March 2013

DQC White Paper Draft 1: A consensus-based data quality reporting framework for observational healthcare data

A consensus-based data quality reporting framework for observational healthcare data

Members of the EDM Forum Data Quality Collaborative[1]

Draft: 2012-12-03

1. Objective

The objective of this activity is to develop a consensus-based set of recommendations that improve the reporting of data quality for studies that use observational clinical and administrative data (aka secondary data use) and to ensure transparency and facilitate best practices in secondary data use.

2. Problem Statement

Electronic health records (EHRs) support the capture of detailed clinical, operational, and administrative data as part of routine clinical and administrative processes. Electronic data from administrative sources, billing and claims databases, drug fulfillment, specialized registries and patient-reported data expand the scope and richness of available patient-level data. As EHR use becomes the norm, the availability of electronic data from a variety of practice and patient settings provides an opportunity to transform the spectrum of clinical research, allowing critical insights into the effectiveness of clinical interventions and disease and adverse drug event surveillance in real-world settings. This 'last mile' of research translation focusing on patient-centered, clinically relevant outcomes with real-world patients in real-world practice settings should complement traditional efficacy studies and provide independent evidence in situations where efficacy studies are not feasible. Multi-site comparative effectiveness research, pragmatic clinical trials, and patient centered outcomes research are new models for clinical investigation that have emerged, in part, due to the availability of clinically rich electronic data captured during routine care.

Substantial federal investments in national large-scale distributed research networks and other data sharing initiatives now support the use of EHR data for multi-site retrospective studies. Propelled by the Meaningful Use provisions in the ARRA HITECH act, there is increased interest in leveraging the most current patient-level data available to support healthcare decisions for patients, providers, payers, and policymakers. Increased patient engagement in access to and management of their health care data highlights the growing awareness of the value of EHR data. All of these forces have been unified under the banner of creating collaborative Learning Health Systems that continuously generate new evidence and rapidly respond to that evidence to achieve optimal outcomes with minimal resources. Learning Health Systems leverage EHR data for a broad range of uses including collaborative quality and

[1] In alphabetical order by last name: Jeffrey Brown – Harvard Pilgrim Health Care Institute/Harvard Medical School; Michael Kahn – University of Colorado; Daniela Meeker – Rand Corporation; Meredith Nahm – Duke University; Patrick Ryan – OMOP/J&J; Lisa Schilling – University of Colorado; Nicole Weiskopf – Columbia University; Andrew Williams – Kaiser Permanente Hawaii.

process improvement programs, chronic disease management, public health surveillance, and more formal observational and comparative effectiveness research studies.

Key to achieving any of these desired goals is access to high quality clinical and administrative data that can support the interventions, outcomes and conclusions derived from these real-world studies. Numerous publications illustrate the challenges using of EHR data for comparative effectiveness research, including recording biases, workflow differences, and data quality, such as invalid, inconsistent, or missing data. Even data collected prospectively using dedicated data collection tools and personnel have documented data quality deficiencies. Data combined from various intra-institutional data sources and then combined across multiple institutions may have data quality issues that vary significantly over institutions, data domains (e.g., demographics, observations, medications, laboratory results) and time.

Given the inherent limitations in EHR data, it is critical to develop a comprehensive data quality assessment framework for describing data quality. Associated with this framework should be a comprehensive set of data quality reporting recommendations that are applied to data received from any data source. A common framework is essential for implementing comprehensive data quality standards, defining roles and responsibilities for all stakeholders, describing core desiderata for monitoring, identifying data quality issues, and quantifying improvements in data quality over time. A comprehensive set of data quality reporting recommendations is necessary to promote trust by establishing reproducible processes and improving the transparency and integrity of the appropriate use of EHR data.

The desired outcome from these efforts is to create a consensus-based set data quality reporting recommendations that should be included with all EHR data. These data quality elements can be used by a site's data manager responsible for releasing data for internal use or to a network, by a data analyst responsible for combining data across a data network, by a clinical investigator using a data set for an analysis, and by a consumer of the analytic results derived from a data set. Existing efforts for describing reporting recommendations in specific settings include the Cochrane Collaboration for describing the quality of published evidence (http://www.cochrane.org/) or STROBE recommendations for reporting results from observational clinical studies (http://www.strobe-statement.org/). Our work aligns most closely with STROBE recommendations. But as described below, we have taken a more directive approach that is captured in our focus on defining "**Figure 1 for Data Quality Assessment Reporting**".

Not in scope for this work are recommendations for how data quality could be improved. A related set of literature points to best practices that combine workflows and technologies that attempt to ensure high data quality from the moment of initial observation and recording. The current effort of this white paper and associated workshop focuses on assessing and reporting core data quality features that provide transparency to the strengths and weaknesses of the underlying data acquisition and processing to consumers of these data and associated findings.

3. Defining Clinical Data Quality

Data quality is a complex, multi-dimensional concept that defies a single one-size-fits-all description. Data quality is context-dependent which means the same data elements or data sources may be deemed high quality for one use and poor quality for a different use. A key data quality concept is "fitness for use" -- a term originally used in industrial quality control but

adapted to describe the context-dependent nature of data quality. Data are considered fit for use "if they are free of defects and possess desired features... for their intended uses in operations, decision making, and planning." A related contextual feature is the intended setting for data analysis, which we describe in more detail below. Thus developing a comprehensive framework for assessing and describing data quality that captures multiple dimensions and provides value across different stakeholders will require input from a broad range of data creators, data stewards, data curators, and data consumers.

Information Science professionals published formal models of data quality more than 20 years ago. These publications have focused on business-oriented definitions. Recently, members of the EDM Forum's Data Quality Collaborative have recognized the need to develop similar data quality assessment models for clinical data sets, especially in the context of large multi-institutional distributed research networks that have received substantial funding over the past 5 years, such as the AHRQ ARRA DRN program and the FDA Mini-Sentinel program. A separate document provides an evolving list of data quality dimensions that have been proposed in recent publications.

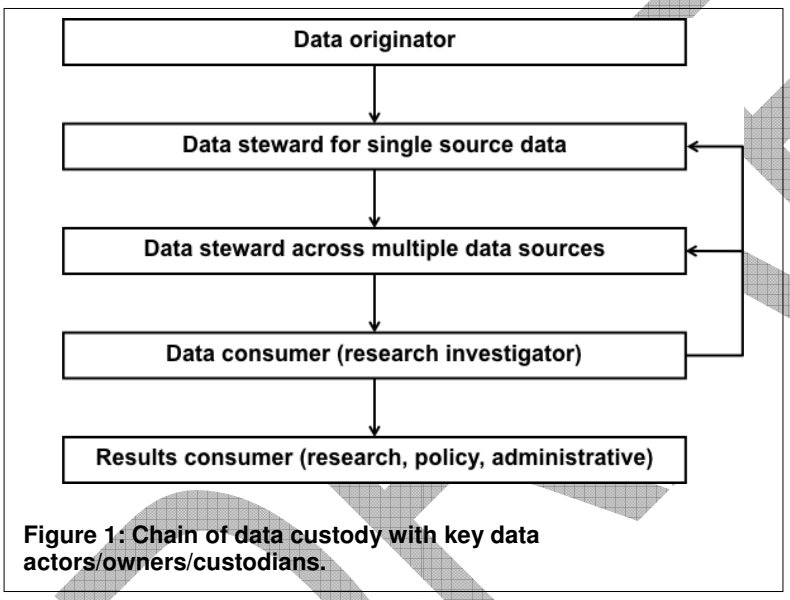

Figure 1: Chain of data custody with key data actors/owners/custodians.

Figure 1 highlights key data "actors" who have one or more roles in managing observational data and in evaluating the quality of data that are entrusted to their oversight or use. Each data custodian illustrated in Figure 1 has unique roles and responsibilities regarding data management, data quality, and data quality assessment and evaluation. Different aspects of data quality assessment may apply at each level in this "chain of data custody." At the end of each arrow is a "data consumer" who obtains data from the "data source" in the previous box. A goal of the workshop is to develop data quality assessment reporting recommendations for each step along this chain and to determine how assessments from earlier stages are added to or incorporated into the data quality assessments of the next step.

4. Defining "Table 1 for reporting data quality"

Members of the EDM Forum Data Quality Collaborative are engaged in a wide range of large-scale multi-institutional data networks spanning prospective regulated clinical trials, pragmatic clinical trials, and observational outcome studies. As hands-on practitioners, we have seen a multitude of individualized ad-hoc approaches to assessing clinical data quality. In addition, we are aware of substantial data quality "cleaning" efforts that occur behind the scenes that often are not disclosed with the final data sets. In this section, we propose a set of explicit disclosures that focus on increasing the transparency of these important data quality processes.

In most articles that describe clinical studies, “Table 1” provides an overall assessment of the study population, such as age and gender distributions and other key clinical or administrative features that describe the study population or cohorts. We propose a similar “**Table 1 for Reporting Data Quality”** that describes data acquisition and processing. In support of reproducible research, this table should contain key information regarding the sources, processes and measures that provide insight into data quality. We have not limited the proposed contents to a set of elements that would fit within the constraints of typeset journal pages. Rather, we acknowledge the growing use of electronic annexes that allow detailed information to be associated with a publication. As a long-term objective, we seek to create a computer-readable data structure that allows detailed data quality assessment metrics to be incorporated into analytic programs that use the associated data sets.

We have aligned our proposed framework with the STROBE reporting recommendations. The STROBE Statement consists of a checklist of 22 items that focus on specific elements of an article (title, abstract, introduction, methods, results, and discussion) reporting results from observational studies. STROBE identifies two “Item Numbers” that are directly relevant to this activity:

- Item 8: Data sources/measurements: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.
- Item 12(c): Explain how missing data were addressed

Our work extends the STROBE reporting requirements around these two existing reporting categories with explicit structured reporting requirements to meet these recommendations.

For data quality, we organized data quality recommended reporting elements around four components: data definition, data acquisition, data processing, and data element characterization:

Data definition refers the intended meaning of each data element present in the data set. Data definition is the heart of reproducibility. Specification of data definition at the data element rather than the data model level (how data are stored in a database) emphasizes necessity of unambiguous specification of data elements and frees investigators and research teams from concern for the data model in which the data are stored or exchanged.

Data acquisition refers to the mechanisms by which data were observed or collected and recorded in the database. Operational definition specifies the way in which the data values were obtained, e.g., Blood pressure measured as the average of three sequential blood pressure readings at five-minute intervals with a clinic manual cuff. Data acquisition methods are sometimes subjective and can significantly impact data quality and interpretation of the data.

Data processing refers to the transformational processes that have been applied to the data. Data processing includes all operations performed on data after the original acquisition. Complete transparency of operations performed on data is necessary for reproducible research.

Data elements characterization: Researchers are responsible for ensuring that the available data are “fit for use” for the purpose of the study. Because fit for use is context-specific, there cannot be a single set of data characterizations that are sufficient for all studies. However, we propose a set of recommendations that should be present in all data quality assessment reports to provide the assurance and transparency that the data do meet minimal "fitness for use" features. Our recommendations include data quality assessments designed to highlight the characteristics of the entire database, and also data quality checks that target core data elements unique to a specific study. The researcher should undertake additional study-specific characterizations and these activities should be reported with the same specificity and transparency as the core quality assessment measures.

Our recommendations extend to all data custodians (Figure 1) along the data transformation process, with each custodian adding their data documentation as they pass data sets along the chain to the next custodian or user. If our recommendations are adopted, a data set will accumulate additional data documentation over its lifetime, ensuring that the needed details are available to all subsequent downstream data consumers. Appendix provides our straw-man recommendations for “**Table 1 for Reporting Data Quality**” in the same checklist model used by STROBE. Appendix is intended to be the starting point for discussion at the EDM Forum Data Quality Workshop.

Participants in the EDM Forum will help refine the proposed "Table 1 for Reporting Data Quality” starting with the straw-man proposal in Appendix Through continued iterative engagement beyond the workshop, we seek to leverage the broad interests of the Workshop participants to develop and publish a consensus statement and guidance document for data quality assessment reporting that adds transparency and value to each data consumer in the chain of data custody (Figure 1). A long-term goal is to create data quality assessment tools that support both the computational and reporting aspects of data quality assessment in a uniform manner.

Appendix

A straw man "**Table 1 for Reporting Data Quality**" is provided for starting the conversation among workshop participants:

	Item#	Recommendation
DATA CAPTURE DOCUMENTATION		
1. Original data source		
Data origin	1	The source of the original or "raw" data prior to any subsequent processing or transformation for secondary use. Examples would be "clinical practices via EHR", "interviewer-administered survey", or "claim for reimbursement"
Data capture method	2	A description of the technology used to record the data values in electronic format. Examples would be "EHR screen entry", "automated instrument upload"
Original collection purpose	3	A description of the original context in which data were collected. Examples would be "clinical care and operations" or "research", and in which facilities
2. Data custodian information		
Data provenance	4	Organization responsible for obtaining and managing the current data set. Examples could be "PBRN", "Registry", Medical group practice", State agency
Data custodian's database model	5	How was the data structured into a data model? URL to documentation
Data custodian's data dictionary	6	What are the data definitions used for data elements? URL to documentation
DATA PROCESSING		
Data extraction, including use of natural language processing, specifications	7	How was the data obtained from the data originator to the data compiler? Examples would be medical record abstraction guidelines or Natural Language Processing algorithms. URL to documentation
Mappings from original values to standardized values	8	How were original data values altered to conform to data model? Documentation of source values and logic/mappings used to transform from original to standardized values

Data management organization's data transformation routines, including constructed variables	9	Description of how the original data was altered in management and to create analysis data sets. URL to documentation. The documentation should allow an independent reader to trace a value in the data back to the source and should explain all operations performed on the data.
Data validation routines	10	Documentation of all data validation rules to which the data were subjected. Rules should identify both data elements and validation algorithms. URL to documentation
Audit trail	11	Documentation of all changes made to data values, user/system making the change and date/time of the change. Reason for the change should be evident from data transformation routines.
DATA ELEMENTS CHARACTERIZATION Documentation		
Data model verified	12	For required study elements verify format, proper storage, and that required elements are not missing. For each verified data element: Format, units of measure, precision, rounding rules: Pass/check/date
Single element data domain descriptive statistics	13	Available or not (#/% missing) Continuous distributions - min, max, range, etc., Categorical – frequencies & proportions by category)] If a specific distribution is anticipated, goodness-of-fit tests.
Temporal data	14	Start before stop dates and times, Distribution of intervals between successive measurements. For time-series -- changes in adjacent values, expected directionality in changes Conformance to state transition / sequencing rules
Multiple variables cross validations	15	Conformance to data model cardinality rules Conformance to data model primary/foreign key rules Conformance to cross-variables dependency rules Conformance to co-occurrence rules Conformance to co-measurement rules (two distinct measurements of the same observation) Conformance to mutual exclusivity rules
STUDY SPECIFIC DATA QUALITY Documentation (as applied by investigators or analytic team)		
Data cleansing/customization	17	Study specific additions to Item# 9

Data quality checks of key variables used for cohort identification	18	Study specific additions to Item# 13-15
Data quality checks of key variables used for outcome categorization	19	
Data quality checks of key variables used to classify exposure	20	
Data quality checks of key confounding variables	21	