Validating the Underlying Data: A Measurement Building Block"

Barbara Rudolph, Ph.D., MSSW

Consultant to NAHDO

NAHDO's Goals

- Assist states in determining how best to validate their data
- Provide guidance to states which are bringing up new data systems
- Share expertise with new collaboratives that have formed—e.g., Chartered Value Exchange's
- Make sure states get "credit" for high quality data
- Assure quality and efficiency metrics based on high quality data

Background: Vision for 21st Century

- NCVHS and NCHS Report: Shaping a Health Statistics Vision for the 21st Century
- Report published in 2002—but still valid today
- Defined major gaps and issues in health statistics development, evaluation, and improvement

Defined Issues

- Insufficient connections between suppliers of data and producers of data
- Lack of geographic detail and other specificity
- Lack of timeliness in making data available
- Existing data are unnecessarily difficult to locate, access and use

Defined Issues

- Resource constraints have placed performance and usefulness of major data collection systems at risk (current budget cycle—places systems at even greater risk)

Reflections on this--

- Insufficient focus on developing consensus approaches and standards for data collection, validation methods, software and specifications for individual data systems
- Insufficient focus on new methods for analyzing data, data technologies (web-enabled, auto-editing) that could improve data quality
- Insufficient training on health statistics—not one school of public health offered an introductory course in health statistics to masters and doctoral students—need expanded in-service training as well
- Current participants have been ineffective in dealing with decision-makers who influence how much data is available

Reflections on state of affairs

- Health statistics lacks an overall framework for addressing data processing and data-sharing that could facilitate integration and linkage of data
- Timely production of valid and reliable health statistics—but understanding the cost of “perfection”

Today, we focus on one aspect

- Validation of the underlying data—sharing ideas and resources regarding data edits, validation activities
- NAHDO has been working on a summary of data editing/validation practices
- Request from purchasers to address how state data is validated…

NAHDO Summary of Data Validation Practices

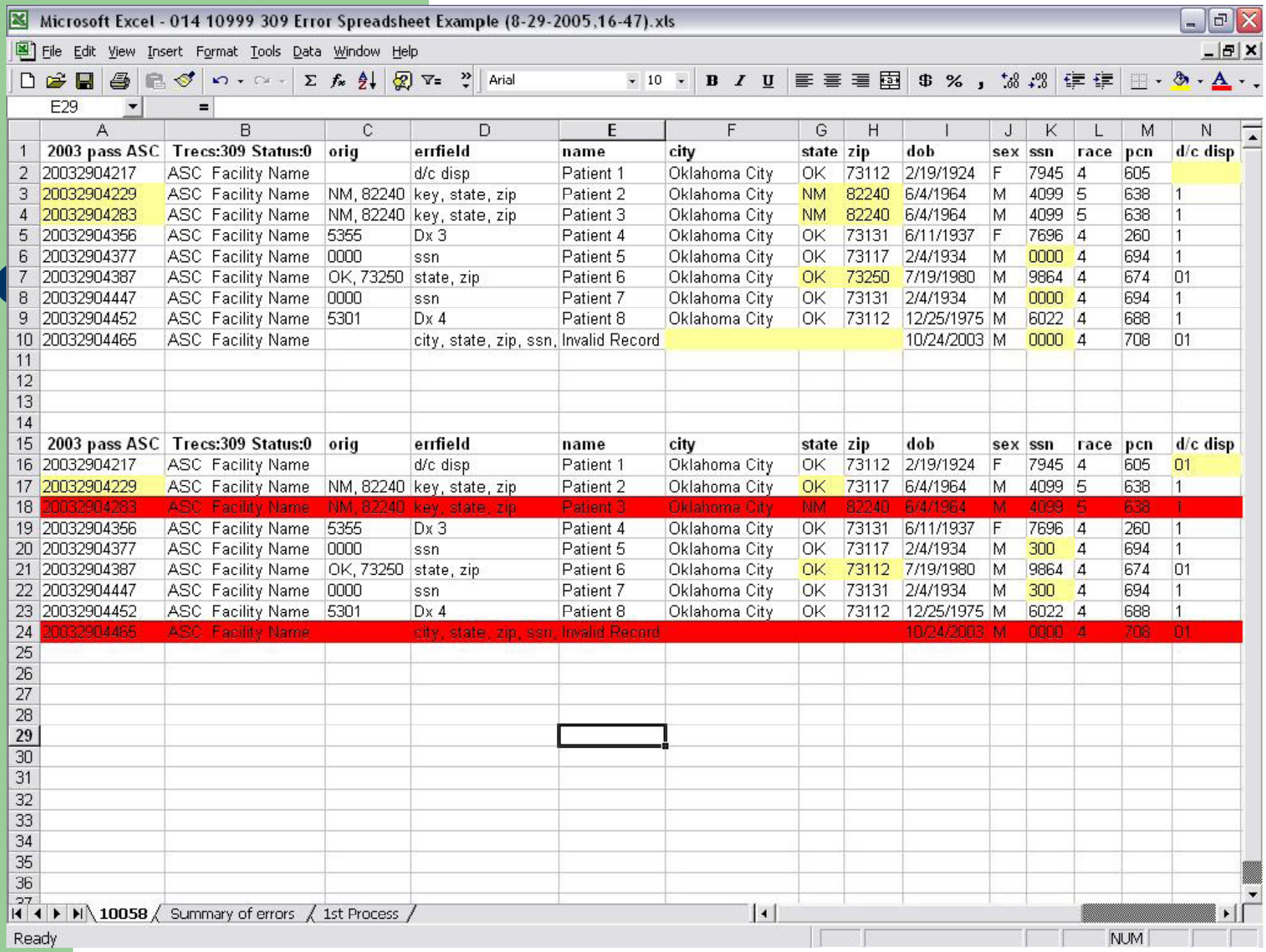
- All agencies should edit data submissions for overall data accuracy (system edits for missing information, invalid fields, implausible fields (range checks), clinical consistency (clinical edits), demographic errors, or other errors causing linkage or grouping failures (linkage failures).
- The data agency should establish data quality thresholds and work with data suppliers to achieve field accuracy and overall data quality requirements.
- The data agency should adopt specific field edits for newly-added fields (Present on Admission, Ecodes, Race/Ethnicity/lab values.)

NAHDO Summary of Data Validation Practices

- The data agency should require data suppliers to update/replace records rejected/failed by edit programs in a timely manner. May establish an accepted error tolerance.
- Post production edit routines (trends in coding and volume of submission) should be incorporated into the data management process.
- Tools, such as fines for non-compliance and/or exemption/waivers should be negotiated with data suppliers experiencing special reporting problems with plans to correct these.
- Agency should normalize the edited data and construct an analytic file, with documented recoding/aggregation/enhancements (age grouping, suppression, severity adjustment fields, readmit field)

NAHDO Summary of Data Validation Practices

- The agency should provide for a validation period for data suppliers to review and correct their submissions and sign off on the accuracy of the final data.
- Data Users Guide—providing information on how data is checked for quality; types of edits used; errors found; confidence in quality of data elements.
- Error reports to the providers with substantial errors—fines or public reporting of data errors.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	2003 pass ASC	Trecs:309 Status:0	orig	errfield	name	city	state	zip	dob	sex	ssn	race	pcn	d/c disp
2	20032904217	ASC Facility Name		d/c disp	Patient 1	Oklahoma City	OK	73112	2/19/1924	F	7945	4	605	
3	20032904229	ASC Facility Name	NM, 82240	key, state, zip	Patient 2	Oklahoma City	NM	82240	6/4/1964	M	4099	5	638	1
4	20032904283	ASC Facility Name	NM, 82240	key, state, zip	Patient 3	Oklahoma City	NM	82240	6/4/1964	M	4099	5	638	1
5	20032904356	ASC Facility Name	5355	Dx 3	Patient 4	Oklahoma City	OK	73131	6/11/1937	F	7696	4	260	1
6	20032904377	ASC Facility Name	0000	ssn	Patient 5	Oklahoma City	OK	73117	2/4/1934	M	0000	4	694	1
7	20032904387	ASC Facility Name	OK, 73250	state, zip	Patient 6	Oklahoma City	OK	73250	7/19/1980	M	9864	4	674	01
8	20032904447	ASC Facility Name	0000	ssn	Patient 7	Oklahoma City	OK	73131	2/4/1934	M	0000	4	694	1
9	20032904452	ASC Facility Name	5301	Dx 4	Patient 8	Oklahoma City	OK	73112	12/25/1975	M	6022	4	688	1
10	20032904465	ASC Facility Name		city, state, zip, ssn,	Invalid Record				10/24/2003	M	0000	4	708	01
11														
12														
13														
14														
15	2003 pass ASC	Trecs:309 Status:0	orig	errfield	name	city	state	zip	dob	sex	ssn	race	pcn	d/c disp
16	20032904217	ASC Facility Name		d/c disp	Patient 1	Oklahoma City	OK	73112	2/19/1924	F	7945	4	605	01
17	20032904229	ASC Facility Name	NM, 82240	key, state, zip	Patient 2	Oklahoma City	OK	73117	6/4/1964	M	4099	5	638	1
18	20032904283	ASC Facility Name	NM, 82240	key, state, zip	Patient 3	Oklahoma City	NM	82240	6/4/1964	M	4099	5	638	1
19	20032904356	ASC Facility Name	5355	Dx 3	Patient 4	Oklahoma City	OK	73131	6/11/1937	F	7696	4	260	1
20	20032904377	ASC Facility Name	0000	ssn	Patient 5	Oklahoma City	OK	73117	2/4/1934	M	300	4	694	1
21	20032904387	ASC Facility Name	OK, 73250	state, zip	Patient 6	Oklahoma City	OK	73112	7/19/1980	M	9864	4	674	01
22	20032904447	ASC Facility Name	0000	ssn	Patient 7	Oklahoma City	OK	73131	2/4/1934	M	300	4	694	1
23	20032904452	ASC Facility Name	5301	Dx 4	Patient 8	Oklahoma City	OK	73112	12/25/1975	M	6022	4	688	1
24	20032904465	ASC Facility Name		city, state, zip, ssn,	Invalid Record				10/24/2003	M	0000	4	708	01
25														
26														
27														
28														
29														
30														
31														
32														
33														
34														
35														
36														

What can we learn from others?

- A number of researchers and others have addressed data quality in articles, white papers, and data documentation
- Angeles and MacKinnon modeled an algorithm to address integration of data for public websites
  - Defined quality criteria
  - Assigned metrics to rate the criteria

Angeles and MacKinnon

Quality Measurement and Assessment Models including Data Provenance to grade Data sources

Angeles, Pilar
PhD Student
Heriot-Watt University
School of Mathematical and Computer Sciences
Edinburgh
United Kingdom
pilar@macs.hw.ac.uk

MacKinnon, Lachlan
Director of Postgradute Study
Heriot-Watt University
School of Mathematical and Computer Sciences
Edinburgh
United Kingdom
lachlan@macs.hw.ac.uk

Abstract:

User priorities, data inconsistencies and quality differences between participating data sources have not been fully addressed, in the process of data integration.

We propose a Data Quality Manager (DQM) to establish communication between the process of integration, the user and the application, to deal with semantic heterogeneity problems.

The aims of the project are to handle inconsistencies at data value level. To achieve this we identify Data Quality Criteria, drawn from an existing research, and their accompanying metrics.

Based on these criteria, we can distinguish between different data sources based on their quality within the context of a user query.

The Data Quality Manager contains a Reference Model, a Measurement Model and an Assessment Model to define the quality criteria, the metrics and the assessment methods, respectively. A Combinatorial Algorithm has been developed, to consider the provenance of the data sources and the user quality priorities to grade the relative data quality within the system, to resolve extensional inconsistencies.

1 Introduction

Data quality degradation can derive from conflicts during the integration process, which may complicate the task of deciding which data to trust. In order to determine data quality, we are developing a Data Quality Manager [7], which utilises a number of criteria, including data provenance, to grade data sources.

The Data Quality Manager will establish the basis for taking decisions on extensional inconsistencies within the databases through the following steps:

Based on certain criteria of quality, the measures of the participating data sources are stored in a repository.

Data Provenance determines the original source of data and compares their quality, in order to provide an appropriate measure of relative quality.

A combinatorial algorithm is used to identify relative quality between data sources considering data context and user quality priorities.

The intension of this paper is to establish the Measurement and Assessment Models from existing body of knowledge, to support the development of the Data Quality Manager. Moreover, to present how Data Provenance provides with a mechanism to compare data sources in context in any user query situation.

The organization of this paper is as follows: Section 2 is concern with the Data Quality Reference Model. Section 3 relates to the Quality Measurement Model.

What can we learn from others?

- One construct they use is “Reputation”—defined as the extent to which data are trusted or highly regarded in terms of their source or content. They include “correction of mistakes” as a key item in reputation

- Usefulness of the data is the degree of value to the user, which is dependent on relevance, accuracy, completeness and timeliness

Study of billing errors--Malone

- Colorado Foundation for Medical Care (QIO) reviewed billing data to reduce number of errors—specifically for “outpatient billed as inpatient”
- In five hospitals—they estimated that 9% of the bills were incorrect—with a cost to Medicare of approximately $5,600 per error. (One hospitals had an error rate of 16.53%)
- Review how hospitals code observation stays and one day hospitalizations—provide clear instructions—run edit checks on % of one-day stays

Informatica
White Paper

Performance Management, Data Governance, and Data Quality Metrics

Establishing a business case for data quality improvement hinges upon the ability to document the pains incurred by data flaws in running the business. The tasks of segmenting them across impact dimensions and categorizing each impact within lower levels of a hierarchical taxonomy facilitates researching negative financial impacts specifically attributable to "bad data." Reviewing the scale of the data failures based on their corresponding negative financial impacts suggests ways to prioritize the remediation of data flaws, which in turn relies on data quality tools and technology.

However, the challenge in employing the concept of "return on investment" for justifying the funding of an improvement project is the ability to monitor, over time, whether the improvements implemented through the project are facilitating the promised positive impacts. So in contrast to the approach used to establish the business case, we can see that if business performance, customer satisfaction, compliance, and automated logistics are all directly tied to ensuring high quality data, then we should be able to use the same kinds of metrics to evaluate the ongoing effectiveness of the data quality program. Documenting this approach, standardizing its roles and responsibilities, and integrating the right tools and methods are the first key tasks in developing a data governance framework.

Positive Impacts of Improved Data Quality

The business case is developed based on assessing the negative impacts of poor data quality across a number of high-level categories: decreased revenues, increased costs, increased risk, and decreased confidence. Since a proactive approach to data governance and data quality enables the identification of the introduction of flawed data within the application framework, the flawed processes that are responsible for injecting unexpected data can be corrected, eliminating the source of the data problem. As we eliminate the sources of poor data quality, instead of looking at the negative impact of poor data quality, let's consider the positive impacts of improved data quality namely: increased revenues, decreased costs, decreased risks, and increased confidence.

Business Policy, Data Governance, and Rules

Not only did the impact analysis phase of the business case process identify impact areas, it also provided some level of measurement and corresponding metrics. For example, Figure 1 shows an example of how data errors introduced at an early stage of processing contribute to various business impacts. The missing product identifiers, inaccurate product descriptions, and inconsistency across different systems contributed to the list of business impacts shown at the right.

White Paper by David Loshin

- Provides dimensions of data quality and metrics for quantifying data quality performance
- Positive impacts of improved data quality—more data sales, increased confidence in data, decreased risks
- Dimensions include: Uniqueness of record, accuracy (do data values correspond with another source?), consistency within record (gender issues), completeness, timeliness, currency (with the world), conformance with standards,
- Validations—assertion about what must be true
- Ongoing monitoring and process control policies
- Data quality scorecard

The Data Quality Scorecard

- Measure results of data validation rules and incorporate into a data quality scorecard
- Rules applied at record level can be measured by percent of valid records
- Rules applied at data set level can measure number of occurrences of invalidity
- For any metric used—establish thresholds (pass-fail, acceptable, questionable but still usable, unusable)
- Use a dashboard to monitor on an ongoing basis

Linda Remy
Ted Clay
Geraldine Oliva

METHODS TO PREPARE HOSPITAL DISCHARGE DATA

All California hospitals except certain state or federal facilities are required to submit patient discharge data (PDD) summarizing the course of care for each discharged patient to the Office of Statewide Health Planning and Development (OSHPD). These large datafiles contain arrays of diagnoses and procedures, as well as other information to describe the patient, geographic characteristics, and the clinical course of care for every patient discharged in a given year.

OSHPD distributes the PDD to qualified researchers such as the Family Health Outcomes Project (FHOP). The FHOP human subjects protocols permit us to have the confidential PDD, for all discharges and ages, from 1983 forward. Currently we have processed all years through 2000 and are about to start with the 2001 and 2002 files. This document presents an overview of the methods we developed to create the core files we use as the source for the different PDD-based research and data products that FHOP distributes.

CONFIDENTIAL MASTER FILES

The confidential PDD includes the following data elements: Social Security Number (SSN), dates (birth, admission, discharge, procedures) and 5-digit ZIP of patient residence. We were permitted to obtain the confidential PDD with these confidential elements because some of our work involves linking PDD records either within the file in a given year or over multiple years, or to other data such as the Vital Statistics mortality files which we access using an encrypted SSN we create. To protect confidentiality, work with these files is done on stand-alone work stations. Access to the files is restricted to two key members of the FHOP research team.

Files with the SSN never reside on the work stations. We developed an algorithm to create an encrypted SSN (SSNC) that is applied as the raw PDD are read into SAS. The encryption method uses a random number process to reassign digits, while maintaining the ability to make soft-linkages to correct for data entry errors. The algorithm includes a second routine to create a second identifier (SSNCN) based on SSNC if available or other data if SSNC is not available. With exact dates of birth, admission and discharge, admission source, and disposition, and the patient's ZIP code of residence, SSNCN allows us to "soft-match" likely transfers and readmissions. Once records are linked for a given study, a unique ID is created and the SSNCN is dropped from the analytic files. This method is described elsewhere.[1] SSNs were not available before 1990. Thus linkage is more reliable for certain years and conditions than for others. We subsequently pursued extending the linkage methodology to children and young adults, and added extensive reliability checks to the procedure, given the increased uncertainty.[2]

The first sequence of programs reads the data, encrypts the SSN, creates SSNCN, does minimal edits, and creates some variables. Details vary from year to year depending on the structure of the incoming data. We check results with various diagnostic listings and descriptive statistics. For example, we may identify unformatted values for which we need to update format libraries. We also may identify data errors that need to be addressed in subsequent programs.

The very large PDD file is output into a main file of core data elements (MAINyy), with extra diagnoses (OTHRDXyy) and procedures (with their associated dates) (MORPROyy) removed to smaller files to reduce computing storage and overhead needs. The three files can be linked as needed using a unique discharge record identifier (YR_OBS) and SSNCN to link cases for the same person. The program also outputs a small file of records with date errors (DTERRSyy), and a small file with other variables we do not want (XVARSyy).

References cited

- Shaping a Health Statistics Vision for the 21st Century, Final Report, November 2002, DHHS Data Council, CDC, NCHS, NCVHS.
- P. Angeles and L. M. MacKinnon, *Quality Measurement and Assessment Models including Data Provenance to grade Data Sources*, Proc. ATINER Conference 2005, Athens, June 2005.
- S.M. Malone, Billing Error Reduction Project: A Hospital Payment Monitoring Program Special Study.
- D. Loshin, Monitoring Data Quality Performance Using Data Quality Metrics: A White Paper. Informatica. November 2006.
- L. Remy, T. Clay, G. Oliva. Methods to Prepare Hospital Discharge Data. Family Health Outcomes Project, UCSF, June 2004.

Slide Presentations

- http://www.academyhealth.org/files/2008/tuesday/virginiaab/6_10_2008_8_00/andrewsr.ppt#293,22,
- http://www.ok.gov/health/documents/HCI_Inpatient_Outpatient_Surgery_Data_Training.pdf