QUALITY METRICS FOR HEALTHCARE DATA: AN ANALYTICAL APPROACH

Rema Padman
The H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213
(412) 268-2180
rpadman@cmu.edu

Melissa Tzourakis
Medicode Inc.
5225 Wiley Post Way, Suite 500
Salt Lake City, UT 84116
(801) 536-1050
mtzoura@medicode.com

ABSTRACT

Data quality is a priority for health information management because of the many critical uses and users of the data. The computer-based capture and processing of various types of information ranging from patient records to claims data has heightened the need to assess and measure the quality of the captured data to better manage the data gathering and scrubbing processes. Completeness, correctness, consistency and timeliness are generally used to provide a framework for thinking about data quality. However, there has been little work on linking healthcare application-specific metrics to these more general concepts. In this paper, we report on an ongoing study to develop an analytical approach to address this problem. The perspective used is of an information processing organization that receives healthcare data from multiple, disparate sources. A healthcare claims processing example is used to illustrate the approach.

1.0 Introduction

Data quality is of critical concern to healthcare organizations that depend on computer systems to create, modify and utilize data. In this new era of managed care and integrated delivery systems, with its emphasis on improving cost efficiency without risking quality of care, the explosive advances in information technology has created the need for a methodologically driven analysis that integrates healthcare data requirements with the information technology capabilities. As data and information become key strategic resources, poor quality can have significant consequences on the ability of organizations to fulfill their mission [1, 3 , 7]. For example, survey data of state health databases indicate that the data quality is so highly suspect that they cannot be used for making policy decisions or conducting cost, utilization and related analysis [5, 6]. Thus, integrated healthcare organizations and their associated information systems face a major problem in ensuring the quality, integrity and validity of their data. While both manual and automated mechanisms are currently utilized to ensure that the data are reasonably accurate [4, 7], they are not cost-effective. Lack of methods and tools for measurement and evaluation of quality are factors. Consequently, there is significant potential for developing analytical methods that identify and categorize the classes of problems arising in healthcare data in order to facilitate the development of solutions for them. A fundamental prerequisite for such analytical work is measurement. In other words, how should the quality of healthcare data be measured?

Metrics such as completeness, correctness, timeliness and consistency are generally used to provide a framework for thinking about data quality [1, 2, 7, 8, 9]. While these are relevant to healthcare data, they need to be extended and modified to meet the needs of the application context. In this paper, we develop analytical approaches to address this problem from the perspective of an information processing organization (IPO) that receives vast amounts of healthcare data from multiple clients. Given the disparate, non-homogeneous and nonstandard data that are received, these organizations face tremendous challenges in converting this data into a usable product. Hence, there is a need for these IPOs to consider how data quality requirements should be stated, measured and benchmarked so that appropriate feedback can be provided to the data suppliers.

Informed by practice, the research described in this paper has significant value for health data organizations, public and private, as they struggle to satisfy reporting requirements mandated by state and federal regulations as well as by competitive market pressures. It facilitates the comparison of the quality of data from various sources prior to combining them. It also serves as a stepping stone to establishing benchmarks for data quality within the industry. Once these benchmarks are established, employers, regulators and other third parties can more easily measure the data they are collecting and determine prior to exhaustive analysis, whether or not a data set is of sufficient quality for the study or analysis proposed. Healthcare, like many other industries, is an information driven industry. It has invested billions of dollars in developing data warehouses and purchasing decision support applications, but has spent little time and few resources in ensuring that the data is at a level of quality that is appropriate for the requirements of these tools. This research can also be extended to explore whether a given measure can help an organization define what their information quality, or lack thereof, is costing them.

The paper is organized as follows. Section 2 details the specific problem that we address in this paper, followed, in Section 3, by the conceptual model and methodology used in developing and analyzing the specific metrics that are important in the healthcare environment for IPOs. Section 4 presents the results of our preliminary study and section 5 concludes with the directions for future research on this problem.

2.0 Background and Problem Description

Vast amounts of healthcare data are collected, aside from clinical reasons, for administrative purposes such as reimbursement, financial transaction and cost analysis, employer reporting, physician profiling and utilization analysis. Since the origin of healthcare data is an encounter between a patient and a healthcare provider, data sources generally include the claim forms that are completed at the time of the patient visit, the physician's notes, and the documented patient history. From each encounter, the provider, which may be a physician, hospital, laboratory, and so on, will record the service rendered (an office visit, lab tests, details of inpatient stay), conditions of service (the diagnosis, date, place of service), patient information (sex, age, patient history, insurance information), and clinical information (result of tests, prognoses, consultation

notes). All of this information is useful to the provider(s) and appears in a variety of forms, from paperwork to billing systems.

Some standardization exists in the way data is captured. Most of this has come about in recent years due to the requirements imposed by the government's Medicare program. When the provider submits a "claim" to the payor for services rendered to the patient, it has to include many data elements based on standard code definitions for the procedure or service performed and the diagnosis, in the form of ICD-9, CPT-4, and DRG codes. Once implemented, these coding rules form the basis for payment to the provider while concurrently being captured as reimbursement data by the payor. Data collection and submission thus becomes the responsibility of the provider, mostly clinicians, who find the administrative and reimbursement requirements burdensome. The complexity of correct coding, with its more than 8000 procedure codes and 16,000 diagnosis codes, creates further problems due to the annual change in the codes and their methodologies and, more importantly, the inadequate knowledge and training of the provider's administrative staff.

The necessity to contain costs and maximize productivity and value have forced healthcare providers and payors to turn to their data and decision support tools to validate their cost and quality initiatives. For health plans to effectively monitor activity within their networks, they must have an accurate picture of encounters that take place between provider and patient. Ideally, the correct data to analyze the effectiveness and quality of these interactions would be the original clinical data by the provider at the point of service. However, due to a lack of extensive automation and standard methodologies for collecting clinical information, this data are difficult to obtain. Consequently, the healthcare industry has had to accept the fact that available reimbursement or claims data are most representative of the encounter.

Unfortunately, encounter data can further degrade from the payor's handling of the claim. Data may not be captured accurately or completely. System maintenance files and insured and enrolled information may be incorrect, referral data may be missing (causing an incorrect denial), data may be keyed incorrectly, or a contract may not be set up correctly. In addition, there are many other points within a payor's claims adjudication system where the data may further lose

quality. Claims operations focus on the correctness of the financial payment that results from the claim. To do this, they use clinical editing systems that assess combinations of procedure and diagnostic codes, automatically notifying the entrant of a suspicious or incorrect combination.

However, recent changes in health plan analytical and reporting demands have forced organizations needing information on medical management, enhanced account reporting, and marketing, to build data warehouses of historical data. As users begin accessing this data, the enthusiasm to provide information has quickly changed to frustration. Many times, the numbers retrieved from the warehouse do not match other reports. Data is found to be incomplete and inconclusive and there is generally a lack of understanding by analysts as to the anomalies and inconsistencies in these data. Furthermore, because of the industry's use of new reporting and analysis tools, it is unclear if the quality deficiencies are caused by the tools, how they are implemented, or the data.

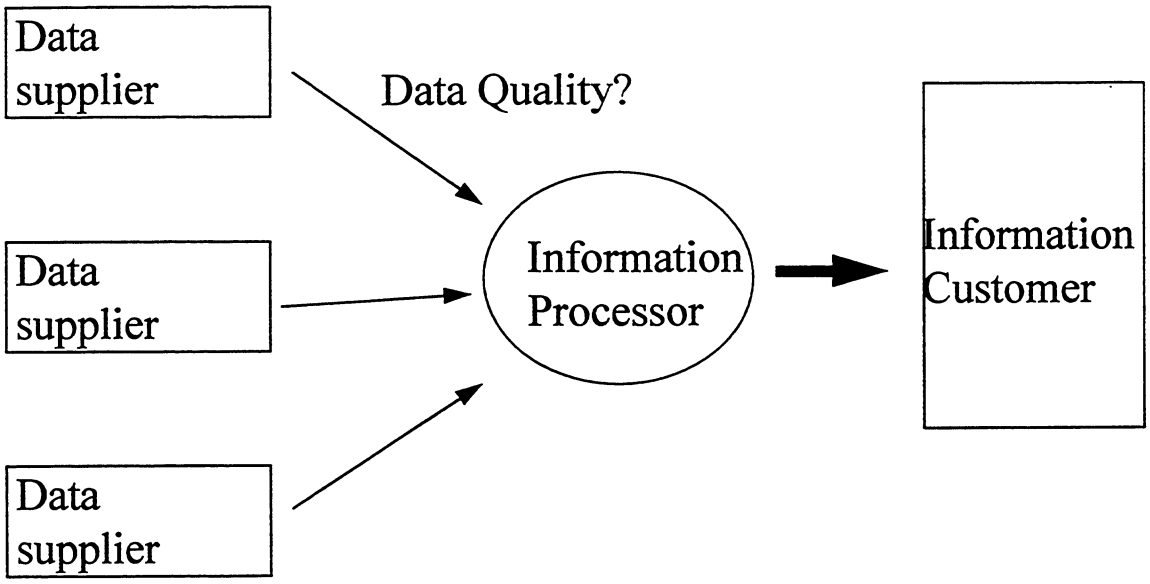

Figure 1. Supply Chain Model of Data

Our approach to this problem is based on a supply chain model of the data. Figure 1. illustrates the value chain in the processing of the data from suppliers to consumers. The information processor, receiving the data from several suppliers, is concerned about the management of the supply chain, with particular emphasis on the quality of the data. Setting

standards, providing feedback on the data quality and working with the suppliers in the continuous process of improving data quality are some of the solutions. In particular, given the large number of data elements about which data is collected, they would like to determine the critical attributes for which data quality should be measured. Also, drawing on the research on data quality metrics in the general literature, they would like to develop appropriate metrics specific to healthcare data. These metrics, if accepted by the healthcare industry, can be utilized for creating reports that can help identify quality problems and improve the usability of the data. Recognizing that overall quality of the data is a subjective measure based on the perception and requirements of a particular user or set of users, we combine the metrics into a single measure that is dependent on a ranking scheme of the data elements and metrics that is determined by the user. However, if the same priority ordering is used by a health data organization on the data received from multiple clients, our procedure enables a comparison of the data from different sources on a consistent and quantifiable measure of quality.

We illustrate this procedure using claims data from Medicode, Inc., a large healthcare data and information processing organization. Medicode's primary focus is on its clinical and technical expertise in coding, reimbursement, and statistical analysis of claims data for the healthcare industry. It is also a primary source of UCR (usual, customary and reasonable) pricing databases, containing more than 350 million records of data used by payor and managed care organizations to price healthcare services and adjudicate claims. As an organization that receives large amounts of healthcare data for processing, the availability of the data and the concerns for its quality provide a significant opportunity to validate our methods and their value to the healthcare industry.

3.0 Methodology

3.1 Conceptualization

In the following, we propose a formal framework for managing data quality and apply it using real data from a large healthcare IPO. To compute the quality of the data, we first need to address the following questions. What should be the unit of analysis along the data supply chain? How should the quality of the unit of analysis be measured? Should it be multi-dimensional

versus uni-dimensional? How should the measure be used to manage data quality along the supply chain? We argue that a unit of analysis should be the data set for the IPO since they receive data from suppliers at this level. At the same time, this unit of analysis also facilitates the ability of IPOs to provide evaluation and feedback to the suppliers on the quality of the data.

We thus propose the following definitions, illustrated in Figure 2.:

Unit of analysis: Dataset

Dataset: A named set of records R1,..., Rm

Record: A named set of attributes A1,...,Ak

***Attributes*:** Have domains and formats

Domain (A) -- Range of valid values for attribute A

Format (A) -- The format in which values of attribute A are stated

Records\Attributes A1 A2 A3

R1

R2

R3

Figure 2: Composition of a Data Set

3.2 Metrics

The metrics used to measure data quality also have to be defined according to the characteristics of the data set exchanged between the supplier and IPO. This definition thus pertains to the particular features of the data and application. In the following, we define these metrics using our formal framework and specify how we measure them. The application of these metrics to real data is described in section 4.2. We argue that while accuracy, completeness, consistency, and timeliness are necessary measures, given the proliferation of proprietary standards at all levels of healthcare data, it is very important to define and measure compliance

with open standards as an additional dimension of data quality. We discuss this in greater detail in section 4.1.

1. Accuracy: property of *an attribute in a record*
   Value accuracy: Value(A) = KV(A), where

   KV(A) is the known value of A from other independent sources. This is the way in which claims data is audited, as shown in Figure 3. This metric is, in general, difficult to measure since the original documents may not be available or the process may be manual and cumbersome and consume significant time and resources.

   Measurement: 0 if attribute A in record R is accurate, else 1.

Records\Attributes	A1	A2	A3
R1	0	0	0
R2	1	0	0
R3	0	0	1

   Figure 3: Example of Measurement of Accuracy

2. Compliance With Standards: property of *an attribute in a record*
   Value compliance: value(A) is in domain(A)
   Format compliance: format(A) is in F(A)
   F and domain are specified by standards (e.g., EDS or McData)

   Measurement: 0 if attribute A in record R is compliant, else 1 (similar to Figure 3).

3a. Syntactic Consistency: property of an attribute in a record
   An attribute may be accurate in format but may not be consistent throughout the data set. e.g., variability in the number of digits in the patient-ID.

Measurement: 0 if attribute A in record R is syntactically consistent, else 1 (similar to Figure 3).

3b. Semantic Consistency: property of a record

Captures validity of value fields in multiple attributes of a record. e.g., claims data record shows a 45 year old woman underwent treatment for prostate cancer.

Measurement: 0 if a record is semantically consistent, else 1 (Figure 4).

Records	Consistency
R1	0
R2	0
R3	1

Figure 4: Example of Measurement of Semantic Consistency

4. Completeness: property of a record

A record is not complete if an attribute value is missing.

Measurement: 0 if a record is complete, else 1 (similar to Figure 4).

5. Timeliness: property of an attribute in a record

An attribute is not timely if the standards used to assign its value are not current.

Measurement: 0 if a record is timely, else 1 (similar to Figure 3).

These definitions can be applied to attributes or records, as appropriate, to compute the level of error in each attribute/record on each data quality dimension. This information can then be aggregated using a simple rule to arrive at a quality indicator for the data set.

3.3 Procedure

1. Identify all the attributes in the data set
2. Develop a critical subset that can be used in the quality computation
3. Generate a random sample of the data
4. Compute $quality_i^a$ = (1- error rate) along each quality dimension i for each attribute/record a, where error rate = $\sum$ cell value in the table for i = 1,...4 over all the records
5. Aggregate quality of the attribute over all dimensions, $AQ^a = (\sum quality_i$ * number of records in sample)/(Number of records*number of dimensions for which quality was measured for that attribute)
6. Aggregate quality for each dimension, $AQ_i = (\sum quality_i^a$ * number of records)/ (number of records * number of attributes for which quality is measured on that dimension

Using this procedure, each data set can be assigned a quality indicator along each dimension or attribute. Data sets can be benchmarked and compared along these multiple dimensions. An alternative simple rule for computing aggregate quality may be as follows: Consider a record in error if at least one attribute is in error on at least one dimension of data quality. Then compute quality = % records in error in the sample. While this rule provides a single indicator of quality for the data set, it may overstate the poor quality considerably. In the next section, we illustrate our procedure with real claims data from Medicode, Inc.

4.0 Application: Sampling of Data and Identification of Critical Attributes

A randomly selected sample of 500 records for a single client were drawn from the Medicode Extended Data Set format (EDS). The duration of the database spanned three years, the records being selected from 1993-1996. The EDS format contains 35 attributes, ranging from provider information, service details, procedure and diagnosis information, and patient identification information. The next step involved reducing the number of attributes to those that were critical with respect to their use in reporting and analysis. This reduction was necessary to handle the complexity and increase the value of the data quality computation.

Twenty-eight attributes were selected as critical from the set of data elements in EDS format. These attributes were selected for several reasons.

(a) Criticality: The attributes selected represent those fields that are regularly collected by payor organizations, are readily available from the Health Care Financing Administration's (HCFA 1500) or Medicare's Uniform Billing (UB-92) forms or the data is typically stored within the claims adjudication systems. They were also selected based on their use and value in financial transactions and cost analysis, employer reporting, and quality of care, physician profiling and utilization analysis.

(b) Attributes used in Standards: There has been much effort by many parties to establish a standard data set. To date, 17 have been proposed. Of these 17, the one that will most likely be mandated at least for Medicaid and Managed Medicare reporting is HCFA's McData (Medicaid-Medicare Common Data Initiative Steering Committee) set. This is comprised of 28 elements. The elements currently not collected or submitted in the EDS set that are not measured in this study include discharge patient destination, days since admission, early and periodic screening, and diagnostic and treatment services indicator. Some of the elements such as date of service and admission/discharge date are represented by a single field in the EDS data set.

4.1 Applying the Defined Measures of Data Quality

As mentioned earlier, the four dimensions of data quality, timeliness, accuracy, consistency and completeness proposed in the literature [1, 7, 8] were the initial intent of the study. However, as we looked at the characteristics of healthcare data, several things became apparent. One, as the data was submitted by several different clients that covered a three-year time span, timeliness of the data could not be measured adequately from the perspective of the currency of data. Timeliness, as we have defined it, is from the perspective of the data supplier, and refers to the use of the most timely standard for coding methodologies. Coding methodologies, for example CPT codes and ICD-9 codes, change on an annual basis. If claims are submitted with deleted codes, they should be considered an error on the timeliness dimension. While this is a significant

component of data quality, we exclude this dimension in this study. In a follow-up study, we plan to automate checking the data with the coding reference standard to better determine the value of timeliness.

In addition, we use adherence to standards as another dimension of quality. Several industries have not only tried to establish standard formats for their transactions, but have also established standard reference data that allows for the codification of information. Healthcare has attempted to do this for many years and it is a continuing effort. Since 1983, HCFA has proposed standards around coding on the HCFA 1500 and UB-92 data collection forms. These standards include the HCPCS (HCFA Common Procedure Coding System) Levels I, II and III which codifies the services performed during an encounter. It includes the CPT procedural coding (Level I), HCPCS Level II which codifies supplies and non-physician services and Level III, which is variable as it codifies local standards at a state level for Medicaid or Medicare billing.

Standardization is a necessity when combining disparate databases for the purposes of comparison or aggregating analytical studies. Without the use, mandate or enforcement of standards, many of the recent efforts on the part of employers and regulators to utilize data in this way to measure and compare health plans will be successfully challenged. Compliance with existing standards, whether they are mandated or not, as a dimension of data quality will further efforts to compare and contrast encounter data across all healthcare entities. Adherence to standards also reduces or eliminates the need for the user to crosswalk data from various data sources. Cross walking is time consuming and much of it is manual because the coding systems are so variant due to their homegrown nature, and therefore expensive. In order to reduce analytical costs associated with combining and scrubbing disparate databases, the dimension of adherence to standards would be very important.

4.2 Critical Attributes and Their Errors

In the following, we briefly describe some of the critical attributes that were considered in this study and the dimensions measured for each. Due to the unavailability of the original claims, timeliness dimension was dropped from further analysis for several attributes. In

addition, the set of critical attributes also was reduced to 11 attributes from 28 because it was either not measured or not collected.

1. Rendering Provider ID - This field is often encrypted upon submission, to protect the physician from the dissemination of their Federal Tax ID number which is commonly used as a unique identifier within the health plans. HCFA is currently working on implementing unique physician identifiers (UPINs), but as of yet not been implemented and acceptance will be slow without a mandate. We also did not have access to the actual claim for reasons of confidentiality. This will also be the case when data sets are collected for decision support purposes. Therefore, on this field we measured for completeness of the attribute by noting its presence, or lack of it, in all the records. Consistency was defined by consistent use of the same format for numbering. As long as the provider can be uniquely identified, it is not considered inaccurate, but inconsistency in the length of field and the numbering system can lead to more minor problems around selecting general queries and having to account for field length variation.

2. The referring provider number was also measured in the same way. The biggest problem is the lack of completeness of the field in general by the clients studied. Many times this information is not always submitted and since it was not necessary for financial transactions, many systems do not collect or require collection of the field.

3. Provider specialty was measured for completeness, consistency, accuracy and adherence to a standard. The HCFA standard is most accepted as the de facto standard, but is also not mandated. The client considered in this study did not utilize the standard which resulted in low scores on this field. Also, there was inconsistency as to the collection of the field and many were blank. This field is critical in that it allows data to be separated by specialty which allows for more specific analysis on utilization by specialty since utilization patterns within a specialty are more comparable.

4. Patient ID was difficult to quantify on quality as a stand alone field. Often, it is encrypted since the ID within the payor systems utilize the SSN. The critical aspect of this field is that it allows the analyst to differentiate data specific to one patient or an episode of care for one patient. To quantify the quality of this attribute, we looked at the relationship of this field with the date of birth (DOB) and sex fields to ensure that the number was unique to that patient (i.e., for each unique ID number, the DOB and Sex should be consistent for all records of that ID).

5. Claim ID identifies the claim submitted. It allows the user to identify and group services as submitted by the provider. It is also used in reimbursement. This is often an internally generated code by the payor. We have quantified its quality based on two factors. One is by comparing with the physician ID to ensure that the services represent those performed by only one physician or provider. This ensures proper reimbursement as well as analysis by claim. We also ensured that there was consistency in the field length as well as numbering of the claims. The "from" and "to" dates cannot be measured in terms of accuracy as to the dates representing the true date of the service without referencing the submitted claim or encounter record. We did ensure that the dates fell within the time frame requested (the three year submission) as well as ensuring that the "from" date predated the "to" date.

6. The Place of Service (POS) was measured for completeness and adherence to standard (HCFA POS code). Value accuracy was difficult to measure without access to the actual claim form. We also measured the use of POS consistency.

7. Procedure Code was measured for completeness, since it is always required, as well as consistency and adherence to standards. We expected to see the HCPCS Levels I, II and III in this field. The client had also submitted the UB-92 related revenue codes in this field even though that field was separate on the critical elements list. Accuracy again was difficult to measure as we did not have access to the claim form. The procedure field is critical in identifying the service(s) performed during an encounter.

8. The Primary and Secondary Modifier fields were not utilized by the client. This field was not expected to be complete in all instances since the modifiers are associated with the procedure code field and the standard values are set by the same entities establishing the procedure codes. This attribute is critical for reimbursement as well as in understanding any extenuating circumstances associated with the service(s) provided. It can identify numerous situations such as multiple procedures, co-surgery or rental or purchase of an item.

9. The Units field identifies the number of services provided. Multiple services can be represented with one code on one record, given that units are applied, rather than on separate line items. This was measured for consistency, accuracy and completeness. Standards are not applicable for this attribute. All line items would be expected to have, at a minimum, one unit. This was not always the case for the sample analyzed. The client submitted a 0 for many of the fields which creates a problem when trying to use this field to calculate values in doing any frequency or utilization analysis.

10. The Diagnosis field is also very critical in that it identifies the condition of the patient and the diagnosis rendered at the time of the visit. ICD-9 codes are utilized as standard values for this field. This field establishes medical appropriateness of the service for reimbursement as well as allowing the user to analyze medical conditions across the data sets to establish costs and utilization by medical conditions. The additional diagnosis fields are not always utilized since some patients have a single diagnosis at the time of the visit. But one would expect to see the additional diagnosis fields utilized to some degree.

4.3 Results

Table 1 summarizes the aggregate quality along each data quality dimension and for each attribute computed using the procedure outlined in section 3.3. The first two columns list the names of the critical attributes in the two standards used in this study. Columns 3 and 4 report which of these attributes were either not measured or not collected. Columns 5-8 display the

quality for each attribute on each dimension, and the last column reports the weighted quality measure AQ_i^a. The final row of the table summarizes AQ_i.

Based on a sample of 500 records, we note that more than 25 percent of the records have quality problems on every dimension. For the attributes, quality varies considerably, from perfect scores for attributes such as dates of service to poor quality for provider specialty. Given the extensive use of procedure and diagnosis codes by the industry, it is not surprising to see higher levels of quality for these attributes. The net summary may be that data used for reimbursement purposes, such as these codes and dates of service, can be considered to be of reasonably good quality. However, the data that is necessary for decision support purposes, such as provider specialty, still have a long way to go to reach acceptable levels of quality.

5.0 Conclusions

This paper summarizes our preliminary research on the measurement and computation of data quality for large healthcare data sets from the perspective of an information processing organization. Such organizations, receiving vast amounts of data from multiple suppliers in varied formats, adhering to different standards, using disparate nomenclatures for recording encounter and claims information, face significant difficulties in cleaning this data for further use. As the number of data suppliers increase, along with increased computer-based capture of their data and proliferation of standards in the absence of national standards, mandated or otherwise, the complexity of the data quality problem and its management challenges increase exponentially. We propose a structured, quantitative approach to addressing this problem by identifying critical attributes and defining the rules by which data quality can be measured. The major challenge now is to develop the appropriate methodologies to combine the multiple dimensions used in this study into a single value that can be used to compare data quality across many data sets. Future studies will also lead to the development of rule sets that will allow automation of the procedures for computing quality.

McData Set	EDS	Measured	Not collect	Completed	Accuracy	Consistenc	Standard	Total
Recepeint ID	Patient ID	Yes		1	0.738	0.738	N/A	0.8253
Receipeint Name	N/A	No	X					
Recipeint DOB	DOB	With Pt ID						
Insured ID	Subscriber ID	Yes		0	0	0	N/A	0
Facility ID	Rendering Provider ID	Yes		1	1	0.792	N/A	0.736
Plan ID	Plan Type	No						
Physician ID	Rendering Provider ID	Yes						
Specialty Code	Provider Specialty	Yes		0.12	0.12	0.12	0.12	0.12
Provider Location	Provider Zip	No						
Place of Service	POS	Yes		1	1	1	0	0.75
Principal Diagnosis	Diag	Yes		1	0.934	0.972	0.906	0.953
Other Diagnosis	Diag2-Diag6	No						
Procedure Code	Proc	Yes		1	0.992	0.992	0.982	0.9915
Early and Periodic Screening	N/A	No	X					
Date of Service	From/To Date	Yes		1	1	1	N/A	1
Units of Service	Units	Yes		0.9	0.9	0.9	N/A	0.9
Attending/Referring Physican	Referring Physician	Yes		0	0	0	N/A	0
Performing Provider ID	Rendering Provider ID	Yes						
Provider Type	Vendor Type	No						
Type of Bill	N/A	No	X					
Admission/From Date	From Date	Yes		1	1	1	N/A	1
Discharge/To Date	To Date	Yes		1	1	1	N/A	1
Discharge Pt Destination	N/A	No	X					
Revenue Code	Submitted under Proc	Yes						
Begin Date	From Date	Yes						
End Date	To Date	Yes						
Days Since Admission	N/A	No	X					
National Drug Code	NDC	No						
		Total		73.71%	70.60%	68.65%	50.05%	68.72%

Table 1: Aggregate measures of data quality

References

1. Ballou, D.P., H.L. Pazer. Modeling data and process quality in multi-input, muli-output information systems. Management Science, 31(2):150-162, 1985.

2. Ballou, D.P., H.L. Pazer. Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff. Information Systems research, 6(1):51-72, 1995.

3. Ballou, D.P., G.K. Tayi. Methodology for allocating resources for data quality enhancement. CACM, 32(3):320-329, 1989.

4. Brodie, M. Data quality in information systems. Information Management. 3:245-258, 1980.

5. Gold, M., Burnbauer, L, Chu, K. How Adequate are State Data to Support Health Reform or Monitor Health System Change? Inquiry, 32:468-475.

6. Physician Payment Review Commission, 1997 Annual Report to Congress. Issues Concerning Data Reporting by Health Plans, Chapter 8:161-176.

7. Redman, T. Data Quality: Management and Technology, New York Bantam Books, 1992.

8. Wang, V.C. Storey, C.P. Firth. A Framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering. 7(4):623-639, 1995.

9. Wang, D.M. Strong. Beyond Accuracy: What data quality means to data consumers. JMIS, 12(4):5-34, 1996. QUALITY METRICS FOR HEALTHCARE DATA: AN ANALYTICAL APPROACH - PAGE 4