Building Capacity for Health Informatics in the Future
F. Lau et al. (Eds.)

doi:10.3233/978-1-61499-742-9-161

Better Data Quality for Better Healthcare Research Results - A Case Study

Robert HART[a,1] and Mu-Hsing KUO[a]

[a] *School of Health Information Science, University of Victoria, BC, Canada*

Abstract. Electronic Health Records (EHRs) have been identified as a key tool to collect data for healthcare research. However, EHR data must be of sufficient quality to support quality research results. Island Health, BC, Canada has invested and continues to invest in the development of solutions to address the quality of its EHR data and support high quality healthcare studies. This paper examines Island Health's data quality engine, its development and its successful implementation.

Keywords. Electronic Health Record, Data Quality, Business Intelligence, Health Information

1. Introduction

Reliable/quality data is foundational to quality research results [1, 2]. This brings us to the question what is data quality and how do we measure it? Data quality is the assessment of data to determine its viability or fitness for a given purpose [3-7]. A proper assessment of data quality will examine the data from several perspectives or dimensions including validity, accuracy, completeness, relevance, timeliness, availability, comparability, consistency, duplication, integrity and conformity [8]. Any organization that values or places significant importance on its data and information assets from an operational or strategic perspective needs to examine and measure its data quality. In health care, where the electronic health record has a direct impact on patient care, this must be stressed. Evaluation frameworks are essential to this [9, 17].

2. Foundation Elements for Data Quality

For an organization to address data quality, certain foundational elements are required that provide support to accomplish this goal. These include Organizational Commitment, Master Data Management, Metadata Management and other aspects of data architecture necessary to support information system [3, 5, 6, 8]. Data quality is dependent on data architecture for defining the information within the business systems. In some aspects, data quality can be considered part of data architecture as the definition of the data and the rules applied to that data are where the assessment of data quality begins. A proper review of these areas is not within the scope of this paper but understanding the need for these components is necessary.

[1] Corresponding Author: Robert HART; Email: Robert.Hart@viha.ca

2.1. Organizational Commitment

To support data quality, an organization must undertake the development of processes to capture, measure and take corrective action. To do this, a commitment must be present at the highest levels of an organization to foster a culture of quality [6]. Without the recognition that data quality is critical to the organization, it will never be possible to address data quality issues, as no organizational support will exist.

At Island Health it was recognized that data quality is a concern not only at the level of direct patient care but also for the management of health services. To support this, Island Health began the development of a system to capture, measure and report on data quality several years ago. Efforts within the organization to develop procedures to address data quality have always existed and continue to this day.

2.2. Data Architecture and Metadata

Data architecture involves defining the data and its management within an organization and its information systems [11, 12]. The information captured as part of data architecture is the foundational metadata that data quality is built upon. All computer information systems have metadata, if only at the minimum level of a relational database data dictionary. For any attribute of a computer system, it is necessary to define what that attribute is, what it represents, how it is used, how it relates to other attributes and what represents valid use of that attribute.

Extensive metadata standards exist [10] and are managed by multiple organizations. The health profession is supported by many of these such as the Canadian Institute for Health Information [15, 16] or the World Health Organization with extensive resources available from them. These resources are a valuable asset and provide information for the development and support of data quality.

For the purposes of capturing data quality we need metadata [4, 10, 13] in order to define the data quality rules and measures. The definition of the business entity containing the error down to the field attribute has to exist in our metadata repository. Island Health has developed a metadata management tool and is moving forward with a commitment to this system inclusive of data quality.

2.3. Master Data Management

Master data management is the term applied to the governance of the critical data of an organization [17, 18, 19]. This involves policies, standards and processes to manage the base reference data of corporate computer systems. Within healthcare there is a strong culture of master data management represented in the foundational standards that our health systems are based upon. These include standards such as ICD-10-CA [15, 16], The Canadian Classification for Health Interventions (CCI), Residential Care Assessment (RAI-MDS 2.0) and many more.

The health profession has long supported standards for capturing and measuring a patient's health even prior to the development of health information systems. These standards have continued to support systems such as our Electronic Health Records and are an invaluable resource.

2.4. Data Profiling

Though not a foundational component, an early step in developing a data quality system and a valuable tool in its development is data profiling [4, 5, 9, 14]. This involves the systematic evaluation of data and information for a business application. Several commercial tools are available that provide data profiling abilities. In its most basic form, this can be done directly with SQL against the data although this can be cumbersome and difficult.

Data profiling will typically examine individual data fields in a business system and provide information on the mean, mode, standard deviation and other aspects of those fields. For character data it can provide pattern analysis and frequency counts and is frequently used for fields such as postal codes or phone numbers. In more sophisticated analysis it can examine multiple fields to view relationships between them or compare data against reference information and patterns [21]. The purpose of data profiling is to tell us what our data is. Combined with the other foundational elements that define what our data is supposed to be, we have all the pieces necessary to build a data quality system.

3. A Case Study- Island Health Authority Data Quality Engine

Island Health built its data quality system inside its business intelligence and data warehouse. This is a practical and versatile approach as it is unconstrained by the source computer systems and allows the evaluation of data quality across multiple business areas with no impact to the source system. The data quality system begins with the metadata catalog. This follows the design suggested by both Maydanchik and Olson [3, 4]. The catalog is a hierarchical database structure as shown in Figure 1 where information objects are stored in a parent-child relationship inside the metadata object table with descriptive information in a child attribute table. These attributes follow defined standards and are based on the type of metadata object they relate to. As an example a database column would have attributes representing column name, definition, datatype, valid values, reference domain information or business rules governing its use.

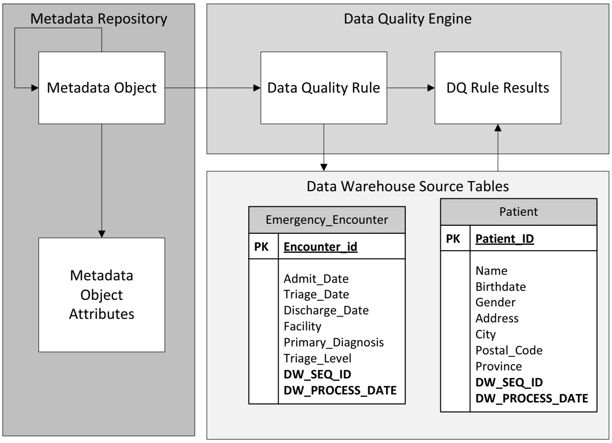

Figure 1. Metadata Repository, Data Quality Engine and Data Warehouse.

The metadata catalog is a repository of information that describes the business systems and data used by Island Health. An example of elements stored in the catalog would be a patient or emergency encounter table as shown in Figure 1 along with the individual columns in these tables. Metadata could also include business rules that stipulate that an encounter cannot exist without a patient or an admit date, that a patient record must have both birthdate and gender populated, or that an admit date can’t be after the discharge date for an encounter.

The data warehouse replicates the data from Island Health source systems. The source system tables in our metadata catalog are represented in the data warehouse allowing for analysis and reporting on this data. That analysis includes evaluation of the data by the data quality engine.

This analysis is enabled by the capabilities within the data warehouse. Every table that is represented in the data warehouse has additional columns added for auditing and processing purposes. These include a data warehouse timestamp and a unique record identifier DW_SEQ_ID that is distinct for every source record in the database. By leveraging this identifier it becomes possible to identify every record and every data quality error associated to that record in the database.

The Island Health data quality system is a rule-based process [3, 4, 5] where source system data is evaluated against a series of validation rules stored in the data quality rule table. These data quality rules are SQL expressions that select the unique identifier for the source record, the date the rule violation was identified, the date it was corrected and the rule it violates. This data is captured and stored in the system.

Using the captured data quality errors, Island Health has built a data quality star schema based on Kimball’s design [6, 7] to provide performance metrics for the management of data quality within Island Health as shown in Figure 2. This star schema is used to provide analysis and reporting of our data quality and is the final component of the data quality system. Error analysis and reporting is possible by date, data quality rule and metadata repository object.

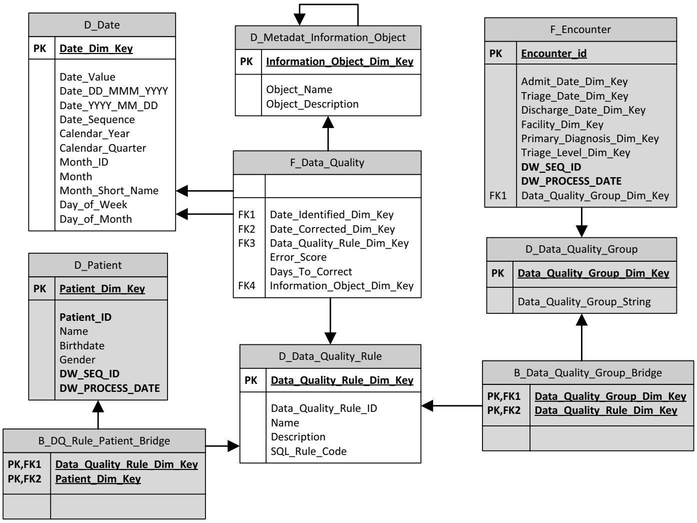

Figure 2. Error star schema and associative star schema structures.

Equally important to our data quality star schema are the associative structures. These are built into the system and relate our data warehouse dimension and fact tables to our data quality dimension as in Figure 2 for Patients and Encounters. By leveraging the unique DW_SEQ_ID we can relate our data quality rules to the records that violate the rule and allow us to report on our errors by dimension or fact table.

4. Island Health Home and Community Care Implementation

Island Health has successfully employed its data quality system in support of multiple business areas. The data quality system was first deployed in home and community care in order to capture and identify data issues related to provincially mandated reporting. The home and community care area benefits from a number of well-documented standards and supporting systems. These include standardized assessments, service provisions and well-documented business practices, all of which provided extensive supporting metadata.

Part of the provision of home and community care services is mandated reporting at both federal and provincial levels. This reporting is well defined and requires high (95%) levels of acceptance for quality testing. To meet these requirements, Island Health utilizes its data quality system to examine its patient demographics, financial assessments and service provision information. Currently over 100 individual validation tests are performed on this data. These tests are performed at a field level, record level and across tables to validate the information. Even systematic checks (21), such as those performed in evaluation of patient assessments in long term care facilities, are possible, although that particular study is more conducive to a profiling exercise.

To address the data quality issues and support improved reporting and analysis, the home and community care area initiated a business process whereby their data is assessed on a regular basis. All data is evaluated against the data quality rules and the results are captured and stored in the system. Any records designated as having failed validation are reported back to the business area and distributed to the responsible staff for correction. Once corrections have been made the data is re-assessed and submitted. The results of provincial reporting for 2016 are shown in Table 1. Currently Island Health data for home and community care has an acceptance rate of better than 99.9%.

Table 1. Home and Community Care Submission Errors by Record Type

	PATIENT INFORMATION			SERVICE INFORMATION			FINANCIAL INFORMATION		
SUBMISSION 2016	Total	Accepted	Rejected	Total	Accepted	Rejected	Total	Accepted	Rejected
JANUARY 28	14741	14719	22	20076	20059	17	10219	10206	13
FEBRUARY 25	15366	15353	13	20909	20898	11	10798	10788	10
MARCH 31	15786	15772	14	22099	22083	16	10857	10847	10
APRIL 21	14789	14776	13	19350	19338	12	10572	10563	9
MAY 19	15539	15527	12	21189	21178	11	10806	10798	8
JUNE 16	15348	15337	11	20846	20835	11	10566	10558	8
JULY 14	15381	15371	10	20822	20816	6	10578	10572	6
AUGUST 11	15254	15244	10	20475	20469	6	10555	10550	5

5. Discussion and Conclusion

The development of a data validation system inside Island Health shows what is possible for an organization to achieve in the area of data quality. A rule based system to examine electronic health records built upon quality metadata standards and data profiling can validate an organizations data for reporting, research, analysis and the provision of patient care.

The data validation system employed at Island Health has identified errors in all aspects of the Electronic Health Records. Those errors have been and continue to be addressed. Patient demographic, assessment and service records have all had data quality issues identified and those errors have been corrected. Even system related errors in the calculation of assessment measures and data conversion issues have been identified as part of this process.

The commitment of Island Health staff in the home and community care area is an example of what can be achieved with a culture dedicated to data quality. The initiation of system processes to validate data quality combined with formal business processes to distribute and correct validation errors shows the support of management and staff towards data quality.

References

[1] J.B. Byrd, et al., Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: The VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. *Am Heart J* **65**:3 (2013), 434–440.

[2] R.S. Sreenivas, et al., Quality of big data in health care, *Int J Health Care Qual Assur* **28**:6 (2015), 621 – 634.

[3] A. Maydanchik, *Data Quality Assessment*, Technics Publications, Bradley Beach, NJ 07720 USA, 2007.

[4] J.E. Olson, *Data Quality Assessment*, Morgan Kaufman, San Francisco, CA 94104-3205 USA, 2003.

[5] D. Loshin, *The Practitioners Guide to Data Quality Improvement*, Morgan Kaufman, Burlington, MA 01803 USA, 2011.

[6] R. Kimball, *An Architecture for Data Quality*, white paper, 2007. Retrieved Sept 12, 2016 from http://www.kimballgroup.com/2007/10/white-paper-an-architecture-for-data-quality/.

[7] R. Kimball, *The Data Warehouse Toolkit*, John Wiley and Sons, New York, NY 10158-0012, 2002.

[8] C. Batini, et al, Methodologies for data quality assessment and improvement, *ACM Compu Surveys (CSUR)* **41** (2009), 1-52.

[9] T. Norris, K. Kerr. *Improving Data Quality in Health Care, Health Information Systems: Concepts, Methodologies, Tools, and Applications*. Medical Information Science Reference, Hershey. PA 17033 2010:218-225.

[10] M. Aljumaili, et al. eMaintenance ontologies for data quality support. *Journal of Quality in Maintenance Engineering* **21**:3 (2015), 358-374.

[11] A. Clyde, Metadata. *Teacher Librarian* **30**:2 (2002), 45.

[12] M.R. Kogalovsky, Metadata in computer systems. *Prog Comput Soft* 39:4 (2013), 182-193.

[13] M. Aljumaili, et al. Metadata-Based Data Quality Assessment, *V I N E J Inform Knowledge Manag Syst* **46**:2 (2016), 232-250.

[14] A. Gupta, *Data Profiling*, Encyclopedia of Database Systems, Springer US, Boston, MA, 2009

[15] Canadian Institute for Health Information, InterRai, Continuing Care Reporting System RAI-MDS 2.0 Output Specifications, 2011-2012

[16] Canadian Institute for Health Information, Canadian Coding Standards for version 2012 ICD-10-CA and CCI (revised September 2012)

[17] K. Williams, K. Robinson, A. Toth, Data quality maintenance of the Patient Master Index (PMI): a 'snap-shot' of public healthcare facility PMI data quality and linkage activities. *Health Inform Manag J* **35**:1 (2006), 10-26.

[18] J. Wilkinson, Multidomain master data management for business success. *Software World* Nov. 2009: 10+. Business Collection. Web. 12 Sept. 2016.

[19] M.A. Poolet, Master data management: a method for reconciling disparate data sources. *SQL Server Magazine* Jan. 2007: 27+. Business Collection. Web. 12 Sept. 2016.

[20] C. Dalton, M. Allen, *Multi-Domain Master Data Management*, Morgan Kaufman, 2015.

[21] J.P. Hirdes, et al., An evaluation of data quality in Canada's Continuing Care Reporting System (CCRS): secondary analyses of Ontario data submitted between 1996 and 2011. *BMC Med Inform Decis Mak* **13** (2013), 27-27.