Methods for examining data quality in healthcare integrated data repositories

Vojtech Huser †
National Library of Medicine, National Institutes of Health
8600 Rockville Pk, Bld 38a
Bethesda, MD, 20852, USA
Email: vojtech.huser@nih.gov

Michael G. Kahn
Department of Pediatrics, University of Colorado
13001 East 17th Place MS-F563
Aurora, CO 80045 USA
Email: Michael.Kahn@ucdenver.edu

Jeffrey S. Brown
Department of Population Medicine,
Harvard Medical School and Harvard Pilgrim Health Care Institute
401 Park Drive, Suite 401 East
Boston, MA 02215USA
Email: jeff_brown@hphc.org

Ramkiran Gouripeddi
University of Utah, School of Medicine
Salt Lake City, 84102, Utah, USA
Email: ram.gouripeddi@utah.edu

This paper summarizes content of the workshop focused on data quality. The first speaker (VH) described data quality infrastructure and data quality evaluation methods currently in place within the Observational Data Science and Informatics (OHDSI) consortium. The speaker described in detail a data quality tool called Achilles Heel and latest development for extending this tool. Interim results of an ongoing Data Quality study within the OHDSI consortium were also presented. The second speaker (MK) described lessons learned and new data quality checks developed by the PEDsNet pediatric research network. The last two speakers (JB, RG) described tools developed by the Sentinel Initiative and University of Utah's service oriented framework. The workshop discussed at the end and throughout how data quality assessment can be advanced by combining best features of each network.

Keywords: Data Quality, Evaluation Methods, Visualization, Observational Research.

† VH work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC)

1. Introduction

Large Integrated Data Repositories (IDRs) have become indispensable for clinical research. Recent emergence of Common Data Models (CDMs) facilitated creation of tools that provide syntactic integration (shared information model) and in some cases also semantic integration (shared set of target terminologies used by structured data). Retrospective data analyses are increasingly being executed on multiple datasets, and distributed research networks are creating reusable tools that streamline data wrangling, data repository maintenance, and data analytics. Examples of large, well-coordinated IDRs developed using a CDM and distributed network approach include the Health Care Systems Research Network (multi-purpose research network of 18 sites), the FDA Sentinel Initiative (17 sites representing billions of medical encounters to support medical product safety surveillance), and PCORnet (over 70 sites with millions of encounters to support clinical research). Each of these distributed networks has a unique approach to addressing data quality, including some shared approaches, and each has developed tools to facilitate data quality querying.

In 2016, a group of data quality researchers called Data Quality Collaborative (DQC)[1] published a milestone article that introduced a harmonized terminology and framework for data quality assessment (DQA).[2] Another recent study, published in 2017, described experience with regular assessment of data quality within a large pediatric data research network.[3,4] A similar summary exist for the Sentinel network.[5]

This paper provides a summary of the current state of the art and future trends presented at a conference workshop focused on examining current and novel methods in assessing data quality.

2. Data Quality within the context of the OHDSI consortium (presented by V. Huser)

2.1. *Achilles Heel Data Quality Tool*

Formulation and refinement of the Observational Medical Outcomes Partnership (OMOP) data model since 2009 provided a data standardization that opened the possibility of creating data quality assessment (DQA) tools that could work unmodified across multiple datasets or multiple healthcare institutions. The OHDSI consortium created in since 2014 a tool, called Achilles Heel that included several data quality checks focused on OMOP data model conformance and some DQA checks. It was sub-component of a tool, called Achilles that also provided data characterization functionality. This tool provides general level data quality assessment as well as model conformance checking and is being actively extended with new functionality.

In September 2017, OMOP CDM workgroup agreed on extending the model with a METADATA table that would allow capturing unstructured (as free text) and structured description of data. The new METADATA table, being part of core data model, opens new feature possibilities for data quality tools (for example recognizing clearly general population datasets from clinical trials datasets and running appropriate data quality checks depending on the dataset type).

2.2. *OHDSI network study evaluating Data Quality*

To advance the analysis of data quality of sites within the OHDSI network, in 2016, OHDSI community initiated a new study focused on comparing data quality measures within the network.[6] This study builds on previous study comparing Achilles Heel outputs at several OHDSI sites.[7] The study introduced a smaller subset of dataset measures (compared to full set of measures generated by the Achilles tool) that contain less detailed data about site and thus possibly encouraging site participation in measure comparisons. The study also deals with quantifying the amount of data that has not been fully mapped to standard concepts (target terminologies for a given data domain). See Figure 1. Such data is typically present in the dataset but mapped to concept with a concept_id of zero. The Procedure and Observation domains were found to have most unmapped data across 10 OMOP datasets compared in the study to date.

2.2.1. *Empirical Thresholds*

Reaching consensus around what constitutes a high quality dataset (in general context, not for a given study) is difficult. Given lack of consensus, the approach chosen by the DataQuality study was to use empirical threshold, where datasets that are below 10th percentile receive a data quality notification (the lowest level of data quality error). For example, for a measure of 'percentage of outpatient visits', within the network, the measure varies from 33% to 92%. Using this approach, datesets below 43% of outpatient visits (below 10th percentile) are marked as likely not containing visit details about all spectrum of care (possibly come from a delivery network that manages mostly hospitals and not a full spectrum of outpatient care).

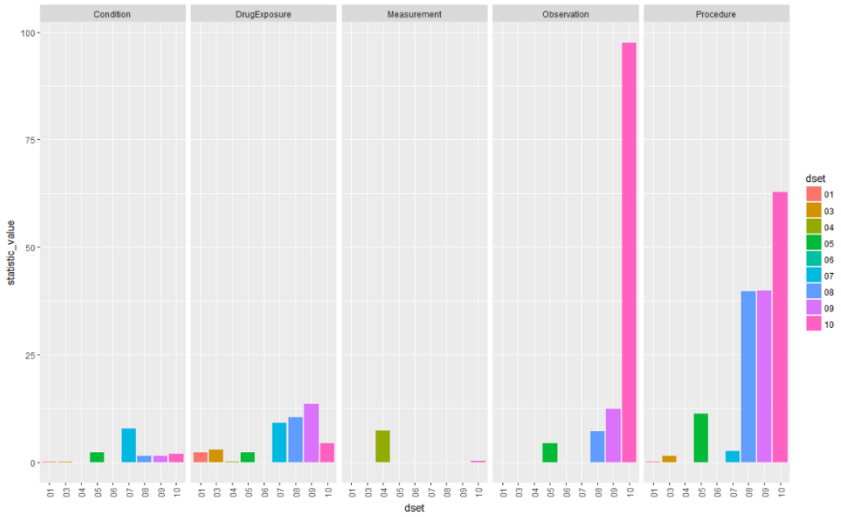

Fig. 1. Percentage of unmapped data (shown on x-axis) by data domain across 10 OMOP datasets in the Data Quality study.

3. Data Quality within the context of the PEDSNet (presented by M. Kahn)

PEDSnet is one of eleven national clinical data research networks sponsored by the Patient Centered Outcomes Research Institute that combines electronic health record (EHR) data from eight large

free-standing pediatric research hospitals.[8,9] All eight institutions implemented the OHDSI OMOP common data model for harmonizing data across different clinical environments and EHR systems. In the process of creating PEDSnet, an extensive data quality assessment program was constructed that examines data submitted by each organization and across all PEDSnet data partners. As the data quality program expanded in scope, findings related to different interpretations of how data were to be harmonized started to appear, resulting in data that were not comparable across institutions, greatly reducing the key objective for investing in a common data model. Different system configurations, workflows, and business processes often were the source of these descripancies. The data quality findings triggered the need to establish detailed data conventions with explicit rules or conventions describing how to transform site-specific data into the common data model to ensure all eight data partners are making the same transformation decisions, especially in unusual or unique situations. This modest document has grown to over 60 pages and continues to grow as new "edge cases" appear or as new data domains introduce previously unknown differences in site-level ETL proceesing. And the types of data quality issues have evolved over time as the ETL processes and network has matured.[4]

Like PEDSnet, other data networks have developed data quality processes. We examined over 11,000 data quality rules used in six large data networks of varying size and maturity.[10] We show the vast differences in data quality rules across these networks. One highly desired goal of the data quality community is to develop sharable/reusable data quality tools. We also describe recent work in developing a common data model for data quality measures that may enable networks to share data quality methods and tools.

4. Data Quality methods used by Sentinel Network (presented by J. Brown)

The US FDA Sentinel System is a medical product safety surveillance network supported by the US FDA.[11] The collaboration includes 17 data partners – 7 health insurers, 9 integrated delivery systems (i.e., insurance and care delivery) and one national hospital system - that have transformed their data into the Sentinel Common Data Model (SCDM).[12] The system includes billions of medical encounters and over 425 million person-years of data. The Sentinel Operations Center (SOC) manages the SCDM and the data quality assurance review[13,14] process. The data quality assurance team has developed and posted online a distributed program to assess network data quality across four levels of data quality checks. The first two levels of checks focus on data model compliance issues such as conformance with data model formats, completeness, validity, and cross-table and cross-variable integrity. For example, the process will check known "always" relationships such as a person with a recorded medical encounter or outpatient pharmacy dispensing must also be in the demographic table. The other checks focus on less concrete metrics focusing on trends and logical plausibility. These include metrics such as the proportion of care visits that are inpatient stays, the number of visits per person per year, the number of medication dispensing per person per year, and overall monthly trends in patients and visits. The SOC developed a set of tools to support the data quality review process; two reviewers conduct each review. The SOC conducts approximately 50 reviews per year, each review involves up to 1,500 individual data checks. The data quality review process has changed substantially since the initiation of the Sentinel pilot (Mini-Sentinel) in 2009.

The data quality approach is now more comprehensive, includes more checks, and now has explicit errors rules built into the distributed quality assurance program that will halt the program and reject a refresh is certain metrics are not met.

5. A Service Oriented Architecture for Assessing Quality of Heterogeneous Health Data (presented by R. Gouripeddi)

Translational research is dependent on secondary use of electronic health data for selection of participant cohorts and assessing real-world effectiveness of different interventions. It is therefore important to assess the quality of data used for these studies for often present data quality (DQ) issues and differentiate between natural and extraneous variations in data[15]. Assessing quality of health data needs to account for semantic and syntactic heterogeneity in health data and the diverse needs of practitioners of DQA. Several data quality conceptual frameworks (DQF) have been proposed across DQ and Information Quality domains for DQA. These DQF are diverse, varying from simple lists of concepts to complex ontological and taxonomical representations of data quality concepts (DQC) for different domains of application. There is a lack of consensus on using these DQF for a comprehensive DQA of a given dataset as well as absence of a "one-framework-fits-all" solution for DQA.[16]

In order to meet these requirements we developed a service-oriented architecture-based (SOA) DQA platform, Open Quality and Analytics Framework[17] (OQAF) consisting of three components:

1. Quality Knowledge Repository (QKR): We extracted DQC, their definitions and applicable measures, their relationships, and the computability of DQC in existing DQF (e.g. Kahn[15], Weiskopf[18] from literature. We identified primitives existing in different DQF to develop a DQ metamodel and implemented it as the QKR. The QKR is based on OpenFurther's Metadata Repository[19] (MDR) which is a standard-based, in-house developed repository that stores metadata artifacts and their relationships.
2. Federated Data Integration Platform: We use the OpenFurther platform (OF) that supports semantically consistent metadata-centric querying of heterogeneous data sources for translational research using an MDR, a terminology/ontology server and various software services (SS) for orchestrating execution of queries.[19]
3. Visualization Meta-Framework (VMF): This provides a repository for storing visualizations for different DQC and their measures. We are currently generating content for the VMF from literature and using crowd-sourced methods.

An end-user characterizes heterogeneous datasets or sources that are distributed or aggregated, by selecting one or more DQC and their measures from any DQF stored in the QKR. At the DQ query execution layer, OF's SS use messaging obtained from the QKR and the VMF along with reusing model and terminology mappings generated for performing federated data queries to perform DQA and visualizations. The content stored in the VMF informs users in the selection of appropriate visualizations for their DQA.

We present the architectural design, implementation, and evaluation of OQAF and its components - open-source, agnostic approach for standardizing DQA.

References

1. Health A. Data Quality Collaborative. 2015; http://repository.academyhealth.org/dqc/. Accessed May 15, 2015.
2. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016;4(1):1244.
3. CHOP. PEDSNet data quality assesment tool. 2017; https://github.com/PEDSnet/Data-Quality-Analysis. Accessed June 5, 2017.
4. Khare R, Utidjian L, Ruth BJ, et al. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc*. 2017.
5. Brown J, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Medical care*. 2013;51(8 0 3):S22-S29.
6. Huser V. OHDSI Data Quality study. 2016; http://www.ohdsi.org/web/wiki/doku.php?id=research:dqstudy. Accessed Jan 2, 2017.
7. Huser V, DeFalco F, Schuemie M, et al. Multi-site Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets. *eGEMs (Wash DC)*. 2016.
8. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578-582.
9. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc*. 2014;21(4):602-606.
10. Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison Of Data Quality Assessment Checks In Six Data Sharing Networks. *Generating Evidence & Methods to improve patient outcomes (eGEMS)*. 2017;5(1).
11. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clinical pharmacology and therapeutics*. 2016;99(3):265-268.
12. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1:23-31.
13. Sentinel. Data Quality Review and Characterization. 2016; https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/data-quality-review-and-characterization.
14. Raebel MA, Haynes K, Woodworth TS, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf*. 2014;23(6):609-618.
15. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. 2012;50 Suppl:S21-29.
16. Kahn MG, Brown JS. Transparent Reporting of Data Quality in Distributed Data Networks. *EGEMS (Wash DC)*. 2015;3(1).
17. Rajan NS, Gouripeddi R, Facelli JC. A service oriented framework to assess the quality of electronic health data for clinical research. Paper presented at: Healthcare Informatics (ICHI), 2013 IEEE International Conference on2013.
18. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics*. 2013;46(5):830-836.
19. Gouripeddi R. FURTHeR: An Infrastructure for Clinical, Translational and Comparative Effectiveness Research. . ***Proc AMIA Symp***. 2013.