverty for most years since 1980. This database contains information on parents and children in divorce and nonmarital cases that have come to court for purposes of divorce, paternity establishment, child support, and child custody. Two to seven years of data have been gathered and recorded for each case. This is a sample of cases, from a sample of Wisconsin counties.

Unemployment Insurance data (UI). The Unemployment Insurance program collects and maintains a history of wage records from employers in Wisconsin for the purpose of providing unemployment benefits to unemployed workers. This history has been maintained electronically for over two decades. Approximately 95 percent of all legally employed workers in Wisconsin are recorded in the UI data system. There are two data sources: wage records and unemployment benefits.

Wage Records. The wage records are reported quarterly by mandated employers. Wage record data is available back to approximately 1988. The full wage record data file is not included in the IRP Data Core, but rather, only wage records that match (on SSN and name, if available) the individuals in the data core.

Unemployment Benefits. We also have access to unemployment benefit data since 2006. This is in the form of weekly cash benefits paid to unemployed workers who continue to report a search for employment. We have access to information on weekly cash benefits, by date of dispersal, and for dates of the covered time period of unemployment. Our current data extend back only to the fourth quarter of 2006, although data prior to this time do exist, and may be requested in the future.

Department of Public Instruction data (DPI). The DPI maintains statewide records for students who attended public K-12 schools. We have access to data beginning with the 2005–2006 school year. We began incorporating this data into the 2018 version of the MSPF.

Appendix B

Some Data-Cleaning Issues for Specific Variables

A. Dates of birth.

Missing information on birth. When birthdates are missing it is often the case that fields are filled in with missing data codes that appear as if they might be actual birthdates. We have found that a variety of missing data codes have been used in a number of our data sources. We know of no system-wide policies, however, that indicate what a missing data code should be, and we have evidence that counties develop their own policies, which may differ from other counties.

Without specifically stated policies as to what codes are used for missing data, we have to search for missing data codes by doing frequencies on the date of birth. There are some specific dates that jump out on a frequency list, as these dates have an extremely large number of observations, compared to dates before and after these dates. An example:

Date of Birth	Frequency
01JAN1900	197
02JAN1900	9
03JAN1900	13
04JAN1900	4
05JAN1900	8
06JAN1900	4
07JAN1900	1

It seems clear from this example that 1/1/1900 is used as a missing code for date of birth, as we would normally expect only about 10 actual births (of individuals from this particular data source) to have occurred on that date, given the frequency of births in the following week of that year. In this situation, we feel free to consider birthdates of 1/1/1900 as missing data codes, realizing that a very small handful may not be missing data, but are actually correct dates. We have found through an examination of the frequencies of dates that the following dates are commonly used as missing data codes in our Wisconsin data sources: 1/1/1900, 2/2/1922, 2/22/1922, 12/22/1922, and 1/1/1950. January 1st of later years also appear to be a record of births in that year for which the month and day are unknown (in many years, about three times more individuals were coded as having birthdates on the first day of the year than would be expected).

In other situations, we can also assume that certain dates are used as missing data codes, but the solution of changing these dates to indicate missing data is not quite so clear. For example:

Date of Birth	Frequency
10JAN1950	190
11JAN1950	48
12JAN1950	47
13JAN1950	63
14JAN1950	61
15JAN1950	57
16JAN1950	42

The date of 1/10/1950 is undoubtedly being used as a missing data code, but if we assume that all birthdates of 1/10/1950 should be considered missing data codes, then we are erroneously classifying as "missing" approximately 50 birthdates that probably are correct, given the frequencies of birthdates on days following January 10th. Also, it may be possible that 1/10/1950 is used as a missing data code for someone who is known to have been born in 1950, but the month and day of birth is unknown. If we set 1/10/1950 to missing, we would then lose the information on the possible year of birth for these individuals.

We also would want to consider that a person possibly born in 1950 is more likely to still be in our data as a person of interest, given our research time period, whereas a person born in 1900 or even 1922 is not likely to be an active participant in our data, and therefore the danger of changing a correct birthdate to missing for these earlier born individuals poses less of a problem.

For the WADC, we have developed the rule that the frequency of birthdates on a given date must be more than five times the frequency that might be expected for that date, given the birthdates in the week following the date in question. However, we ignore this rule when the date of birth was recorded after January 1, 1960, as we believe that the recorded year may give us information on at least the year of birth. Using these rules on our public assistance database (CARES), out of over 4,500,000 individuals, we recoded as "missing" about 70,000 dates of birth, or about 1.6%. This includes over 8,000 (about .18%) instances where the birth date was recorded as prior to 1900. We have recoded all of these dates as missing, as we believe that erroneous birthdates are more misleading than missing birthdates, and lead to fewer ultimately correct matches of individual records.

B. Dates of death.

We have found that missing information for death date often is in the **day** of death, but not the month or year. As such, many dates of death are given as the first day of a given month and year. We preserve this death date, therefore, as we consider it partially accurate, and when matching individuals on the death date, we consider a match on month and year only. We also check for death dates that are recorded as preceding birthdates. A handful of these appears, and are manually checked and corrected.

C. SSNs.

Data cleaning of Social Security numbers involves a search for illegal numbers, and the use of certain numbers as missing data codes.

Searching for illegal numbers. No valid Social Security numbers can have all zeroes in any of the three SSN fields. Valid SSNs (assigned prior to June 2011) also cannot exceed 799-99-9999. Some individuals report SSNs above 799-99-9999, but these may be International Identification Numbers (ITINS), issued for noncitizens for purposes of employment. Many data systems record ITINS in the SSN field. There are also "pseudo-SSNs" issued to individuals in some data systems that are in the 800-00-0000 series, for individuals with no SSN. This was more prevalent prior to the 1980s, when young children and non-working wives often did not have an assigned individual SSN, and was used as a way to track individuals within the data system when there was no available SSN. These numbers were also recorded in the SSN field. We remove these numbers out of the SSN field, as they cannot be used for wage record matching, based on SSNs, or on matching to other data systems.

Some SSN assignment peculiarities.

The rules for SSN assignment have changed over time, with major changes taking place in June 2011. See: http://www.socialsecurity.gov/employer/randomization.html, and http://www.ssa.gov/kc/SSAFactSheet--IssuingSSNs.pdf.

For children in cases from our source data systems born before the early 1980s,[16] and for children who are twins, it is very likely that SSNs will be similar, since applications for SSNs for the children in the family were generally applied for at the same time. This similarity in SSNs, when accompanied by the same set of parents and the same last name, can lead to the problem of false positive matches.

We have also found that the assignment of SSNs for children born in Wisconsin tended to be done in alphabetical order by child's first name for SSN applications from November 2005 through February, 2007. This nonrandom assignment of SSNs can lead to false positive matches, due similarity of SSNs, first names, and dates of birth, for individuals with SSN applications within this time frame.[17]

We use a "spelling distance" function in SAS (SPEDIS) to find SSNs that are similar, and that therefore might indicate typos in data-entry, and a positive match. However, the SPEDIS function assigns higher "penalties" for differences between digits at the beginning of the SSN, compared to the last four digits of the SSN. Since SSNs were assigned geographically until 2011, with Wisconsin SSN-applicants being assigned numbers in the range of 387–399, we have found that reversing the SSN prior to using the SPEDIS function helps to identify typos more accurately, compared to making false positive matches based on similar SSNs.

[16] In the mid-1980s, children's SSNs began to be required for federal tax reporting, and many parents applied for SSNs for all of their children simultaneously.

[17] It is unknown whether this nonrandom assignment was done nationally or regionally.

D. Names.

Matching on names is critical, but the most difficult match-merging task. One problem is in the wide use of nicknames; another is the problem of misspellings or typos; yet another is the need to parse names into appropriate component parts (remove "Dr.," "Mr.," etc., from the first name, remove suffixes such as "Jr." and "Sr." from last names).

One solution for misspelled names or nicknames is to standardize the names, so that names such as "Pat," "Patricia," "Patricai," "Patti," and "Patty," are reset to a common version, such as "Patricia." We maintained a list of name-standardization pairs in our early work on the MSPF.

But then there comes the complication of what to do with "Patrick," "Patrice," and "Tricia," which may or may not be variations on the name "Patricia." Another solution is to access a name-matching table that can look something like this:

Pat	Patricia
Patti	Patricia
Patty	Patricia
Patrice	Patricia
Tricia	Patricia
Pat	Patrick
Pat	Patrice
Patti	Patrice
Patty	Patrice
Tricia	Trish

This method is slightly different from "name standardization," in that names are not actually assigned a standardized version, but are being matched in ways known to be common (such as, Pat = Patrick, or Pat = Patricia). We maintain a table of this type for our current data core work, and use this method for matching variations of nickname, legal name, and possible misspellings.

We also transpose first and last names, and check for first-to-last, and last-to-first name matches, as well as first-to-middle and middle-to-first name matches, as middle names are often used as nicknames. We also search for name matches that include embedded versions of the name (such as "John" = "JJohn," or "Jones" = "SmithJones").

We use SAS software in our data handling, and SAS provides the capability of checking for similarly sounding names (through a function called SOUNDEX), and for slightly misspelled names (through a function called SPEDIS, meaning "spelling distance").

Hyphenated names present additional problems, and usually are found in last names. We maintain the original spelling of the name, but also attempt matches on variations of the name, such as separating the two parts of the hyphenated last name, or eliminating the hyphen, such as:

Brown-Medcalf	Brown
BrownMedcalf	Brown
Brown	Brown
Medcalf	Brown
Brown-Medcalf	Medcalf
BrownMedcalf	Medcalf
Brown	Medcalf
Medcalf	Medcalf

After applying standardization procedures, we then calculate degrees of commonness to both the first and last names. Matches on the name "Pat Brown" would be considered somewhat suspect, since both the names "Pat" and "Brown" are common. But a match on the name "Pat Medcalf" or "Zetty Brown" would be considered much more likely, due to the uncommonness of one of the names. A name match on "Zetty Medcalf" might require no other corroborating matching identifiers, as the uncommonness of both first and last names allows us to be fairly sure of the certainty of the match.

We calculate degrees of commonness/rarity, based on frequencies in our prior year MSPF, for all first names, surnames, and first and surnames together. We began calculating the commonality of first and surnames together for MSPF2016. When we look at the commonality of first and surnames separately, it is necessary to calculate the degree of commonality for Hispanic and Asian subpopulations. When using the commonality of the full name, it is not necessary to calculate name commonality within ethnic groups (see Chalabi & Flowers, 2014).

E. **Place name and birthplace coding schemes.**

We standardize birthplace names into county codes, state codes, and country codes, often having first to parse birthplace text fields into city name, county name, state name, country name.

1) **Country codes.** We use a 2-digit alphabetic coding scheme from the Wisconsin CARES data system to standardize place of birth country names. This coding scheme is similar but not identical to a 2-digit alphabetic coding scheme maintained by the International Organization for Standardization (ISO).[18]

2) **State codes.** We use the 2-digit alphabetic U.S. postal codes to standardize state names. Much of our birthplace data include these codes, but we have found examples of caseworkers (incorrectly) guessing at the assignment of codes, such that Arizona becomes "AR" (the code for Arkansas) instead of the correct "AZ." More seriously, for many Minnesota-born Wisconsin residents is the miscoding of a Minnesota birthplace as "MS" (Mississippi), instead of the correct "MN." Alternatively, a code of "IN" can mistakenly be assigned by a case worker to someone born in India, which is then understood to mean "Indiana."

[18] http://www.iso.org/iso/country_codes/iso-3166-1_decoding_table.htm

3) **County codes.** We use the 2-digit numeric system adopted by most State of Wisconsin agencies, numbering the counties from 1 to 72, with tribal units given higher numbers. Wisconsin has the unfortunate situation, however, of having two different county coding schemes, brought about by the addition of Menominee County in the 1950s as the 72nd county. The original county coding system was often "fixed" by simply placing Menominee County at the end of the county list (county = 72). The more standard current scheme inserts Menominee County alphabetically into the county order. This means that all counties that appear after Menominee alphabetically have different county numbers in the two coding schemes. These are the kinds of mundane, locally specific coding, cleaning, and standardizing issues that programming staff members have to be aware of, and take the time to standardize.

We often find inconsistencies in place name and codes and much work is spent resolving many of these inconsistencies, particularly for individuals with little other identifying information.

F. Twins/multiple births.

Twins present special problems when trying to match on partial identifying information. The last names of twins, as well as the dates of birth, birthplaces, and identifiers of mothers and fathers are identical. Moreover, the first names are often similar ("Patricia" and "Patrick"). SSNs are also sometimes close in actual number (since requested at the same time). PINs in various data systems are also often similar, as family members are sometimes given consecutive numbers. For these reasons, we often mistakenly confuse twins as a match—indicating a single individual.

Sometimes data systems will indicate in text fields that specific individuals are twins, and we have also found that some data systems have a special algorithm in assigning PIN numbers to twins/triplets, which allows us to identify twins within the household or family. In addition, we scour case information to identify children with the same birthdates, last names, and parents, but with differing first names, to identify twins. We record all of these hints of possible twin status and use it when match-merging. For individuals with a twin indicator, we require a higher degree of first name and SSN matching to prevent false positive matches.

Appendix C
Documentation of Participation Files

List of participation data files available to researchers with the 2018 MSPF:
-- generally monthly
-- quarterly, for UI wage record
-- per event date, for child protective service reports and substantiations

IRPID = IRP-constructed unique personal identifier of individuals
IRPcaseID = IRP-constructed unique identifier of cases

File names and content:

W21997 – W22018 by IRPID and IRPcaseID
TANF cash benefit and case type code

AFDC1984, AFDC1988 – AFDC1998 by IRPID and IRPcaseID
AFDC cash benefit and case type code

FS1984, FS1988 – FS2018 by IRPID and IRPcaseID
FoodStamp/FoodShare/SNAP dollar amount, *N* members, indicator of individuals included

CC1997 – CC2018 by IRPID and IRPcaseID
Dollar amount of child care subsidy, dollar amount of co-pay owed, *N* children, and indicator of child included

CCEBT_CSCHLD_PDTO_PROVLOC by IRPIDchild and IRPcaseID
Dollar amount case/family paid per child at a provider-location, starting October 2016

CCEBT_PROVLOC_RECVD_PERTYPE by IRPIDchild and IRPcaseID
Total transaction amount, credit/debit, per provider-location, starting October 2016

MED1988 – MED2018 by IRPID and IRPcaseID
Indicator of MA eligibility, number of covered members, type of medical assistance

CTS1997 – CTS2018 by IRPID and IRPcaseID
Caretaker Supplement cash benefit

HSUB by IRPcaseID
Housing subsidy data from CARES, by date, per report

SHELcostCARES by IRPID
Shelter costs with indicator for homeownership vs. renting, by date, per report

UIwage1998 – UIwage2018 by IRPID

UI wage record of dollars earned, by quarter

AvailSSN
Indicator of SSN availability for all UI wage data

UIbenefits2006 – UIbenefits2018 by IRPID
UI unemployment benefits (dollar amount), by quarter, starting in 4th quarter 2006

UIunemplspells2006 – UIunemplspells2018 by IRPID
UI unemployment spells, by week, starting in 4th quarter 2006

SSICARES1995 – 2018 by IRPID
SSI/SSDI/SS dollar amount, as recorded in CARES, by calendar month

DOC1990 – DOC2018 by IRPID
Indicator of full or partial months of incarceration in Wisconsin State Prison System

MILWJAIL1993 – MILWJAIL2018 by IRPID
Indicator of full or partial months of incarceration in Milwaukee County Jail

CSord1997 – CSord2018 by IRPcaseID, IRPIDpayor
Child and family support orders, and alimony (maintenance) orders (dollar amount), by month

CSord_api1996 – Csord_api2018 by IRPcaseID, subaccount, IRPIDpayor
Child or family support and alimony (maintenance) orders on arrears, past support, and interest accounts, by month and sub-account

CSordHO by IRPcaseID
Held-open child support orders

CSpay1997 – CSpay2018 by IRPcaseID, IRPIDpayor
Total dollar amount of all payments made to the child support system, by case, by payor, by month

CSrec1997 – CSrec2018 by IRPcaseID, IRPIDpayee
Total dollar amount of all receipts through the child support system, by case, by payee, by month

CSrecdetail1996-CSrecdetail2018 by IRPcaseID, IRPIDpayee
Total of all receipts through the child support system, by month, including receipts to the State and for ALL subaccounts

CSenforceevts by IRPcaseID
History of all child support enforcement events

KIDScasetyp by IRPcaseID
Indicator of IV-D status, begin and end dates

Arrears1996 – Arrears2018 by IRPcaseID, IRPIDpayee
Dollar amount of child support and related arrearages as of the end of the calendar year, by case, by payee

SACscreened by IRPID
All screened-in CPS reports, by victim (child), per eWiSACWIS case, per referral period, per maltreator type (parent or non-parent) (2004–2018)

SACsubst by IRPID
All substantiated CPS allegations, by victim (child), per eWiSACWIS case, per referral period, per maltreator (2004–2018)

SACsvcintks by IRPcaseID
All service intake calls, by service intake event (2004–2018)

SACoutofhome by IRPID
Indicator of child removal from home and placement out-of-home, by child, per calendar month (1990–2018)

SACplacements1985_2018 by IRPID
Details of out-of-home placement, by event

SACvolkcar1997_2018 by IRPID
Voluntary Kinship Care, per calendar month

SACvolcarlines1997_2018 by IRPID
Details of voluntary kinship care, by event

Appendix D

New Administrative Data Sources Planned or Under Development

County incarceration data from counties other than Milwaukee County. County jail data are difficult to include in the WADC, as there is no statewide electronic database serving all Wisconsin counties. We currently include the Milwaukee County Jail data, as this is Wisconsin's largest county. In the future, we hope to include other large urban county jail databases.

Unemployment Insurance Data (UI). The Unemployment Insurance program collects and maintains a history of wage records from employers in Wisconsin for the purpose of providing unemployment benefits to unemployed workers. We have thus far included unemployment benefit data back to 2006. This is in the form of weekly cash benefits paid to unemployed workers who continue to search for employment. We have access to information on the dollar amount of weekly cash benefits, by date of dispersal, and dates of the covered time period of unemployment. We hope to extend our coverage of unemployment benefit data and unemployment spells prior to 2006 in future versions of the MSPF data system.

Department of Motor Vehicles (DMV). We are considering requesting access to information from the Department of Motor Vehicles on driver's licenses and vehicle registration. One reason for our interest in these data is the high quality of demographic information available from the DMV. Some of our researchers are also interested in measuring transportation options for purposes of employment for low-income individuals.

Vital Records. Vital records of birth, death, and marriage are important sources of high-quality data. We have been given access to samples of birth records for limited research purposes, but have not been given access to these data for linkage with other administrative data systems, or for more open-ended research questions. We have not yet requested access to marriage or death records, but vital records remain as possible future sources of high-quality data, with good definitions of important populations.

Medicaid Claims and Encounter. We have had records of Medicaid enrollment as part of the WADC since its beginning, but have recently signed a data sharing agreement with the Wisconsin Department of Health Services also to include Medicaid claims and encounter data in future versions of the MSPF and WADC.

Housing Data. We are currently pursuing funding to increase the amount of housing information in the WADC, with the intent of adding information from Public Housing Authorities (PHAs), the State's Homeless Management Information System (HMIS), and a partnership with The Eviction Lab at Princeton University.