QUALITY ISSUES IN THE USE OF ADMINISTRATIVE DATA RECORDS

Aileen Rothbard
University of Pennsylvania

Contents

Introduction

This paper examines the quality issues associated with employing administrative records from state and local agencies for use in monitoring, planning, evaluating, and integrating health and human services information for policy purposes. The objective of this paper is to provide a practical illustration of the key issues on data quality that are important in assembling and integrating administrative records for use in an integrated data system (IDS). A previous paper in this working series reviews the research related to these data quality issues (Boruch, 2012). Administrative record quality and integrated data systems are reviewed in another paper in the (Culhane, Fantuzzo, Rouse, Tam, Lukens, 2010).

Data quality must be addressed at multiple levels within the same program or agency and when integrating information from different sources. Thus, quality begins at the initial data entry step associated with an agency and continues through the single agency stage to the level at which an integrated data system processes and combines data elements from multiple agencies. With the full implementation of the Affordable Care Act (ACA) beginning in 2014, there will be an increasing demand for integrated data which is comprehensive in nature and incorporates patient level information on behavioral health, medical, and related support services. This type of data will be important for evaluating the medical care home programs that are poised to develop as part of the ACA. The Federal government has provided financial incentives for providers to implement electronic health records, which will provide information for a superhighway that includes comprehensive healthcare information for treatment personnel and consumers alike.

Administrative data systems, like Medicaid and Medicare eligibility and service claims records, state and county event or encounter service data, housing and homeless shelter information (i.e. Homeless Management Information System-HMIS), arrest and incarceration data, human service data from the Department of Health and Human Services (DHHS), and the like, are generally designed for internal reporting and reimbursement by a single agency or system. The use of these data in decision-making requires personnel with particular expertise in data management, evaluation design, and statistical skills. The advancements in computing technology have enabled providers/agencies to collect and report on service use in a cost-efficient manner. States, as well as the federal government, are requiring service providers to send them administrative data from programs at an increasing level of detail for purposes of funding accountability and for monitoring the effectiveness of programs. These data are frequently used for measuring performance, patient outcomes, and quality of care (i.e. Healthcare Effectiveness Data and Information Set, or HEDIS, measures). A major challenge of these data are the interpretability, coherence, and accuracy or quality of data items that are being integrated across programs and longitudinally over time.

Although there are many challenges using administrative data, there are also great opportunities. Because these secondary sources of information are already being collected for other purposes, they are relatively inexpensive to use for evaluation, especially in longitudinal studies that track individual patients over time and across providers (Motheral & Fairman, 1997; Quam et al., 1993) They are also a source of information on a large number of cases lending greater power for purposes of statistical inference (Motheral & Fairman, 1997; Garnick, Hendricks, & Comstock, 1994; Lohr, 1990). This makes them valuable in conducting population-based studies, detecting variations in practice patterns, and identifying specific

quality of care problems and health disparities that warrant further investigation (Iezzoni, 1997; Ballard & Duncan, 1994). Additionally, they are beneficial in studying low prevalence disorders, such as schizophrenia, or rare events, where there is high service use and costs for a small percent of the population. When records have at least one similar personal identifier which is unique (e.g., social security number; first and last name), they can be readily linked and aggregated across organizations and systems to build a comprehensive-client level history that should be useful in treating patients with chronic co-morbid conditions receiving treatment in different facilities or programs. They can also be used to monitor inappropriate drug utilization by clients and questionable provider prescription patterns.

Although there are advantages to using these types of data, limitations and challenges exist with respect to access or acquiring large administrative data files, data management, data integration, and, most importantly, data quality issues that are present at each step of the process. In order to determine which secondary sources have strong value, the user must consider the scope, authority, and audience that the information is based on (Cooper & Schindler, 2006).

This paper provides an overview of "data quality" issues and the factors associated with them. These attributes will be applied to administrative data, with a focus on enrollment or eligibility and service claims data used in monitoring the cost and utilization of health care services. Personal health records are briefly discussed; however, they involve other quality assessment strategies not explored in detail in this paper. Issues of accuracy, comprehensiveness, and validity will be discussed and recommendations will be made for using administrative records given the current state of quality found in these data systems.

Secondary Data

Using secondary data for administrative, reporting, or research purposes entails multiple activities of which assessing and insuring quality is a major factor. There are, however, other essential activities that are interrelated with data quality issues that are integral to the process.

- **Data diagnosis** involves initially assessing the data to understand its quality challenges. Data Profiling refers to inspecting data for errors, determining inconsistencies, checking for data redundancy, and completing partial or imperfect information. Profiling also includes a clear description of who the sample population is and the representativeness of the records to the universe that is being captured. Before any data set can be used, the number of records per quarter or year should be examined and compared to other administrative reports or other years of data. If record or person numbers differ, further discussion is required to understand the source of the discrepancies. Additionally, frequency distributions of all variables should be examined to assess missing values, incorrect codes, outliers, etc. These records should be corrected or in some instances set aside in certain types of analyses. Duplicate records should be removed based on a predetermined set of criteria of what constitutes an exact replica. For example, numerous records for the same hospital stay are sometimes found in a data set based on billing practices of hospitals. These records, if not aggregated, can reflect multiple episodes for an individual. This problem sometimes occurs for residential treatment programs or other long term facility stays.
- **Data Integration** is the process of matching, merging, and linking data for a wide variety of sources from disparate platforms. Matching or linking is a way to compare data so that similar, but slightly different, records can be aligned. Matching may use "fuzzy logic" to find duplicates in the data. For example, it often recognizes that 'Bob' and 'Robert' may be

the same individual. It might also find links between husband and wife or children at the same address. Finally, it can be useful in building a composite record, taking the best components from multiple data sources, and constructing a single super-record. For example, the most frequent name may be chosen as the true name when there are multiple records that should have "same" information for an individual. *An example of linking data like mental health and substance abuse treatment for a single individual being treated in different systems is illustrated using a program known as "Link King".*[1] *This program can be used for probabilistic links when all identifiers are not available and for deterministic links when identifying information is of good quality.*

- **Data Augmentation** is the process of enhancing data information from internal and external data sources and involves the addition of any piece of related data. *Examples, such as geocoding for name and address, can match data to US and Worldwide postal standards; phone numbers; contact information; common system wide identifiers associated with a case number at an agency, etc. all represent augmentation practices.*

- **Data Monitoring** is making sure that data integrity is checked and controlled "over time". Monitoring involves identifying variations in the data that require examination as to the cause. Software, based on certain algorithms, can be used to auto-correct variations if an error is involved. If the result is not inaccurate data, further exploration is required to understand changes in patterns due to policy or reimbursement or organizational changes which are not quality related.[2]

---

[1] *Link King:* http://www.the-link-king.com

[2] *For example, the Agency for Healthcare Research and Quality (AHRQ) and the Health Resources and Services Administration (HRSA) have an initiative to monitor the health care safety net* (http://archive.ahrq.gov/data/safetynet/billings.htm). *The monitoring tool, "Tools for Monitoring the Health Care Safety Net," aids administrators and policy makers in assessing local health care safety nets. It can be used to*

Finally, the process of ensuring **Data Quality** is essential to all of the above components. This process needs careful consideration because any mistake can create errors along the way in all of the other activities.

Definition of Data Quality

Data quality is the perception of data's fitness to serve its purpose in a given context. Data are considered to be of high quality "if they are fit for their intended uses in operations, decision making, and planning" (Juran & Gryna, 1993). Additionally, the data are deemed of high quality if they correctly represent the real-world construct to which they refer. Data quality and data volume are generally negatively correlated in that as data volume increases, the issues of *internal consistency* within a database becomes more problematic, regardless of fitness for use for any external purpose. For example, in longitudinal databases or a database where there are many records for an individual within a set time period, a person's gender, race, and birth date (DOB) can often differ between records. The more often information on a person is re-entered, the more likely the probability that differences will be found due to entry miscoding, different people entering the information, and a host of other sources of potential inconsistency. Determining the accurate set of data elements becomes more difficult especially with high users of services.

Even though there is wide agreement on the need for good quality measures to assess the soundness of the data being used, most methods of assessing quality are generally ad hoc, informal and not very rigorous. The process of defining the way data quality is conceptualized and operationalized in an agency and the practice of writing down the methods being employed

---

estimate the size of uninsured populations, to present administrative information to policymakers, to help states assess the financial stability of their provider institutions as well to examine the health outcomes for the populations served.

to assure this task be successfully implemented is a first step in enhancing data quality.

Components of Data Quality

A recent report prepared in 2013 by the Inter Agency Subcommittee under the Federal Committee on Statistical Methodology (FCSM) provided a list of items for administrative data that cover data quality concerns as well as a framework for assessment: The following section describes the data components that are essential for assuring data quality.

Accessibility is the availability of data in a warehouse and the ease of retrieval for monitoring, reporting, and analysis purposes. Accessibility of administrative data, whether within or external to a program or agency, is often a problem (Bright, Avorn, & Everitt, 1989; Potvin & Champagne, 1986; Ray & Griffin, 1989). Internally, problems of sharing and confidentiality often prevent data from being used even between divisions of an agency that has different departments. Furthermore, technical problems associated with record retrieval and record transfer frequently occur between divisions that sometimes serve the same clients through different programs. This occurs predominantly when systems are proprietary and not web-based.

Making data accessible to outside or external entities has similar problems as the internal ones. Most agency data has not been prepared for external purposes, thus policies and procedures for data access by outside groups are unclear. Additionally, there is a paucity of written documentation which can lead to potential errors in interpreting the data. Data dictionaries, needed by outside users to effectively analyze the data, are often inadequate. States, as well as other organizations differ in their approach to providing data to outside persons with a confusing set of rules for obtaining data files. Most data requests are individually negotiated; some occur formally through the development of a letter of agreement or memo of understanding (MOU) for a particular project; others occur informally through the development of a working

relationship between the evaluator or outside group and the data management administrators within the agency.

Once access has been negotiated, other challenges exist. Data retrieval has become more difficult with the use of complex data warehousing systems and staff that are often overworked or not permanent employees but contractors operating the databases. Outsiders requesting data items cannot simply request a flat file with a defined set of data elements. Whether it is the contractor or the internal staff of the agency, consultations regarding data specifications are required before requests for data extracts can be prepared. Unfortunately, these individuals are extremely busy and have little time to work with outside evaluators. Thus, individuals requesting data need to understand the structure of the numerous data tables in the data warehouses and be clear about their data specifications. The lack of adequate knowledge of where the data element exists in the data tables and the format required often results in multiple requests for data before the appropriate information is provided. Furthermore, if the specific data extracts requested by evaluators have never been analyzed by the state, there is no way of checking the accuracy of the results, which is perhaps the biggest challenge faced by outside evaluators. Finally, access or the availability of data for multiple purposes by a diverse group of users is extremely important as it increases visibility of potential errors resulting in better quality.

Security or Confidentiality is the degree to which data is properly protected. HIPAA regulations and confidentiality concerns about the use of personal data that has health care information and individual personal identifiers have increased dramatically. This requires greater effort in insuring that data are secure and results in the need for proper authorization and documentation to access the information. In recent years, when requesting secondary data records that were created for administrative, not clinical purposes, attorneys at both the

requesting agency and the provider agency are involved in writing memos of agreement for protection of the data records. Increasingly, this creates a delay before data sharing can occur, and can make the data less relevant for operational or policy decision making.

Relevance is the degree to which data accurately address the objective of the project or analysis. A data element should provide information at the appropriate level of aggregation and temporality needed to answer the question of interest. For example, administrative data that report information at the agency or organizational level may not be appropriate in assessing a program level issue. To be relevant, it is necessary to know which inputs (staffing, service type and volume, etc.) are associated with which outputs. Thus, data elements must be positively correlated or directly related to one another.

A data element may provide only partial information on a program activity or service cost. For example, when there are multiple funders in the delivery of services, the data element generally will only provide information on what the entity itself contributes versus what the actual contribution may be. For example, 30 minutes of treatment at a cost of 100 dollars may be only 50% of the actual treatment time and cost that is provided to a client if the data element does not capture the complete time and cost. **Interpretability** of data requires that a data dictionary exists to ensure meaningful use of the data items. Good documentation is essential for all variables that are to be used in data analysis. Data dictionaries that are current and clearly worded are essential in understanding the meaning of the information being used. When possible, especially when integrating data from different organizations, the source of the data is important to know. Is the information self-reported or, in the case of clinical information, from a professional? Does it come from other documents? The quality issue in this case is whether the variable(s) can be assumed to represent the same information. Variation in the definition of like

variables in an administrative record may be the result of differences in localities or regulations and clarification is needed to ensure correct interpretation.

Meaningful interpretation also requires an understanding of the target population that the data are representing. This means a clear definition of who the individuals are and the percentage of that particular population that the data is collected on. Generally, different agencies have jurisdiction over partial population groups[3]. Interpreting findings from individual data sets on partial populations versus comprehensive population surveys or the universe of a population will vary greatly.

Accuracy/Coherence are related concepts pertaining to data quality. **Accuracy** refers to the comprehensiveness or extent of missing data, performance of error edits, and other quality assurance strategies. **Coherence** is the degree to which data item value and meaning are consistent over time and comparable to similar variables from other routinely used data sources.

Data accuracy, reliability and validity of measures represent another area of quality concern. The most common errors generally found are coding and input errors (Abowd and Vilhuber, 2005). Coding accuracy and completeness with respect to service dates (when investigating the time sequence of events) must be correct before using the information for reporting or analysis. Fortunately, these errors can be reduced through feedback reports and better training of staff that are responsible for the activity. Some agencies offer financial incentives to providers who may be doing the data input in the form of bonuses at the end of a contract year.

Inaccurate or contradictory diagnostic information across visits, missing records, and lack of information regarding the severity of illness may confound analyses when data are used for

[3] i.e., the uninsured versus those receiving publicly funded services through state programs or through Medicaid or Medicare

program evaluation (Motheral & Fairman, 1997; Rosko, 1988; Sena & Pashko, 1993). To establish utility of this information for performance or evaluation, traditional reliability and validity assessments should be carried out on administrative files. For example, clinical chart data can be compared to administrative files, or different administrative files can be compared to each other. The following studies document differences in the reliability of data sources as follows:

Agreement on diagnostic accuracy of clinical chart data has ranged from 54% (Schwartz et al., 1980) to 100% (Walkup, Boyer, & Kellerman, 2000). Lurie, Popkin, Dysen, Moscovice, and Finch (1992) found that over 86% of clients with a diagnosis of schizophrenia in their administrative record had clinical information supporting that diagnosis. More studies of this type are needed as the reliability and validity of these databases are established for clinical data elements.

Agreement between different administrative data sets range from 74% to 96% (Parente et al., 1995). Agreement between dates of service provision often vary between data sources. The percent agreement was 67.1% for the date a case was opened using a provincial database in Canada (Robinson & Tataryn, 1997), but was as high as 99.2% for admission rates in a study that used Medicare records (Demlo, Campbell, & Brown, 1978).

Accuracy of data is enhanced when data are used extensively and fed back to providers, managers, and policymakers. Policy analysis using reliable data can be further enhanced when data are linked. Studies that link data sets allow for combining patient, organizational, community, insurance company, and provider information into one analysis (Rothbard et. al, 1990). Linkages between Medicaid records and vital statistics (Bell, Keesey, & Richards, 1994) allow utilization, birth, and death data to be tied together in outcomes studies. A study by Coffey

et al. (2001) demonstrates the process of linking state mental health, substance abuse, and Medicaid data. This linkage study involves the collection of information from disparate systems and funding streams which offers important opportunities for policy analyses and client outcome studies. The purpose of linked data is to create a comprehensive dataset from various sources that give a more complete picture of an individual's service history. Unfortunately, any weakness in the accuracy of elements from an individual dataset carries over to the linked data. When a linked data set has variables that do not match, information is lost for those data elements.

The types of practical problems involved in working with secondary data that has missing elements can sometimes be addressed methodologically by incorporating advanced statistical and econometric methods that capitalize on the longitudinal nature of data, as well as the large sample sizes inherent in this type of data (Fink, 1998; McDonald & Hui, 1991). An example of this was a study using maximum likelihood estimation survival models to determine community tenure, as well as the patterns of care in and out of state hospitals, over a ten year period (Stern, Merwin, & Holt, 2002).

Timeliness is the degree to which data can be used in a suitable fashion resulting in the capability to link information that is temporally related. The lag time between when data is collected and when it is available for analysis can affect the usefulness of the data for management and evaluation purposes. Equally difficult is the problem that occurs when data require correction or health claims are denied and then re-instated. Often Medicaid and Medicare data are 90 days behind as they are related to billing cycles. Additionally, these records often require changes in codes for payment, and new replacement records may take several months to be accepted by payers. Removal of original records is required to deal with

duplication. Although service records that are not tied to reimbursement are more readily available, they often are not as reliable or accurate.

Administrative data collection is not ordered for evaluation purposes, thus a person may have their data collected (e.g. housing status, level of functioning) before or after an intervention of interest to an evaluator or policy maker. This often results in large variations in time between baseline information, intervention start up and follow up for subjects. Interpreting the results of analyses using administrative data is challenging especially when information is collected at different time periods and is not comparable for all subjects.

Addressing Data Quality Challenges

Data Management is a broad term that refers to how data is structured or organized in a file, how it is stored (medium used), and what methods are used to protect it (firewalls, backups, encryption). Data quality is greatly affected by the way data is "managed," how accuracy is verified and consistency of information is addressed. The procedures and practices that support these processes must be well articulated and valued within an organization. Quality assurance methods are required to verify the accuracy of collected data. Data must be maintained by regularly examining the information through diagnostic analysis, "cleaning" the data that falls out of the boundaries, as well as unduplicating records and ensuring that the data elements are standardized so that all data elements report the same item in the same way. This requires feedback reports to providers comparing their estimates with others in the system, as well as record reviews done on charts to check for similarities. Discrepancies, other than minor ones, signal the need to assess the input and output data processes to determine the source of the differences. Service visit and episode counts, client counts, and descriptive data on the

population of interest should be consistent with routine planning or monitoring reports when doing an evaluation.

Good data management practices require up to date and detailed data dictionaries, data models, information on how data and process flows within and between organizations, detailed specifications, regular audits and controls and encryption methods.[4]

Oversight is another important component of good data management processes. Good oversight involves a data steward, data custodian, or data management task council that oversees the data management decision-making process. This is important when agencies are involved in sharing or exchanging data. Such stewards have responsibility for making sure that data elements have clear and unambiguous definitions, duplicates are eliminated, values are clearly enumerated or coded, and documentation is sufficient to allow suitable usage.

In sharing data within or between agencies it is important to have solid data management procedures with defined quality practices built into the process from the very beginning. Unfortunately, many systems have the quality built in much later, after the cost of correction in time and service improvement makes the changes necessary. The important lesson is to have quality engineered into every phase of the data management process from the start in order to avoid the costs of correction and failed decision-making. Data elements must always be added, refined, or amended and reconstructed, particularly when used in longitudinal trend analyses, as the purpose of data changes over time to address new reporting and monitoring needs as well as new services.

McDonald & Hui (1991) offer a useful review of data management and methodological problems faced by “researchers” or analysts using large databases. Computer requirements for

[4] http://www.techterms.com/definition/data_management; http://www.iso.com/Research-and-Analyses/ISO-Review/Data-Management-and-Data-Quality-Best-Practices.html

storing and manipulating these large files prove challenging. The service records can be difficult to manage physically as claims systems are organized on a transaction basis with billing and payment records for each reimbursable procedure. Several records may exist for the same service (i.e., when payment claims are denied and then re-submitted) and must be unduplicated for analysis purposes. Adjustments must also be made when there are multiple records for a single episode of inpatient care that spans months. Procedure codes can change over time, new services with a different name but similar function can create tracking difficulties in monitoring care. Finally, data storage and protection can be costly due to the confidential nature of these records. Data management tasks involve careful attention and continuous vigilance when using administrative data, especially when data quality is an important priority.

Verifying Data Accuracy. Management information staff must engage in multiple activities, both externally with those collecting and entering the data, as well as internally with those individuals storing and analyzing the data for planning and policy purposes. Providers should be trained in data entry procedures and should have standardized definitions of all data items. To confirm that the data being collected from external sources is correct, software applications should be developed that check all data fields for formatting errors, field type and size, missing data, and checks for valid codes. Additionally, when systems only collect admission and discharge information for outpatient programs, missing records on client discharge or disenrollment from a program can lead to incorrect length of stay information. This is problematic when monitoring performance in systems that are not claims or event based. To address this, data submission reports can be generated for the data provider with total number of errors in each field and a total percent of accuracy as well as identification of outliers with questionable length of stay.

The provider can be asked to submit a corrective action plan by a specified deadline that details steps for correcting the data before their next submission. A follow-up email and/or phone call can be made to the provider if they did not meet the deadline and/or if not all corrections were made. This type of information should be made available to other providers or data system integrators so they are aware of the limitations of the data.

Retrospective audits are another approach for verifying accuracy. When possible, checking all or a sample of the data against the original source is useful. When the accuracy of the measure you are planning to use is found to be poor, consider dropping the variable or using a proxy measure in its place. For example, use the history of substance abuse treatment as a proxy for co-morbidity if the records for drug and alcohol use are not identifiable due to confidentiality issues.

Verifying Data Consistency Once data is considered accurate, several internal tasks are required to enhance data quality. Data cleaning practices are essential. This involves eliminating duplicate records, resolving differences in data elements among multiple records of the same individual or event in the same data base (date of birth, gender, diagnosis, procedure), resolving inconsistent data elements across data bases when linking more than one data source (age, diagnosis, et al), cross walking data elements over time within and between data sources, and constructing a new data element that is comparable within, between, and longitudinally over time.

When there are different values for data elements, which should be similar, decisions need to be made to rectify the inconsistencies. Standardized rules should be developed and implemented to insure the integrity of the data elements that are immutable (date of birth, ethnicity, etc.). For example, this could include choosing the most frequent value, choosing the

value found on the most recent record, etc. and constructing a variable in all the records that reflects this decision. Alternatively, this could consist of choosing the record or measure that is perceived to be most accurate based on its source of information (i.e. in reporting of demographic information found in a death certificate record).

Variation in data elements frequently occurs in monitoring events over multiple years within the same data source. Changes in data systems often result in new coding schemes for the same variable. Additionally, what is kept in warehouse data tables frequently changes. For example, in the first case, a variable such as case management may take on different forms and meaning over time, and the new data element may specify the inclusion of information that was formerly in a separate data element. The change in the specification of the variable may require recoding or constructing a new variable for prior time periods.

Dealing with data consistency occurs again when integrating information across systems. Variation in data element definition requires the **"**cross walking" of data elements if data sources are to be properly integrated. This process can be tedious as it requires clear definitions of the data element for each system and an ability to reconstruct data elements when the information differs. Information generally needs to be aggregated to a higher or more general level when there are large discrepancies in the variable definitions. For example, some agencies may differentiate the type of outpatient programs they provide to a greater degree than others. Thus, a variable such as community outpatient services may be used to describe any program that is outpatient or ambulatory in nature. In integrating data across systems, more subtle information is sometimes lost in an attempt to be comparable as well as comprehensive in capturing all service data. Another example is when age is used in one system and date of birth in another, or when race categories differ, with some systems being much more specific than others. Information on

service volume or amount of units of care also varies and must be reconciled before integrating information. In some cases, the integrated file may only have the presence of the visit or episode of care and not the specific intensity of care, otherwise comparable volume issues cannot be resolved.

Tools to Assess Data Quality

Several quality assessment tools have been developed to provide a roadmap for those MIS staff involved in working with administrative data. In 2013, the Inter Agency Subcommittee under the Federal Committee on Statistical Methodology (FCSM) issued a report that focused on quality assessment (QA) and a tool to use that has a framework in developing a metric. The QA tool provides the user with a structure for requesting data that will be appropriate for the purposes of the designated project and of known quality. Three stages of assessment are described which reflect the needs of the user at different stages in the process. During what is called the Discovery Phase, the user is provided with a series of questions that explore the relevance, accessibility, and interpretability of the data in order to write a Memorandum of Agreement (MOU) to obtain the data. The Initial Acquisition and Repeat Acquisition Phases follow. The dimension of relevance is no longer an issue in these stages, but the issues of accessibility, interpretability, coherence, accuracy, and institutional environment are more prominent. An Appendix of Questions and a Data Dictionary Template are provided in the publication to give the reader a framework to follow in measuring the quality of the data they are using (Iwig, Berning, Marck, & Prell, 2013)

There are also many data tools available to implement the tasks related to translating raw data from external sources to a data set that is accurate. These data tools offer a series of steps for improving data which may include some or all of the following issues discussed previously:

data profiling, data augmentation, matching, parsing and standardization, and data monitoring. A number of vendors make tools for analyzing and repairing poor quality data available in-house. Service providers can clean the data on a contract basis and consultants can advise on fixing processes or systems to avoid data quality problems initially or once they occur. ISO 8000 is the international standard for data quality (Benson, 2008; 2009). The tool is used to do (1) data profiling, which is the diagnostic approach to determining data quality; (2) standardization of the data so that the elements are similar within and between records; and (3) geographic coding devices for provider and client data using US postal standards. ISO 8000 is being developed by ISO technical committee TC 184, Automation systems and integration, sub-committee SC 4, Industrial data.[5] However, like other ISO and IEC standards, ISO 8000 is copyrighted and is not freely available.

MIT has a Total Data Quality Management program, led by Professor Richard Wang, which produces a large number of publications and hosts a significant international conference in this field (International Conference on Information Quality, ICIQ) .[6]

The United States Health Information Knowledgebase (USHIK) is a metadata registry of healthcare-related data standards funded and directed by the Agency for Healthcare Research and Quality (AHRQ) with management support in partnership with the Centers for Medicare and Medicaid Services. AHRQ provides and maintains this metadata registry of health information data element definitions—values and information models that enable browsing, comparison, synchronization, and harmonization within a uniform query and interface environment. The U.S. Health Information Knowledgebase is populated with the data elements and information models

[5] http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/

[6] Wang's books on information quality include Quality Information and Knowledge (Prentice Hall, 1999), Data Quality (Kluwer Academic, 2001), Introduction to Information Quality (MITIQ Publications, 2005), and Journey to Data Quality (MIT Press, 2006).

of Standards Development Organizations (SDOs) and other healthcare organizations, in such a way that public and private organizations can harmonize information formats with existing and emerging healthcare standards. USHIK employs a metadata registry methodology based on international standards in order to promote interoperability and comparability. USHIK is housed and funded by the Agency for Healthcare Research and Quality with CMS and VA as strategic inter-agency partners.

Regardless of which tools are used, the major issue involves communication and coordination within and between organizations if data quality is to be promulgated. First, the source of data for each agency must be considered with respect to its accuracy. Second, the process of collecting and inputting data should be examined, and points at which data can be compromised should be documented and monitored. If data is to be used across agencies and systems, data standardization should occur where definitions and coding are made to be similar. A crosswalk algorithm should be developed when elements differ sufficiently to warrant aggregation of data information for purposes of integration. A data steering committee can be used to discuss and set interagency standards of definition and measurement or coding.

Uses of Administrative Data in Research and Evaluation

The challenge of using administrative data for purposes other than the specific one(s) they were created for (i.e., reimbursement, descriptive statistics for reporting requirements, etc.) requires knowledge of the data structure, meaning, and quality by the parties providing the data and those using the data, particularly when these are different groups of people. A case example is presented below that illustrates the various components of data quality that need to be addressed to do an evaluation study that includes secondary data from multiple sources. In this example, university researchers used data from several different agencies to create a data set that

could answer questions on the impact of restructuring community based services for persons with serious mental illness. The example is taken from a state which had reached a settlement agreement with the Department of Justice (DOJ) related to the Olmstead Act. This law requires that individuals with mental disorders be treated in an integrated community setting that is considered the least restrictive environment that meets their needs (Bazelon, 2010; Department of Justice, 2012).

The study population of 8,000 individuals consisted of persons with mental disorders that met criteria for serious and persistent mental illness served by the publicly funded mental health sector. These included state and local services, Medicaid services, housing/homeless services provided by multiple agencies, and employment services. The subjects were identified from secondary data sources and came from various departments at the state level, as well as outside agencies that were not formally required to report to the state Office of Mental Health (OMH), and in fact had restrictions based on confidentiality issues, in some cases. The type of data that was required to evaluate the impact of the settlement agreement on community services involved the following information.

- *Institutional admissions and discharges from psychiatric hospital settings, jails, shelters.*
- *Number of people served in the community compared to those served in institutional settings.*
- *Length of stay of individuals in institutional settings*
- *Readmission rates, including number of days elapsed between discharge and readmission.*
- *Number of individuals, in both institutional and community settings, who are on waitlists to receive community-based services*
- *Medicaid dollars spent on community-based services versus funds dedicated to institutional services*

- *State dollars spent on community-based services versus funds dedicated to institutional services*
- *Community-based housing, determined by the existence of supportive housing programs and the number of housing vouchers and subsidies, available to consumers*
- *Access and effectiveness of comprehensive community crisis services based on number of people treated in these programs and the reduction in subsequent hospital admissions based on other alternative*
- *Presence of evidence-based practices, including Assertive Community Treatment teams, supported employment programs, and peer support services*

Multiple sources and types of data were needed to construct the above measures and required the creation of an integrated person level data file for the 8000 individuals who were the focus of the evaluation and intervention. The time period was a four year span that began in 2010 and went through 2014. The data records were comprised of admission and discharge records as well as annual continuation records from all treatment facilities that had contracts with the State Office of Mental Health. These records were not reimbursement claims but demographic and service information on who was treated, type and place of treatment, and prior treatment history. Information on amount of treatment for outpatient services required creation of a length of stay measure from admission and discharge records. Claims or encounter data were used from the Medicaid files for persons enrolled in Medical Assistance (MA). Arrest information on admission and discharge came from the Department of Corrections (DOC) and shelter data from providers that had contracts to provide services to people involved in the People Acting to Help (PATH) federal grant program that provided services to those who were homeless. Employment data came from the Bureau of Vocational Rehabilitation (BVR) and Labor Statistics (BLS) and vital statistics records from the Department of Health were used to access mortality data. The following case study is used to illustrate data quality issues (outlined in Section 4. Components

of Data Quality), and the challenges based on providing accurate information for decision making in real world settings.

Prior to the evaluators receiving data for this study, **access** to data required memorandums of understanding (MOU) between each of the agencies providing the data and the state OMH. The inter-agency MOU is often as difficult to negotiate as external data sources, despite the fact that the data resides in the same Agency or Division. Depending on the legal issues regarding data sharing, gaining access to data was often a tedious and time consuming process. This process, described in a previous paper in this series (Cutuli et al., 2011), is a necessary step before any data can be accessed. In the case of this project, MOUs and other legal documents had to be developed between criminal justice (DOC) and the OMH and between the Medicaid program and the OMH (Petrila, 2010). The BVR and BLS would only agree to reveal the aggregate number of people receiving the designated services, thus individual records were not available.

Once, however, legal issues were settled, the data providers had to be able to send the information in a secure fashion and in a format that was readable by the receiver. To ensure **security,** transfer mechanisms had to be created and the movement of the data files secured so that the information could be sent to the receiver. Multiple conversations were required between evaluators and various technical staff to certify that the records were sent safely and efficiently, and that the records were complete and unchanged in the transfer. Record and variable counts were required before and after the transfer as some data sets had administrative header records that needed to be deleted. This situation is common as data systems vary in their structure.

Technical problems associated with record retrieval and record transfer also occurred between divisions that served the same clients, as the information was kept in different systems.

The file sizes needed to be specified prior to receipt so that the sender and receiver could determine the best methods for dealing with the data records and make proper arrangements in advance with their respective computer system teams. Finally, decisions were needed on whether or not it was cost effective to have the agency or vendors, who also had agency data for other purposes, provide the data. For example, the Medicaid claims processing vendor for the state was approached to provide the data in a more user friendly, less costly, and time consuming process than the agency staff themselves were able to produce because of agency time constraints and their lack of software required to do the data retrieval from their own warehouse in an efficient manner. Once access and transfer issues were decided, comprehensive documentation had to be provided on how the data files would be protected for confidentiality purposes. The last step in this process involved the successful transfer and reading of a test file that comprised all the requested variables requested by receivers.

The next issue related to data quality was the **relevancy** of the records to the evaluation question. The analysis was being done to determine the impact of the settlement agreement on creating an integrated community system of care where individuals would be less likely to use inpatient and other emergency-type services. The time period for the project was 2010 through 2014, encompassing both pre and post the implementation of new services meant to address the problem. The OMH wanted to have the data on a person-level so services could be linked to the individual in the target group and followed longitudinally as new programs were implemented. This would allow program and policy makers to determine if the changes being made in their system were effective and for whom. This required information on service type, admission date, place of service, and tenure in a program pre and post entry into the target group.

The data sources being used and what they represented were, in this case, relevant to answering the questions as they were direct indicators of the use of new services by the subjects involved in the evaluation and the consequences or outcomes. The mental health treatment data were comprehensive at a program level (ACT team) and a service or procedural type level (group counseling, intensive case management, etc.). Some external data sources did not contain information on whether or not an individual had a mental health disorder (criminal justice or homeless shelter records). Thus, we could not identify individuals in those systems that might fit the criteria of the target population (SPMI) unless they could be linked with the records of individuals who were already in the mental health system. This meant that we missed potential at-risk individuals in those systems. Likewise, housing information on subsidized residential arrangements for persons who had serious mental illness and were not being served by the public mental health system was incomplete, as the mental health disorder identifier was not in the housing data base. Furthermore, there were many non-profit organizations and housing programs at federal, state, and local levels, other than the Office of Mental Health, that provided housing and supports to individuals with behavioral health disorders, but did not report who was receiving housing to any centralized state agency. The housing information used in this evaluation was limited to those individuals who were already known to the public mental health system and received their housing through the state mental health agency.

The homeless data came from an agency that provided mental health and other support services through People Acting to Help (PATH). This federal program, funded by the Substance Abuse and Mental Health Services Administration (SAMHSA), was contracted out to a local agency by the State Office of Mental Health (OMH). The data kept by the agency designated that an individual had received a contact associated with a homeless episode. Identifiers for this

PATH group were frequently invalid which made matching unreliable. Although these individuals were considered part of the target population, their information was incomplete and not particularly relevant for the purposes of the evaluation.

Another problem causing difficulty when looking for service patterns and intensity of care was **missing records** or the lack of discharge records from community outpatient services for clients that were no longer in service. Because the services provided by OMH were not in the form of a paid claim each time a person had a visit, which was the method used by Medicaid for services rendered, length of stay in a particular service required a discharge record with a date. For a significant number of people, this was not done at the time the person stopped coming for services. Instead, a discharge record was completed administratively to deal with the problem of people having admissions to multiple programs and no discharge from others. Data quality was thus compromised around the accurate intensity of care over time and length of stay in treatment received by the population

Another missing service variable was the use of an emergency room when a person was seen in a medical emergency department and then admitted to a psychiatric hospital, where funding was from Medical Assistance. The charges for the emergency room were incorporated in the inpatient stay episode, and a separate claim was not recorded. For Medicaid clients, evaluators found that a revenue code in the inpatient records could be used to determine if the admission was through the emergency room. Also, emergency use by those in the state-funded OMH system was not available due to confidentiality issues between the OMH program-funded hospital emergency crisis program and OMH. Only aggregate emergency use data was reported to OMH monthly.

Verifying the **level of error** or actual mistakes in the data sets required the matching of records from different sources on the same variable. The most difficult variable to verify was the individual identifier. Even though the state created a unique identifier for all individuals served in the state system that all providers were to access when admitting a client (state hospital, community psychiatric hospitals, Medicaid services, community outpatient services), multiple identifiers were found for the same person when matched with date of birth, gender, social security number. Names were inherently difficult to match as the spelling differences were great. Mistakes in identifiers were found within a department and between agencies. Using programs such as "Link King"[7] as well as manual examination helped eliminate the duplications and provided improved accuracy for identifying unique individuals within and between data sources. This process is labor-intensive and requires constant upkeep when constructing a longitudinal data file spanning several years from various sources.

Duplication of records has been mentioned several times in this paper and can lead to overestimation of service use if not corrected. Duplication of information in this case example was both within and between data sets. OMH collected information on inpatient admissions and outpatient care in facilities which also treated individuals with Medicaid insurance benefits. In this instance, the records were frequently duplicated when the two data sets were combined, requiring the removal of the second record. Additionally, Medicaid claims are sometimes denied and the newly submitted record can appear again in a new service file several months later. Thus, data needs to be continually checked for the presence of the same information over the time period of the evaluation. As a result, the data set on the target population required continual updating with respect to duplications and identifier issues. Multiple identifiers needed to be retained in order to be comprehensive in searching for services across the various data sets.

[7] http://www.the-link-king.com

With respect to **timeliness** of data, complete Medicaid claims are generally available three to six months following the performance of the service. New claims for the same service can be resubmitted months later and must be reconciled within the data set. The state data is checked for missing variables and errors in certain fields. The shelter data from the homeless program was extremely unreliable time-wise and had many missing fields. Also, associating service information that is continuously generated and up to date to housing status, employment status, and level of functioning that is generally collected at admission to a program, makes it difficult to construct temporally-related information.

In general, gathering data from multiple sources and matching it to create a comprehensive longitudinal data file that is "current", or timely, is challenging, even if data were timely and automatically generated and transmitted on a regular time schedule. Thus, the use of these types of secondary data for operational or management purposes is currently questionable. However, these data can be employed for monitoring systems and can provide a more complete understanding of access, use, and patterns of care within and across agencies. This can lead to program changes and better outcomes. A data warehouse and query system are being developed by the state to try to accomplish this so that individuals on the target list who were arrested, homeless, or discharged from an emergency room or inpatient housing setting are identified daily and follow-up can be more immediate.

Conclusion

Data can be a powerful tool in management decision-making, policy development, and improvement in quality of treatment and services for a population. The United States does not have a tradition of developing case registries for monitoring disease and identifying causes of disorders, as is found in other countries. Privacy issues are of utmost importance and create a

challenge in doing research using secondary administrative data files where individual information must be linked. Standardization of data is also uncommon, which results in each agency developing the data measures and values that are most suitable to their particular situation or need.

Despite the problems associated with using secondary data for evaluation purposes, these data are a rich source of information that can be used in a positive way to improve efficiency and effectiveness of service delivery in the social welfare system. The volume of secondary data being collected today and the computer capacity to store, retrieve, and manipulate large volumes of records using sophisticated queries and software provides an opportunity to do comprehensive evaluation studies that were not possible previously. However, the ability to link data and form comprehensive histories that can be used to answer pressing questions requires that the "quality" of the data being used is appropriate for the situation.

The integrated data system (IDS) concept is fairly new for the social sciences but has exciting possibilities for evaluation and planning. The data the IDS receives should be assessed for quality first by the depositing agency through the checklist of items reviewed in this article. However, the IDS must also do its own assessment to determine the extent to which the data received are good or not (e.g., elements with too many missing cases, duplications, inconsistent records). In cases where bad data is suspected, IDS staff will need to collaborate with their partners to bring this to the attention of the depositing agency, and work together to correct data production problems for data collected in the future. Working to address this can be a shared responsibility between the IDS and the depositing agency, although the role of the IDS might be to bring the problems to light and to assist the agency in discerning what they need to do to improve data production quality. The depositing agencies should be asked to provide

documentation of their data collection forms and the quality assurance checks they use for verification purposes so the IDS has a better understanding of their procedures and potential weaknesses.

When there is no way for an IDS to know the quality of the data, they must perform their own audits to see the extent to which the depositing source data match with other records believed to be valid. In all cases, the IDS should make available their audit functions and have written documentation of their data management processes—this includes their data verification processes, aggregation, and processing steps at all levels of the data production process. Methods of eliminating double counting within each point of service and across service organizations in the same time period, as well as identification of drop outs, lost to service, a person who died, etc., should also be documented. Limitations should be noted on interpretation of variables that may be an underestimate due to incomplete or incorrect data.

How much should we focus on the issue of quality and what level of cost is appropriate to ensure that the record systems being used are adequate enough to provide answers to questions posed by managers, planners, and policymakers? Probably the most crucial activity warranting an investment of resources is the issue of a unique identifier. Without the ability to link individuals across data systems, the IDS function cannot be addressed. Next, time and effort should go into developing cooperation among agencies in data sharing and agreed-upon methods of ensuring data privacy and security so that information on the same person can be integrated. Data sharing agreements should be in place to make this a routine procedure. Third, resources should be used to develop standardized measures, values, and definitions on a subset of variables that require matching so aggregation of information is correct. Members of the partnership must also invest time in developing and maintaining good data documentation to ensure the user(s) are

clear about who and what the records represent. Likewise, the IDS members must provide clear documentation of their processing and decision making in aggregating, eliminating, constructing new variables, etc.

How good the data quality is can be looked at both subjectively and objectively. The subjective component is based on the experience and needs of the stakeholders, and can differ by who is being asked to judge it. For example, the data managers may see the data quality as excellent but consumers may disagree. One way to assess it is to construct a survey for stakeholders and ask them about their perception of the data via a questionnaire. The other component of data quality is objective. Measuring the percentage of missing data elements, the degree of consistency between records, how quickly data can be retrieved upon request, and percent of incorrect matches on identifiers (same ID, different SSN, gender, DOB) are some examples. The improvement of organizational data quality requires performing subjective and objective assessments, determining causes of discrepancies, and making recommendations of the procedures needed to improve data quality. There is no simple prescription for good data quality. An on-going commitment to using the fundamental principles discussed in this paper is the best approach to achieve best practice (Pipino et al. 2002).

References

Abowd, J.M., & Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers. *Journal of Business & Economic Statistics*, 23(2): 133-152.

Ballard, D. J., & Duncan, P. W. US Department of Health and Human Services, & Agency for Health Care Policy and Research. (1994). Role of population-based epidemiologic surveillance in clinical practice guideline development. *US Department of Health and Human Services, Agency for Health Care Policy and Research. Clinical practice guideline development: methodology perspectives. AHCPR Pub*, (95-0009), 27-34.

Bazelon. (2010). *Community integration*. The Judge David L. Bazelon Center for Mental Health Law. Washington, D.C. Retrieved from http://www.bazelon.org

Bell, R.M., Keesey, J., & Richards, T. (1994). The urge to merge: linking vital statistics records and Medicaid claims. *Medical Care,* 32(10): 1004-1018.

Benson, P. (2009). ISO 8000 data quality — the fundamentals, part 1, *Real-World Decision Support (RWDS) Journal* 3(4).

Benson, P. (2008). NATO codification system as the foundation for ISO 8000, the International Standard for data quality. *Oil IT Journal*. Retrieved from http://www.oilit.com/papers/Benson.pdf

Billings J. "Using Administrative Data to Monitor Access, Identify Disparities, and Assess Performance of the Safety Net" in Billings J, Weinick R. Eds A Took Kit for Monitoring the Local Safety Net. Agency for Health Care Research and Quality. July 2003. Retrieved from http://archive.ahrq.gov/data/safetynet/billing2.htm

Boruch, R. F. (2012). "*Administrative Record Quality and Integrated Data* Systems". Actionable Intelligence for Social Policy (AISP), University of Pennsylvania.

Bright, R.A., Avorn, J., & Everitt, D.E. (1989). Medicaid data as a resource for epidemiological studies: Strengths and limitations. *Journal of Clinical Epidemiology*, 42, 937-945.

Campbell, K. M. The Link King. Olympia, Washington: Camelot Consulting. Retrieved from http://www.the-link-king.com/index.html

Coffey, R. M., Graver, L., Schroeder, D., Busch, J.D., Dilonardo, J., Chalk, M., & Buck, J. A. Mental health and substance abuse treatment: results from a study integrating data from state mental health, substance abuse and Medicaid agencies. Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, Rockville, MD; 2001 (SAMHSA Publication No. SMA-01-3528).

Cooper, D. R., & Schindler, P. S. (2006). *Business research methods*. (9th ed.). London, England: McGraw Hill Publishing Company UK.

Culhane, D.P., Fantuzzo, J., Rouse, H.L., Tam, V., & Lukens, J. (2010). *Connecting the Dots: The Promise of Integrated Data Systems for Policy Analysis and Systems Reform.* Actionable Intelligence for Social Policy (AISP), University of Pennsylvania. Retrieved from http://www.aisp.upenn.edu/wp-content/uploads/2013/05/Connecting-the-Dots-AISP-Version.pdf

Cutuli, J., Culhane, D., & Fantuzzo, J. (Eds). (2011). AISP Best Practices: 5 Commissioned Papers. Philadelphia, PA.: http://www.aisp.upenn.edu/best-practices/

Demlo, L.K., Campbell, P.M., & Brown, S.S. (1978) Reliability of information abstracted from patients' medical records. *Medical Care*, 16(12):995-1005.

Department of Justice. (2012). *Participation by the United States in Olmstead cases*. Retrieved from via http://www.ada.gov/olmstead

Fink, R. (1998). HMO data systems in population studies of access to care. *Health Services Research*, *33*(3 Pt 2), 741-766.

Garnick, D.W., Hendricks, A. M., & Comstock, C. B. (1994) Measuring quality of care: fundamental information from administrative datasets. *International Journal for Quality in Health Care*. 6(2):163- 77.

Huang, K. T., Lee, Y. W., & Wang, R. Y. (1998). *Quality information and knowledge*. Upper Saddle River, NJ: Prentice Hall PTR.

Iezzoni, L. I. (1997) Assessing quality using administrative data. *Annals of Internal Medicine*. 127(8):2; 666-674.

Iwig, W., Berning, M., Marck, P., & Prell, M. (2013). Data quality assessment tool for administrative data. Prepared for a subcommittee of the Federal Committee on Statistical Methodology, Washington, DC (February).

Juran, J. M., & Gryna, F. M. (1993). Quality Planning and Analysis (3rd ed.). New York: McGraw-Hill.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. Journey to Data Quality. (2006). *Cambridge, MA, USA: Massachusetts Institute of Technology.*

Lohr, K.N. (1990). Use of insurance claims data in measuring quality of care. *International Journal of Technology Assessment in Health Care*, 6, 263-271.

Lurie, N., Popkin, M., Dysken, M., Moscovice, I., & Finch, M. (1992) Accuracy of diagnoses of schizophrenia in Medicaid claims. *Hospital and Community Psychiatry*, 43 (1): 69-71.

McDonald, C. J., & Hui, S.L. (1991) The analysis of humongous databases: problems and promises. *Statistics in Medicine*.10:511-518.

Motheral, B.R., & Fairman, K.A. (1997). The use of claims databases for outcomes research: Rationale, challenges, and strategies. *Clinical Therapeutics*, 19:346-366.

Parente, S.T., Weiner, J.P., Garnick, D.W., Richards, T.M., Fowles, J., Lawthers, A.G., Chandler, P., & Palmer, R.H. (1995). Developing a quality improvement database using health insurance data: A guided tour with application to Medicare's national claims history file. *American Journal of Medical Quality*, 10, 162-176.

Petrila J. (2010) Legal Issues in the Use of Electronic Data Systems for Social Science Research. Actionable Intelligence for Social Policy (AISP), University of Pennsylvania. Retrieved from http://impact.sp2.upenn.edu/aisp_test/wp-content/uploads/2012/12/0033_12_SP2_Legal_Issues_Data_Systems_000.pdf

Pipino, L.L., Lee, Y.L., & Wang R.Y. (2002). Data Quality Assessment. *Communications of the ACM*. 45(4): 211-218.

Potvin, L., & Champagne, F. (1986). Utilization of administrative files in health research. *Social Indicators Research*, 18, 409- 423.

Quam, L., Ellis, L. B., Venus, P., Clouse, J., Taylor, C. G., & Leatherman, S. (1993). Using claims data for epidemiologic research: the concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population. *Medical Care*, 31(6):498-507.

Ray, W.A., & Griffin, M.R. (1989). Use of Medicaid data for pharmacoepidemiology. *American Journal of Epidemiology*, 129, 837-849.

Robinson, J.R., & Tataryn, D.J. (1997). Reliability of the Manitoba mental health management information system for research. *Canadian Journal of Psychiatry*, 42, 744-749.

Rosko, M.D. (1988). DRGs and severity of illness measures: An analysis of patient classification systems. *Journal of Medical Systems*, 12, 257-274.

Rothbard, A.B., Schinnar, A.P., Hadley, T.R., & Rovi, J.I. (1990). Integration of Mental Health Data on Hospital and Community Services. *Administration and Policy in Mental Health*, 18, 91-99.

Sena, M.M., & Pashko, S. (1993). Drug utilization review using a Medicaid claims database. *Clinical Therapeutics*, 15, 900-904.

Schwartz, A.H., Perlman, B.B., Paris, M., Schmidt, K., & Thornton, J.C. (1980). Psychiatric diagnoses as reported to Medicaid and as recorded in patient charts. *American Journal of Public Health*, 70, 406-408.

Stern, S., Mewin, E., & Holt, F. (2002) Survival models of community tenure and length of hospital stay for the seriously mentally ill: a 10-year perspective. *Health Services & Outcomes Research Methodology,* 2,117-135.

Walkup, J.T., Boyer, C.A., & Kellerman, S.L. (2000). Reliability of Medicaid claims files for use in psychiatric diagnoses and service delivery. *Administration and Policy in Mental Health*, 27, 129-139.

Wang, R.Y., Ziad, M., & Lee, Y. *Data Quality*. London: Kluwer Academic Publishers, 2001.