# Part II

# Administrative Data

# 7

# Matching and Cleaning Administrative Data

*Robert M. Goerge and Bong Joo Lee*

This paper addresses the cleaning and linking of individual-level administrative data for the purposes of social program research and evaluation. We will define administrative data as that collected in the course of programmatic activities for the purposes of client-level tracking, service provision, or decision making—essentially, nonresearch activities. Although some data sets are collected with both programmatic and research activities in mind—birth certificates are a good example—researchers usually think of administrative data as a secondary data source in contrast to surveys that are conducted solely for research purposes.

When we refer to administrative data to be used for research and evaluation of social programs, we are referring primarily to data from management information systems designed to assist in the administration of participant benefits, including, income maintenance, food stamps, Medicaid, nutritional programs, child support, child protective services, childcare subsidies, Social Security programs, and an array of social services and public health programs. Because the focus of the research is often on individual well-being, most government social programs aimed at individuals could be included.

Before these administrative data can be used for research purposes, cleaning the data is a major activity as it is in conducting and using any large social surveys. Cleaning is necessary because there are numerous sources of potential error in the data and because the data are not formatted in a way that is easily analyzed by social scientists. We take a broad view of cleaning. It is not just correcting "dirty" data; it is producing a clean data set from a messy assortment of data sets. In this paper, cleaning refers to the entire process of transforming the data as it exists in the information system into an analytic data set.

Record linkage is a major activity in the use of administrative data, especially if the research is longitudinal—and, by definition, evaluation nearly always is. Record linkage is the process of determining that two data records belong to the same individual. Being able to track an individual from one time to the next or across numerous data sets is nearly always necessary when using administrative data, especially because, in most cases, one does not have access to the independent sources of data that can assure that a time 1 measurement and time 2 measurement are of the same person or that an agency 1 record and agency 2 record are for the same individual.[1]

Data cleaning and record linkage are closely related activities. At its most simple level, record linkage is necessary to determine if any duplicate records exist for a single individual or case in a particular data set. Record linkage is used to produce clean, comprehensive data sets from single program data sets. Without accurate record linkage, it is likely that the data on an individual will be incomplete or contain data that do not belong to that individual. And, to do accurate record linkage, the data fields necessary to perform the linkage, typically individual identifiers, must be accurate and in a standardized format across all the data sets to be linked.

## ADVANTAGES AND DISADVANTAGES OF ADMINISTRATIVE DATA

The advantages and disadvantages of administrative data can be identified most easily when they are compared with survey data. However, the comparison of these two types of data collection is a straw man. The research questions that are appropriately addressed are qualitatively different from those appropriately addressed by surveying a subset of the general population. The type of data that one should use for answering a particular research question should be determined by the question. If a comprehensive study of a particular issue requires a survey of a population not covered by administrative data or if important variables are unavailable, other data collection is necessary. However, it is almost always the case that a rich study of a particular issue that identifies or rules out multiple potential causes or correlates of a particular phenomena requires data from multiple sources.

Administrative data, in most cases, are superior to other data sources for identifying program participation—what benefits were provided to whom, when, and in what amount. (The exact reason why people are participating is often missing.) Administrative data are collected on an entire population of individuals

[1] Because of problems with recall, often the individual cannot confirm his or her participation at a particular point in time (see Kalil et al., 1998).

or families participating in a given program. This is advantageous for two reasons. First, it is possible to study low-incidence phenomena that may be expensive to uncover in a survey of the general population. Second, and related to the first, it is possible to study the spread of events over a geographical area; this is even easier if extensive geographical identifiers are available on the data record. Given that information about events is usually collected when the events happen, there is much less opportunity for errors because of faulty recall.

Using administrative data is also advantageous for uncovering information that a survey respondent is unlikely to provide in an interview. In our work, we were relatively certain that families would underreport their incidence of abuse or neglect. The same issue exists for mental health or substance abuse treatment. Although survey methods have progressed significantly in addressing sensitive issues, administrative data can prove to be an accurate source of indicators for phenomena that are not easily reported by individuals—if one can satisfactorily address issues of accessing sensitive or confidential data.

Because the data record for an individual or case is likely viewed often by the program staff, opportunities exist for correcting and updating the data fields. The value of this is even greater when the old information is maintained in addition to the updates. A major problem with administrative data archiving and storing is that when data are updated, the old information is lost when it is overwritten.

As noted, the disadvantages of administrative data are often listed as a contrast to the characteristics of survey data. Although this may be a straw man argument, other legitimate concerns should be addressed when using administrative data. The concerns are related to the choice-, event-, or participation-based nature of the data; the reliability of administrative data for research purposes; the lack of adequate control variables; and the facts that all outcomes of interest are not measured (e.g., some types of indicators of well-being) that data are available only for the periods that the client is in the program, and that the level of reliability of administrative data is uncertain. Also, the data are difficult to access because of confidentiality issues (as far as getting informed consent) and because of bureaucratic issues in obtaining approval. When the data will be available, therefore, is often unpredictable.

Finally, there is often a lack of documentation and information about quality. One must do ethnographic research to uncover "qualitative" information about the condition of the data. There is no shortcut for understanding the process behind the collection, processing, and storage of the administrative data.

## ASSESSING THE QUALITY OF THE DATA AND CLEANING THE DATA FOR RESEARCH PURPOSES

In this section, we present strategies for determining if a particular administrative data set can be used to answer a particular question. Researchers seldom go directly to the online information system itself to assess its quality—although this may be one step in the process. Typically, government agencies give researchers both inside and outside the agency an extract of the information system of interest. This file may be called a "pull" file. It is a selection of data fields, never all of them, typically on all individuals in the information system during a specified period of time created for a particular purpose, usually not specified each time a request for data is made. Any one actual pull refers to a time period that corresponds to some administrative time period—for example, month or fiscal year. These cross-sectional pulls are very useful for agency purposes because they describe the point-in-time caseload for which an agency is responsible. As we will explain, this approach is not ideal for social research or evaluation.

The programming for a pull file is often a time-consuming task that is done as part of the system design based on the analytic needs at the time of the design. Even a small modification to the pull file may be costly or impossible given the capacity of the state or county agency information systems division. The advantage of this practice is that multiple individuals usually have some knowledge of the quality of the pull file—they may know how some of the fields are collected and how accurate they are. The disadvantage is that it probably requires additional cleaning to answer a particular set of research questions.

We cannot stress enough the importance of assessing data sets individually for each new research project undertaken. A particular data set may be ideal for one question and a disaster for another. Some fields in a database that may be perfectly reliable because of how the agencies collect or audit these fields, while other fields may almost seem to contain values entered in a random manner. Also, a particular programmatic database may have certain fields that are reliable at one point in time and not at other points. Needless to say, one field may be entered reliably in one jurisdiction and not in another.

For example, income maintenance program data are ideal for knowing the months in which families received Aid to Families with Dependent Children (AFDC) or Temporary Assistance for Needy Families (TANF) grants. However, because they rely on the reporting of grantees for employment information and there are often incentives for providing inaccurate information, addressing questions about the employment of TANF recipients using income maintenance program data is not ideal. Furthermore, information about the grantee, such as marital status or education, may only be collected at case opening and therefore is more likely to be inaccurate the longer the time since the case opening. Undertaking these tasks of assessing data quality is quite time consuming and resource intensive. The resource requirements are similar to those of cleaning large survey data sets, however, where to go to get information to do the cleaning is often unclear. Often documentation is unavailable and the original system architects have moved to other projects. Therefore, cleaning administrative data is often a task that goes on for many years as more is learned about the source and maintenance of the particular database.

In the following paragraphs, we provide some of the strategies and methods that we use to assess and address issues of data quality in the use of administrative data. The most basic, and perhaps best, of these is to compare the data with another source on the same event or individual. We will end with a discussion of that strategy.

### Assessing Data Quality

Initially, the researcher would want to assess if the data entry were reliable, which would include knowing whether the individual collecting the data had the skill or opportunity to collect reliable information. The questions that should be asked are as follows:

- What is the motivation for collecting the data? Often a financial or contractual motivation produces the most reliable data. When reimbursement is tied to a particular data field, both the payer and the payee have incentives to ensure that neither party is provided with an additional benefit. The state agency does not want to pay more TANF that it needs to pay, and a grantee (or his or her advocate) wants to ensure that the family gets all to which they are entitled. Also, an agency may have a legal requirement to track individuals and their information. Properly tracking the jail time of incarcerated individuals would seem to be one such activity for which one could be fairly certain of the data accuracy—although not blindly so.
- Is there a system for auditing the accuracy of the data? Is there a group of individuals who sample the data and cross-check the accuracy of the data with another source of the information? In some agencies, the computer records will be compared to the paper files.
- Are the data entered directly by the frontline worker? Adding a step to the process of entering the data—having a worker filling out a paper form and then passing it on to a data entry function—allows another opportunity for error and typically also excludes the opportunity for the worker to see the computerized record in order to correct it.
- Do "edit checks" exist in the information system? If there is no direct audit of the data or the data are not entered or checked by a frontline worker, having edit checks built into the data entry system may address some errors. These checks are programmed to prevent the entry of invalid values or not entering anything into a field. (This is similar to the practice of programming skip patterns or acceptable values for data entry of survey instruments.) For example, an edit check can require that a nonzero dollar amount is entered into a current earnings field for those individuals who are labeled as employed.
- What analyses have been done with these data in the past? There is no substitute for analyzing the data—even attempting to address some of the research questions—in the process of assessing the quality, especially when the

administrative data have not been used extensively. A good starting point for such analysis is examining the frequencies of certain fields to determine if there are any anomalies, such as values that are out of range; or examining inexplicable variation by region, suggesting variation in data entry practices; or seeking missing periods of the time series. Substantive consistency of the data is an important starting point as well. One example of this with which we have been wrestling is why 100 percent of the AFDC caseload were not eligible for Medicaid. We were certain that we had made some error in our record linkage. When we conferred with the welfare agency staff, they also were stymied at first. We eventually discovered that some AFDC recipients are actually covered by private health insurance through their employers. With this information, we are at least able to explain an apparent error.

- Finally, are the items in the data fields critical to the mission of the program? This issue is related to the first noted issue above. Cutting checks is critical for welfare agencies. If certain types of data are required to cut checks, the data may be considered to be accurate. For example, if a payment cannot be made to an individual until a status that results in a sanction is addressed, one typically expects that the sanction code will be changed so payment can be made. On the other hand, if a particular assessment is not required for a worker to do his or her job or if an assessment is outside the skill set of the typical worker doing the assessment, one should have concerns about the accuracy (Goerge et al., 1992). For example, foster care workers have been asked to provide the disability status of the child on his or her computerized record. This status in the vast majority of the cases has no impact on the decision making of the worker. Therefore, even if there is an edit check that requires a particular set of codes, one would not expect the coding to be accurate.

We will continue to give examples of data quality issues as we discuss ways to address some of them. The following examples center on the linking of an administrative data set with another one in order to address inadequacies in one set for addressing a particular question.

The choice-based nature of administrative data can be addressed in part by linking the data to a population-based administrative data set. Such linkages allows one to better understand who is participating in a program and perhaps how they were selected or selected themselves into the program. There are some obvious examples of choice-based linking data to population-based data. In analyzing young children, it is possible to use birth certificate data to better understand what children might be selected into programs such as Women, Infants and Children (WIC), Early, Periodic, Screening, Diagnosis And Treatment Program (EPSDT), and foster care. If geographic identifiers are available, administrative data can be linked to census tract information to provide additional information on the context as well as the selection process. For example, knowing how many poor children live in a particular census tract and how many children participate in a welfare program can address whether the welfare population is representative of the entire population of those living at some fraction of the poverty level.

If one is interested in school-age children, computerized school data provide a base population for understanding the selection issues. One example is to link the 6- to 12-year-old population and their School Lunch Program (SLP) information to Food Stamp administrative data to understand who uses Food Stamps and what population the administrative data actually represent. Because SLP eligibility is very similar to Food Stamps (without the asset test), such data could provide a very good idea of Food Stamp participation. The criticism that administrative data only tracks individuals while they are in the program is true. Extending this a bit, administrative data, in general, only track individuals while they are in some administrative data set. Good recent examples of addressing this issue are the TANF leaver studies being conducted by a number of states. They are linking records of individuals leaving TANF with UI and other administrative data, as well as survey data, to fill in the data that welfare agencies typically have on these individuals—data from the states' FAMIS or MMIS systems. Especially when we are studying welfare or former welfare recipients, it is likely that these individuals appear in another administrative data set—Medicaid, Food Stamps, child support, WIC, or child care, to name a few. Although participation in some of these is closely linked to income maintenance, as we have learned in the recent past, there is also enough independence from income maintenance programs to provide useful post-participation information. Finally, if they are not in any of these social programs databases, they are likely to be in the income tax return databases or in credit bureau databases, both now becoming data sets used more commonly for social research (Hotz et al., 1999).

A more thorny problem may be situations in which an individual or a family leaves the jurisdiction where administrative data were collected. We may be "looking" for them in other databases when they may have moved out of the county or state (or country) in which the data were collected. The creation of national-level data sets may help to address this problem simply through a better understanding of mobility issues, if not actually linking data from multiple states to better track individuals or families.

It is certainly possible that two administrative databases will label an individual as participating in two programs that should be mutually exclusive. For example, in our work in examining the overlap of AFDC or TANF and foster care, we find that children are identified as living with their parents in an income maintenance case when they are actually living with foster parents. Although these records eventually may be reconciled for accounting purposes (on the income maintenance side), we do need to accurately capture the date that living in an AFDC grant ended and living in foster care began. Foster care administrative data typically track accurately where children live on a day-to-day basis. Therefore, in studying these two programs, it is straightforward to truncate the AFDC record when foster care begins. However, one would want to "overwrite" the

AFDC end date so that one would not use the wrong date if one were to analyze the overlap between AFDC and another program, such as WIC, where the participation date may be less accurate than in the foster care program.

Basic reliability issues also arise. For example, some administrative databases do a less than acceptable job of identifying the demographic characteristics of an individual. At a minimum, data entry errors may occur in entering gender or birth dates (3/11/99, instead of 11/3/99). Also, data on workers' determination of race/ethnicity might not be self-reported, or race/ethnicity might not be critical to the business of the agency, although this is often a concern of external parties. In some cases, when one links two administrative data files, the race/ethnicity codes for an individual do not agree. This discrepancy may be a particular problem when the data files cover time periods that are far apart, because some individuals do change how they label themselves and the labels used by agencies may change (Scott, 2000). Linking administrative data with birth certificate data—often computerized for decades in many states—or having another source of data can help address these problems. We will discuss this issue below when we discuss record linkage in detail (Goerge, 1997).

### *Creating Longitudinal Files*

As mentioned earlier, the pull files provided by government agencies are often not cumulative files and most often only span a limited time period. For most social research, longitudinal data are required, and continuous-time data—as opposed to repeated, cross-sectional data—are preferred, again depending on the question. Although these pull files may contain some historical information, this is often kept to a minimum to limit the file size. The historical information is typically maintained for the program's unit of administration. For TANF, this is the family case. For Food Stamps, it is the household case. In either program, the historical data for the individual member of the household or family are not kept in these pull files. The current status typically is recorded in order to accurately calculate the size of the caseload. Therefore, to create a "clean" longitudinal file at the individual level, one must read each monthly pull file in order to recreate the individual's status history. Using a case history for an individual would be inaccurate. An example is the overlap between AFDC and foster care discussed earlier. The case history for the family—often that of the head of the household, and which may continue after the child enters foster care—would not accurately track the child's income maintenance grant participation. More on this topic is discussed in the following sections.

### *Linking Administrative Data and Survey Data*

The state of the art in addressing the most pressing policy issues of the day is to use administrative data and survey methods to obtain the richest, most accurate data to answer questions about the impact and implementation of social programs. The TANF leaver studies mentioned earlier, which use income maintenance administrative data to select and weight samples and TANF and other programmatic databases to locate former TANF participants, provide certain outcome measures (e.g., employment and readmission) and characteristics of the grantees and members of the family. Survey data are used to obtain perceptions about employment and fill in where the administrative data lack certain information. Administrative lists have also been used to generate samples for surveys that intend to collect data not available in the administrative data.

Such studies can be helpful in understanding data quality issues when the two sources of data overlap. For example, we worked with other colleagues to compare reports of welfare receipt with administrative data and were able to gauge the accuracy of participant recall. We have some evidence for situations in which it is quite defensible to use surveys when administrative data are too difficult or time consuming to obtain. For example, although childcare utilization data may be available in many states, the data often are so decentralized that bringing them together into a single database may take many more resources than a survey. Of course, this depends on the sample size needed. However, much more needs to be done in this vein to understand when it is worthwhile to take on the obstacles that are more the rule than the exception in using administrative data.

## ADMINISTRATIVE DATA RECORD LINKAGE

A characteristic of administrative data that offers unique opportunities for researchers is the ability to link data sets in order to address research questions that have otherwise been difficult to pursue because of lack of suitable data.[2] For example, studying the incidence of foster care placement, or any low-incidence event, among children who are receiving cash assistance requires a large sample of children receiving cash assistance given that foster care placement is a rare event. The resources and time required to gather such data using survey methods can be prohibitive. However, linking cash assistance administrative data and foster care data solves the problem of adequate sample size in a cost-effective way. Linking administrative data sets is also advantageous when the research

[2] Our discussion of record linkage focuses on its application at the individual level, where research interests require individual-level linkage as opposed to aggregate population overlap statistics. In other words, we address the need to follow the outcome of interest at the individual level, focusing on research questions dealing with temporal data on timing and sequence. Utility of statistical techniques that are developed to estimate aggregate population overlap among different data sets without doing individual-level record linkage, such as the probabilistic population estimation method, is beyond the scope of our discussion in this paper. (For further information on such a technique, refer to Pandiani et al., 1998.)

interest is focused in one particular service area. For example, if one is interested in studying the multiple recurrences of some event, such as multiple reentries to cash assistance, recurring patterns of violent crime, or reentries to foster care, the size of the initial baseline sample must be large enough to observe an adequate number of recurrences in a reasonable time period. Linking administrative data over time at the population level for each area of concern is an excellent resource for pursuing such research questions without large investments of time and financial resources.

When the linked administrative data sets are considered as an ongoing research resource, it is preferable to have data from the entire population from each source database that are linked to each other and maintained. Given the large number of cases needed to be processed during record linkage, the idea of working with data from the entire population could overwhelm the researcher. However, because most data processing now is done using computers, the sheer size of the data files needed to be linked is typically not a major factor in the time and resources needed. On the other hand, the importance of having good programmers with necessary analytic and programming skills cannot be overemphasized for achieving successful record-linking results. Because the amount of skilled programming for a sample file may be equal to the amount needed for an entire large file, the additional cost involved in linking the entire files rather than samples is justifiable in computerized record-linking situations. The advantages that arise from having population data (as opposed to samples from each system or some systems) far exceed the costs involved. When tracking certain outcomes of a base population using linked data, one needs at least the population-level data from the data source that contains information about the outcome of interest.

For example, suppose one is interested in studying the incidence of receiving service *X* among a 10 percent random sample of a population in data set A. The receipt of service *X* is recorded in data set B. Because the researcher must identify *all* service *X* receipt for the 10 percent sample in data set A, the sample data must be linked to the entire population in data set B. Suppose the researcher only has a 10 percent random sample of data set B. Linking the two data samples would provide, at best, only 10 percent of the outcomes of interest identified in the 10 percent sample of the base population A. Furthermore, the "unlinked" individuals in the sample would be a combination of those who did not receive service *X* and those who received service *X* but were not sampled from data set B. Because one cannot distinguish the two groups among the "unlinked" individuals, any individual-level analysis becomes impossible (see Deming and Glasser, 1959, for a discussion of the issue of linking samples and the difficulty associated with it).

## Research Applications of Data Linking

There are four different research applications of linked data sets. Each represents a different set of issues and challenges. The four types of linking applications can be broadly defined as: (1) linking an individual's records within a service system over time, (2) linking different information system data sets across service areas, (3) linking survey data to administrative data sets when the survey sample is drawn from an administrative data set, and 4) linking sample data to administrative data sets when the sample is drawn independent of administrative data.

The first type of linking application is the most common. Typically researchers take advantage of administrative data's historical information for various longitudinal analyses of service outcomes. Often this type of research requires linking data on individuals across several cross-sectional extracts from an agency's information system. Many agency information systems only contain information on the most recent service activities or service populations. Some information systems were designed that way because the agency's activity is defined as delivering services to a caseload at a given point in time or at some intervals. A good example would be a school information system in which each school year is defined as the fixed service duration, and each school year population is viewed as a distinct population. In this case, there is typically no unique individual ID in the information system across years because every individual gets a new ID each year—one that is associated with the particular school year. Even in a typical state information system on cash assistance, case status information is updated (in other words, overwritten) in any month when the status changes. To "reconstruct" the service histories, as discussed in the earlier section on cleaning, one must link each monthly extract to track service status changes.

At times, the information system itself is longitudinal, and no data are purged or overwritten. Even when the database is supposedly longitudinal, a family or an individual can be given multiple IDs over time. For example, many information systems employ a case ID system, which includes a geographic identifier (such as county code or service district code) as part of a unique individual ID. In this instance, problems arise when a family or an individual moves and receives a different ID. Our experience suggests that individuals are often associated with several case IDs over time in a single agency information system. Sometimes individuals may have several agency IDs assigned to them either because of a data entry error or a lack of concerted effort to track individuals in information systems. In any situation outlined here, careful examination of an explicit linking strategy is necessary.

The second type of linking application most often involves situations in which different agency information systems do not share a common ID. Where the funding stream and the service delivery system are separate and categorical in nature, information systems developed to support the functions of each agency are not linked to other service information systems. In some instances, information systems even in a single agency do not share a common ID. For example, many child welfare agencies maintain two separate legacy information systems; one tracks foster care placement and payments and the other records child maltreatment reports. Although following the experiences of children from a report of abuse or neglect to a subsequent foster care event is critical for child welfare agencies, the two systems were not designed to support such a function. Obviously, where there is no common ID, linking data records reliably and accurately across different data sources is an important issue. Also, as in the case of linking individual records over time in a single information system, there is always a possibility of incorrect IDs, even when such a common ID exists. In fact, a reliable record linking between the two information systems that contain a common ID on a regular basis could provide a means to "correct" such incorrect IDs. For example, when the data files from the two systems are properly linked by using data fields other than the common ID, such as names and birth dates, the results of such a link could be compared to the common IDs in the information systems to identify incorrectly entered IDs.

The third type of linking application is when a sample of individuals recorded in administrative data is used as the study population. In such a study, researchers employ survey methods to try to collect information not typically available in administrative data. Items such as unreported income, attitude, and psychological functioning are good examples of information that is unavailable in administrative data. Most often, this type of application is not readily perceived as a linking application. However, when researchers use administrative data to collect information about the service receipt history of the sample, either retrospectively or prospectively, they face the same issues as one faces in linking administrative data in a single information system or across multiple systems. Also, if researchers rely on the agency ID system to identify the list of "unique" individuals when the sampling frame is developed, the quality of the agency ID has important implications for the representativeness of the sample. The degree of multiple IDs for the same individuals should be ascertained and the records unduplicated at the individual level for the sampling frame.

The fourth type of linking application involves cases in which researchers supplement the information collected through survey methods with detailed service information; they do this by linking survey data to service system administrative data after the survey is completed. Because the sample is drawn independent of the administrative data, no common ID is designated between the sample and the administrative data. Here the major concern is the kinds of identifying information that are available for linking purposes from both data sources. In particular, whether and how much identifying information—such as full names, birth dates, and Social Security numbers (SSNs)—is available from the survey data is a critical issue. When the identifying information is collected, data confidentiality issues might prohibit researchers from making information available for linking purposes.

## TECHNICAL ISSUES IN LINKING

### Two Methods of Linking: Probabilistic and Deterministic Record-Linkage Methods

Linking data records reliably and accurately across different data sources is key to the success in the four applications outlined. In this section, we focus on the data linkage methods. Our main purpose is to provide basic concepts for practitioners rather than to present a rigorous theoretical method. Our discussion focuses on two methods of record linkage that are possible in automated computer systems: deterministic and probabilistic record linking.

#### *Deterministic Record Linkage*

Deterministic linkage compares an identifier or a group of identifiers across databases; a link is made if they all agree. For example, relying solely on an agency's common ID when available for linking purposes is a type of deterministic linking. When a common ID is unavailable, standard practice is to use alternative identifiers—such as SSNs, birth dates, and first and last names of individuals—that are available in two sets of data. Researchers also use combinations of different pieces of identifying information in an effort to increase the validity of the links made. For example, one might use SSN and the first two letters of the first and last names. In situations where an identifier with a high degree of discriminating power (such as SSN) is unavailable, a combination of the different pieces of identifying information must be used because many people have the same first and last names or birth dates. What distinguishes deterministic record linkage is that when two records agree on a particular field, there is no information on whether that agreement increases or decreases the likelihood that the two records are from the same individual. For example, the two situations in which, on last name, Goerge matches Goerge, and where Smith matches Smith, would be treated with similar matching power, even though it is clear that because there are few Goerges and many Smiths, these two matches mean different things.

#### *Probabilistic Record Linkage*

Because of the problems associated with deterministic linking, and especially when there is no single identifier distinguishing between truly linked records (records of the same individual) in the data sets, researchers have developed a set of methods known as probabilistic record linkage.[3] Probabilistic record

[3] Presentation of the detailed mathematical process of probabilistic record linkage method is beyond the scope of this paper. Readers interested in the theory should refer to references cited in this section of the paper.

linking is based on the assumption that no single match between variables common to the source databases will identify a client with complete reliability. Instead, the probabilistic record-linking method calculates the probability that two records belong to the same client by using multiple pieces of identifying information. Such identifying data may include last and first name, SSN, birth date, gender, race and ethnicity, and county of residence.

The process of record linkage can be conceptualized as identifying matched pairs among all possible pairs of observations from two data files. For example, when a data file A with A observations and a data file B with B observations are compared, the record-linkage process attempts to classify each record pair from the A by B pairs into the set of true matches (M set) and the set of true nonmatches (U set). First introduced by Newcombe et al. (1959) and further developed by Fellegi and Sunter (1969), the two probabilities for each field that are needed to determine if a pair belongs to M or U are *m* and *u* probabilities. Each field that is being compared in the record-linking process has *m* and *u* probabilities. The *m* probability is the probability that a field agrees given that the record pair being examined is a matched pair. The *m* probability is a measure of validity of the data field used in the record-linkage process because it is essentially one minus the error rate of the field. Thus, one can see that a more reliable data field will provide greater *m* probability. The *u* probability is the probability that a field agrees given that the record pair being examined is not a matched pair. This is a chance probability that a field agrees at random. For example, if the population has the same number of males and females, the *u* probability will be .5 because there is a 50 percent chance the gender field will match when the pair being examined is not a matched pair. Accordingly, a variable such as SSN will have a very low *u* probability because it is very unlikely that different individuals have the same SSN. Although there are many methods to calculate M and U probabilities, recent studies show that maximum-likelihood-based methods such as the Expectation-Maximization (EM) algorithm is the most effective of those developed and tested (Winkler, 1988; Jaro, 1989).

Using *m* and *u* probabilities, Fellegi and Sunter (1969) define weights that measure the contribution of each field to the probability of making an accurate classification of each pair into M or U sets. The "agreement" weight when a field agrees between the two records being examined is calculated as *log2(m/u)*. The "disagreement" weight when a field does not agree is calculated as *log2((1–m)/(1–u))*. These weights indicate how powerful a particular variable is in determining whether two records are from the same individual. These weights will vary based on the distribution of values of the identifiers. For example, a common last name match will provide a lower agreement weight than a match with a very uncommon name because *u* probability for such a common name will be greater than the uncommon name.

Fellegi and Sunter (1969) further showed that a composite weight could be calculated by summing the individual data field's weights. Using the composite weights, one can classify each pair of records into three groups: a link when the composite weight is above a threshold value (U), a non link when the composite weight is below another threshold value (L), and a possible link for clerical review when the composite weight is between U and L. Furthermore, the threshold values can be calculated given the accepted probability of false matches and the probability of false nonmatches (Fellegi and Sunter, 1969; Jaro, 1989). This contrasts favorably with the link or non link dichotomy in deterministic linkage.

Since the seminal work by Fellegi and Sunter (1969), the main focus of record linkage research has been how to determine the threshold values of U and L to improve the accuracy of determining what the threshold weight is for a certain link, as well as the threshold value for a certain non link. Recent development in improving record linkage allows us to take advantage of the speed and cost that computerized and automated linkage confer, such as deterministic matching, while allowing a researcher to identify at which "level" a match would be considered to be a true one (see for example; Jaro, 1989; Winkler, 1993, 1994, 1999).

### Standardization and Data-Cleaning Issues in Record Linking

Regardless of which method of deterministic linking is used, entry errors, typographical errors, aliases, and other data transmission errors can cause problems. For example, one incorrectly entered digit of a Social Security number will produce a nonmatch between two records for which all other identifying information is the same. Names that are spelled differently across different systems also cause a problem. A first name of James that is recorded in one system as Jim and in the other as James will produce a nonmatch when the two records, in fact, belong to the same individual. The data cleaning in the record linkage process often involves (1) using consistent value-states for the data fields used for linking, (2) parsing variables into components that need to be compared, (3) dealing with typographical errors, and (4) unduplicating each source file for linkage.

Because record linking typically involves data sets from different sources, the importance of standardizing the format and values of each variable is used for linking purposes cannot be overemphasized. The exact coding schemes of all the variables from different source files used in the matching process should be examined to make sure all the data fields have consistent values. For example, males coded as "M" in one file and "1" in another file should be standardized into a same value. In the process, missing and invalid data entries also should be identified and coded accordingly. For example, a birth year 9999 should be recognized as a missing value before the data set is put into the record-linking process. Otherwise, records with a birth year 9999 from the two data sets can be linked because they have the "same" birth year. We also find that standardization of names in the matching process is important because names are often spelled differently or misspelled altogether across agency information systems. For example, a first name of Bob, Rob, and Robert should be standardized into a same first name such as Robert to achieve better record-linking results.

The data cleaning and standardization in matching process often requires parsing variables into a common set of components that can be compared. Names may have to be split or parsed into first name, middle initial, and last name and suffix (e.g., Junior). In using geographic information, street names and the form of the addresses must be standardized. This may mean parsing the address into number (100), street prefix (West), street name (Oak), and street suffix (Boulevard).

Because of typographical errors, an exact character-by-character comparison for certain fields used in a record-linking process may miss many "true" matches. A good example is variant spellings of names. For example, character-by-character comparison of a last name spelled as "Goerge" in one data file to a misspelled name "George" in another file would cause disagreement in the last name comparison even though "George" in the second file was a misspelling. In some situations, these types of typological errors can be a serious problem in record linkage. Winkler and Thibaudeau (1991) and Jaro (1989) describe how researchers at the U.S. Bureau of the Census reported that about 20 percent of last names and 25 percent of first names disagreed character by character among true matches in the Post Enumeration Survey. In recent years researchers in the field of record linkage have made substantial progress in developing algorithms to deal with such problems in character-by-character comparisons. As a result, some complex string comparator algorithms also have been developed to determine how close two strings of letters or numbers are to each other that account for insertions, deletions, and transpositions (Jaro, 1985, 1989; Winkler, 1990; Winkler and Thibaudeau, 1991).

In the record linkage process, one critical data cleaning process is to "unduplicate" each source data set before any two data sets are linked. As discussed earlier, often individuals are associated with several IDs because of data entry errors or a lack of concerted effort to track individuals in agency information systems. Obviously, multiple records for the same individual in each data set being linked produce uncertain links because the process must deal with *N* to *N* link situations. Unduplication of the records in a single data set can be thought as "self-match" of the data set. Once a match has been determined, a unique number is assigned to the matched records so that each individual can be uniquely identified. The end result of the unduplication process is a "person file," which contains the unique number assigned during unduplication and the individual's identifying data (name, birth date, race/ethnicity, gender, and county of residence) with a "link file" that links the unique individual ID to all the IDs assigned by an agency. Once each data set is unduplicated in such a way, the unduplicated person files can be used for cross-system record links.

### Accuracy of Record Linking

Regardless of which method is used, the ultimate concern is in the degree of validity and accuracy of the links made. Whether it is a deterministic or probabilistic record-linkage technique that is used, the linking process essentially involves making an educated guess about whether two records belong to the same individual. Because the decision is a guess, it might be wrong. These errors in record linkage can be viewed as making false-positive and false-negative errors. A false-positive error occurs when the match is made between the two records when the two records, in fact, do not belong to the same individual. This type of error is comparable to a Type I error in statistical hypothesis testing. A false-negative error occurs when the match is not made between the two records when they, in fact, belong to the same individual. The type of error is comparable to a Type II error in statistical hypothesis testing.

As with Type I and Type II errors, although the probability of making a false-positive error can be easily ascertained in the linking process, determining the probability of a false negative error is more complex. Because the "weights" calculated in the probabilistic record-linkage method are essentially relative measures of the probability of a match, the weights can be converted to an explicit probability that a record pair is a true match (i.e., 1-false positive error rate). Belin and Rubin (1995) introduced a method for estimating error rates for cutoff weight values in the probabilistic record-linkage process. Many developments also have been made in dealing with linkage errors in post-linkage analysis stages (such as a regression analysis using linked files) (see Scheuren and Winkler, 1993). In the case of deterministic record linkage, an audit check on the matched pairs could provide an estimate of false-positive errors. Estimating the false-negative error rate is much more complex because it conceptually requires knowing the true matches prior to the linking and comparing the linking results to the true matches.

Adding to the complexity, as one tries to reduce one type of error, the other type of error increases. For example, in an effort to reduce false-positive errors, one might use a stringent rule of labeling the compared matches as matched pairs only when they are "perfect" matches. In the process, a slight difference in identifying information (such as one character mismatch in the names) might cause a non link when, in fact, the two records belong to the same individual. Hence, false-negative error rates increase. In the opposite scenario, one might accept as many possible matches as true matches, thereby relaxing the comparison rule by reducing false-negative errors. In this case, false-positive errors increase.

### An Example

In practice, it would be useful to consider false-positive and false-negative error rates as a means to compare different methods of record linkage. One practical issue researchers face is determining which linkage method to use, especially when an ID variable such as SSN is available in the two data sets to be linked. Although most experts agree that probabilistic record linkage is a more reliable method than deterministic linking, it requires extensive programming or the purchase of software, which can be quite expensive. If one does not have ready access to suitable commercial record-linkage software, it may be sufficient for a good programmer to write a quick deterministic linkage program that matches a good deal of the records. There are other situations where there is no apparent common ID and the quality of identifying information in the data is questionable (such as many typographical errors in certain data fields), so that only using probabilistic record-linkage methods will yield acceptable linking results.

We present some empirical data comparing the two methods in the following paragraphs and corresponding tables. The methods compared are a deterministic record link using SSN and a probabilistic link using SSN, full name, birth date, race/ethnicity, and county of residence. We use data from the Client Database and the Cornerstone Database from the Illinois Department of Human Services. The Client Database records receipt of AFDC/TANF and Food Stamps and documents all those who are registered as eligible for Medicaid from 1989 to the present. The Cornerstone database contains WIC and case management service receipt at the individual level. There is no common ID between the two systems, while SSN and other identifying information are available in both systems.

Because both systems serve mainly low-income populations and contain data for a long period of time, we expected a high degree of overlap between the two populations. When the existence of SSN in both systems is examined, we find that about 38 percent of the Cornerstone records have missing SSNs while the Client Database identifies nearly 100 percent of the SSNs. In our first analysis, we excluded the records with missing SSNs from the Cornerstone data. Table 7-1 compares the number of matched and unmatched Cornerstone data records to the Client Database records comparing the deterministic match using SSN and

TABLE 7-1 Comparison of SSN Match (Deterministic) Versus Probabilistic Match (Without Missing SSN)

| | | Probabilistic Matching Number | | | Probabilistic Matching Percent | | |
|---|---|---|---|---|---|---|---|
| | | Nonmatch | Match | Total | Nonmatch | Match | Total |
| SSN Matching | Nonmatch | 74,496 | 45,987 | 120,483 | 61.8 | 38.2 | 100.0 |
| | Match | 5,849 | 438,959 | 444,808 | 1.3 | 98.7 | 100.0 |
| | Total | 80,345 | 484,946 | 565,291 | 14.2 | 85.8 | 100.0 |

the probabilistic match using all other identifying information, including SSN. As shown in Table 7-1, the probabilistic match identified about 86 percent of non-SSN-missing Cornerstone record links to the Client Database. The SSN deterministic method identified about 84 percent of the matches.

Although the percentage of overall matches is similar, the distribution of error types is quite different, as shown in Table 7-1. The false-negative error rate of using the SSN deterministic record-linking method when compared to the results from the probabilistic match is about 38 percent. On the other hand, the false positive error rate is about 1 percent. We checked the results of the probabilistic link from random samples of the disagreement cells (i.e., probabilistic match/SSN no match and probabilistic nonmatch/SSN match) to verify the validity of the probabilistic match. We found that the probabilistic match results are very reliable. For example, we found that most of the pairs in the probabilistic match/SSN no match cell involve typographical errors in SSN with the same full name and birth date. Also, we found that most of the pairs in the probabilistic nonmatch/SSN match involve entirely different names or birth dates. Although the findings might be somewhat different when applied to different data systems, our finding suggests that employing a probabilistic record-linkage method helps to reduce both false-negative and false-positive errors. The findings also show that the benefit of employing probabilistic record linkage is greater in reducing false-negative errors (Type II errors) than in reducing false-positive errors (Type I errors) when compared with a deterministic record-linkage method using SSN.

Next, we included the Cornerstone records with missing SSN in the analysis. The findings are presented in Table 7-2. As one might expect, the probabilistic record-linkage method significantly enhances the results of the match by linking many more records. Compared with the results presented in Table 7-1, the number of matches from the probabilistic match increases by about 210,000 records, representing about 62 percent of matches made among the records with missing SSNs. Again, most of the benefit of using the probabilistic linkage method is in reducing false-negative errors. With about 30 percent of the records showing

TABLE 7-2 Comparison of SSN Match Versus Probabilistic Match (With Missing SSN)

| | | Probabilistic Matching Number | | | Probabilistic Matching Percent | | |
|---|---|---|---|---|---|---|---|
| | | Nonmatch | Match | Total | Nonmatch | Match | Total |
| SSN Matching | Nonmatch | 199,442 | 260,720 | 460,162 | 43.3 | 56.7 | 100.0 |
| | Match | 5,782 | 444,540 | 475,537 | 1.3 | 98.7 | 100.0 |
| | Total | 205,224 | 699,478 | 904,702 | 22.7 | 77.3 | 100.0 |

missing SSNs, the false-negative error rate of the SSN deterministic link method is about 57 percent. From the above results, one can conclude that when SSN information is nearly complete in the two data sets, the added benefit of using probabilistic linking is relatively smaller (although quite significant) and the benefit comes largely from identifying false-negative errors. As the number of records with missing SSN increases, the benefit of employing a probabilistic record-linkage method increases.

Very often in practice, being able to link different data sources involves many other issues than that of the linking method. A key issue is data confidentiality, especially when full names are needed for linking purposes in the absence of a common ID. One possible solution to the confidentiality issue is the use of Soundex codes. Even though Soundex is not a complete method to preserve confidentiality, it provides added protection compared to using actual full names. The Soundex system is a method of indexing names by eliminating some letters and substituting numbers for other letters based on a code. Although experts disagree on what should be the authoritative Soundex system, the most familiar use of Soundex is by the U.S. Bureau of the Census, which uses it to create an index for individuals listed in the Census. Because it is impossible to derive an exact name from a Soundex name, the system can be used to conceal the identity of an individual to some extent. (For example, similar sounding but different names are coded to a same Soundex name.)

The issue in probabilistic linking, however, is how valid a Soundex name is alone compared to using full names. We examine this issue by comparing the two methods involving the same data sets with the other identifying information fixed. The other identifying information variables are SSN, birth date, race/ethnicity, and county of residence. Table 7-3 presents the results of such an exercise. The agreement rate between the Soundex-only method and the full-name method is very high—close to 100 percent. The results suggest that Soundex coded names work equally as well as full names in a probabilistic match. In situations in which full names cannot be accessed for linking purposes, Soundex

TABLE 7-3 Comparison of Full Name Match Versus Soundex Code Match

| | | Full Name Matching<br>Number | | | Full Name Matching<br>Percent | | |
|---|---|---|---|---|---|---|---|
| | | Nonmatch | Match | Total | Nonmatch | Match | Total |
| Soundex | Nonmatch | 256,628 | 221 | 256,849 | 99.9 | 0.1 | 100.0 |
| Matching | Match | 40 | 43,111 | 43,151 | 0.1 | 99.9 | 100.0 |
| | Total | 256,668 | 43,332 | 300,000 | 85.6 | 14.4 | 100.0 |

names might be a good alternative while providing a better means of protecting individual identities.[4]

## CONCLUSION

### Recommendations

We recommend a number of activities in the cleaning of administrative data for research use. These include:

- Examining the internal consistency of the data;
- Examining how the data were collected, processed, and maintained before delivery to the researcher;
- Taking every opportunity to compare with other data sets, either survey or administrative, through record linkage; and,
- Most important, getting to know the operations of the program, not just the collection of administrative data, but also how services are provided so that inconsistencies in the data might be understood better.

We also recommend using probabilistic record linkage and not relying on any one identifier for linking records. We believe our analysis above makes this case. The golden rule of record linkage is that there is no such thing as a unique identifier, because individuals can match on many identifiers. In many cases the same SSN has been provided to two or more individuals.

### Developments in Information Technology That May Improve Administrative Data

Much of what is discussed previously is required because public policy organizations are still, for the most part, in their first generation of information systems. These "legacy" systems are typically a decade or older mainframe installations that do not take advantage of much of today's technology. Data entry in the legacy systems, for example, is often quite cumbersome and requires a specialized data entry function. Frontline workers are typically not trained to do this or do not have the time or resources to take on the data entry task. An exception is in entitlement programs in some jurisdictions, where the primary activity for eligibility workers is collecting information from individuals and entering it into a computerized eligibility determination tool. The development of new graphical user interfaces that are more worker friendly—in that the screens

[4] Popular software programs such as SAS provide a simple method of converting names to Soundex codes.

flow in a way that is logical to a worker—is likely to have a positive effect on data entry both because of the ease of entry and because the worker may be able to retrieve information more easily. If this is the case, the worker will have a greater stake in the quality of the data.

The development of integrated online information systems, where a worker can obtain information on a client's use of multiple programs, also may have a positive effect on the quality of the data. First, the actual job of linking across the programs will likely be an improvement over the after-the-fact linking of records. For example, if an integrated system already exists, when a mental health case is opened for an individual with Medicaid eligibility, his or her records should be linked immediately. This, of course, requires an online record-linkage process for the one case or individual. Even though a researcher would still want to check whether an individual has multiple IDs, the process at the front end will greatly improve the quality of the analytic database.

Many states are now creating data warehouses in order to analyze many of the issues of multiple-program use and caseload overlap. These data warehouses "store" data extracts from multiple systems and link records from individuals across programs. If states are successful in creating comprehensive, well-implemented data warehouses, researchers may not have to undertake many of the cleaning or linking activities discussed in this paper. Government will have already done the data manipulations. The researchers, just as is typically done with survey data, will have to verify that the warehouse was well built. Although this may require some confidential information, it should make it easier to access administrative data.

## REFERENCES

Belin, T.R., and D.B. Rubin

1995 A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* 90(June):694-707.

Deming, W., R. Edwards, and G. Glasser

1959 One the problem of matching lists by samples. *Journal of the American Statistical Association* 54(June):403-415.

Fellegi, I.P., and A.B. Sunter

1969 A theory for record linkage. *Journal of the American Statistical Association* 64:1183-1210.

Goerge, R.M.

1997 Potential and problems in developing indicators on child well-being from administrative data. In *Indicators of Children's Well-Being*, R.M. Hauser et al., eds. New York: Russell Sage Foundation.

Goerge, R.M., J. Van Voorhis, S. Grant, K. Casey, and M. Robinson

1992 Special education experiences of foster children: An empirical study. *Child Welfare* 71:5.

Hotz, V.J., C. Hill, C.H. Mullin, and J.K. Scholz

1999 EITC Eligibility, Participation and Compliance Rates for AFDC Households: Evidence from the California Caseload. Working Paper 102. Joint Center for Poverty Research, University of Maryland and University of Michigan.

Jaro, M.A.

1985 Current record linkage research. In *Proceedings of the Section on Statistical Computing*. Alexandria, VA: American Statistical Association.

1989 Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 84(June):414-420.

Kalil, A., P.L. Chase-Lansdale, R. Coley, R. Goerge, and B.J. Lee

1998 Correspondence between Individual and Administrative Reports of AFDC Receipt. Unpublished paper presented at the Annual Workshop of the National Association for Welfare Research and Statistics, Chicago, August 2-5.

Pandiani, J., S. Banks, and L. Schacht

1998 Personal privacy versus public accountability: A technical solution to an ethical dilemma. *Journal of Behavioral Health Services and Research* 25:456-463.

Scheuren, F., and W.E. Winkler

1993 Regression analysis of data files that are computer matched. *Survey Methodology* 19:39-58.

Scott, C.

2000 Identifying the race/ethnicity of SSI recipients. In *Turning Administrative Systems into Administrative Systems.* Washington, DC: U.S. Department of the Treasury, Internal Revenue Service.

Winkler, W. E.

1988 Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association.

1990 String comparator metrics and enhanced decision rules in the Felligi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

1993 Matching and record linkage. In *Statistical Research Division Report 93/08.* Washington, DC: U.S. Bureau of the Census.

1994 Advanced methods for record linkage. In *Statistical Research Division Report 94/05.* Washington, DC: U.S. Bureau of the Census.

1999 The state of record linkage and current research problems. In *Statistical Research Division Report 99/04.* Washington, DC: U.S. Bureau of the Census.

Winkler, W.E., and Y. Thibaudeau

1991 An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. In *Statistical Research Division Report 91/09.* Washington, DC: U.S. Bureau of the Census.