11-30-2016

Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets

Vojtech Huser
NIH (NLM), vojtech.huser@nih.gov

Frank J. DeFalco
Janssen Research & Development

Martijn Schuemie
Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States; and Observational Health Data Sciences and Informatics (OHDSI) New York, New York, United States

Patrick B. Ryan
Janssen Research & Development

See next pages for additional authors

Recommended Citation

Huser, Vojtech; DeFalco, Frank J.; Schuemie, Martijn; Ryan, Patrick B.; Shang, Ning; Velez, Mark; Park, Rae Woong; Boyce, Richard D.; Duke, Jon; Khare, Ritu; Utidjian, Levon; and Bailey, Charles (2016) "Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 4: Iss. 1, Article 24.
DOI: https://doi.org/10.13063/2327-9214.1239
Available at: http://repository.edm-forum.org/egems/vol4/iss1/24

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets

Abstract

Introduction: Data quality and fitness for analysis are crucial if outputs of analyses of electronic health record data or administrative claims data should be trusted by the public and the research community.

Methods: We describe a data quality analysis tool (called Achilles Heel) developed by the Observational Health Data Sciences and Informatics Collaborative (OHDSI) and compare outputs from this tool as it was applied to 24 large healthcare datasets across seven different organizations.

Results: We highlight 12 data quality rules that identified issues in at least 10 of the 24 datasets and provide a full set of 71 rules identified in at least one dataset. Achilles Heel is a freely available software that provides a useful starter set of data quality rules with the ability to add additional rules. We also present results of a structured email-based interview of all participating sites that collected qualitative comments about the value of Achilles Heel for data quality evaluation.

Discussion: Our analysis represents the first comparison of outputs from a data quality tool that implements a fixed (but extensible) set of data quality rules. Thanks to a common data model, we were able to compare quickly multiple datasets originating from several countries in America, Europe and Asia.

Acknowledgements

This research was supported in part (VH) 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), Reagan-Udall Foundation for the FDA project IMEDS-SA-0011, National Institute on Aging (K01AG044433), and the National Library of Medicine (1R01LM011838-01). The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of their employing institutions. Contributions: VH conceived the evaluation idea, contacted the participating sites and performed the aggregate data analysis. PR, MS, FD and Chris Knoll were initial authors of Achilles and Achilles Heel with later contributions from other OHDSI collaborators (as recorded via GitHub). All authors contributed to the manuscript and participated in generating their site data.

Keywords

Data Use and Quality, Informatics, Electronic Health Record (EHR), Common Data Model

Disciplines

Medicine and Health Sciences

Authors
Vojtech Huser, *NIH (NLM)*; Frank J DeFalco, *Janssen Research & Development*; Martijn Schuemie, *Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States; and Observational Health Data Sciences and Informatics (OHDSI) New York, New York, United States*; Patrick B Ryan, *Janssen Research & Development*; Ning Shang, *Department of Biomedical Informatics, Columbia University, New York, USA*; Mark Velez, *Department of Biomedical Informatics, Columbia University, New York, USA*; Rae Woong Park, *Department of Biomedical Informatics, Ajou University, Suwon, Korea*; Richard D Boyce, *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA*; Jon Duke, *Regenstrief Institute, Indianapolis, IN*; Ritu Khare, *Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia*; Levon Utidjian, *Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia*; Charles Bailey, *Department of Biomedical and Health Informatics, Department of Pediatrics, The Children's Hospital of Philadelphia*.

eGEMs
Generating Evidence & Methods to improve patient outcomes

Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets

Vojtech Huser, MD, PhD;[i,ii] Frank J. DeFalco;[iii] Martijn Schuemie, PhD;[iii,iv] Patrick B. Ryan, PhD;[iii] Ning Shang, PhD;[v] Mark Velez, MD;[v] Rae Woong Park, MD, PhD;[vi] Richard D. Boyce, PhD;[vii] Jon Duke, MD, MS;[viii] Ritu Khare, PhD;[ix] Levon Utidjian, MD;[ix] Charles Bailey, MD, PhD[ix]

ABSTRACT

Introduction: Data quality and fitness for analysis are crucial if outputs of analyses of electronic health record data or administrative claims data should be trusted by the public and the research community.

Methods: We describe a data quality analysis tool (called Achilles Heel) developed by the Observational Health Data Sciences and Informatics Collaborative (OHDSI) and compare outputs from this tool as it was applied to 24 large healthcare datasets across seven different organizations.

Results: We highlight 12 data quality rules that identified issues in at least 10 of the 24 datasets and provide a full set of 71 rules identified in at least one dataset. Achilles Heel is a freely available software that provides a useful starter set of data quality rules with the ability to add additional rules. We also present results of a structured email-based interview of all participating sites that collected qualitative comments about the value of Achilles Heel for data quality evaluation.

Discussion: Our analysis represents the first comparison of outputs from a data quality tool that implements a fixed (but extensible) set of data quality rules. Thanks to a common data model, we were able to compare quickly multiple datasets originating from several countries in America, Europe and Asia.

[i]National Institute of Health, [ii]National Library of Medicine, [iii]Janssen Research & Development, [iv]Observational Health Data Sciences and Informatics, [v]Department of Biomedical Informatics, Columbia University, [vi]Department of Biomedical Informatics, Ajou University, [vii]Department of Biomedical Informatics, University of Pittsburgh, [viii]Regenstrief Institute, [ix]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia

Introduction

Data Ubiquity and Importance

Large health care databases are becoming increasingly important for clinical research and for a learning health care system.[1,2] Organizations typically try to capture longitudinal patient data on diagnoses, procedures, prescribed medications, laboratory results, and clinical notes. To enable advanced data analysis, data from disparate systems are integrated into a single analytical model (e.g., health plan data, electronic health record (EHR) data, and pharmacy dispensing data).[3] The informatics literature typically uses the term *source data* to refer to primary collected data and *target data* to refer to transformed or integrated output data. Data conversion from source to target is often referred to as the "extract, transform, and load" (ETL) process.

Data Quality

While ETL helps with data integration, it can also be a potential source of data quality issues when human mistakes are made in the ETL code. Most data transformation also occurs in multiple stages and can span multiple ETL code files written by a variety of developers and teams. Depending on the ETL process involved, *ETL data errors* typically affect all source system data or some consistent part of it, e.g., when birth dates of the mothers of newborns are incorrectly loaded into the newborn's record, or when a multisite data set has some subset of patients assigned to an incorrect location. A special type of an ETL data error is a *mapping error* that results from incorrect transformation of data from the source terminology (e.g., Korean national drug terminology) into the target data model's standard terminology for a given domain (e.g., RxNorm ingredient terms or Anatomic Therapeutic Class terms). Finally a third type of error is *source data error*, which occurs when the error is already present in the source data due to various causes, such as a human typo created during data entry or an incorrect default value assignment (e.g., birth year of 1900 assigned to patients with missing birth year data). Some source data errors may be typos, and those typically do not follow a consistent pattern.

In some cases source data errors may affect a large number of patients (e.g., missing coding or loss of data),[4] and it can be difficult or impossible to distinguish ETL errors from source data errors.[5,6] In recent years, the biomedical informatics community has increasingly adopted common data models (CDMs) shared across many organizations,[7] because they allow the same analytical code to be executed on multiple distributed data sets. In some cases, adherence to a CDM is a prerequisite for participating on a grant (or research network). Wider adoption of CDMs[8,9] also facilitates development of data quality tools that can be easily applied across multiple data sets.

Prior Literature

Data quality has been a subject of several past studies. Vlymen et al.[10] proposed a set of metadata to document data about data in a Primary Care Data Quality initiative in the domain of kidney disease patients. De Lusignan et al.[11] defined several data quality concepts and emphasized the ability to identify the origin of any data cell within the final analysis data set. Data quality is often addressed within established research networks, but the full methodology and the actual data quality evaluation scripts may be available only to researchers participating in the network. For example, Health Care Systems Research Network defined quality checks for data in their Virtual Data Warehouse.[12] Similarly, the Mini-Sentinel network[7] has defined a series of quality checks. Recently, the Data Quality Collaborative (DQC) published a 20-item list of data quality recommendations[13] that cover the areas of

Volume 4 (2016) Issue Number 1

(1) data capture documentation, (2) data processing and provenance documentation, (3) data elements profiling, and (4) analysis-specific data quality documentation. See Table 1 for areas descriptions and example recommendations. The DQC also advocates for publishing data quality metrics together with any observational data analysis, but points out that doing so can have unintended consequences, such as withdrawal of consortium data partners due to exposure of traditionally internal-only data quality indicators. Most recently, DQC proposed a data quality CDM[14] that builds on common elements of several prior data-quality frameworks.

Objective

In this paper, we compare outputs from the Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (called Achilles) Heel—a data quality tool developed by the Observational Health Data Sciences and Informatics (OHDSI)[15] collaborative as it was applied to 24 large health care data sets at seven different organizations. In contrast to previous studies, our analysis is not an attempt to introduce a new data quality framework, rather it builds on the previous classification created by the DQC.[13] Within the DQC's 20-item list, our study and the evaluated tool can be classified under three recommendations in the third area of data element characterizations: recommendation 13 (single element data descriptive statistics), recommendation 14 (temporal data quality checks), and recommendation 15 (multiple variables cross validations).

Background

Observational Health Data Sciences and Informatics (OHDSI) Consortium

The OHDSI is a multi-stakeholder, interdisciplinary collaborative that is striving to bring out the value of observational health data through large-scale analytics. OHDSI's vision is to build a research network that aggregates the data of one billion patients and to generate evidence about all aspects of health care.[16] In November 2014, OHDSI published version 5 of an OMOP Common Data Model (CDM) CDM that specifies a target data model that includes clinical domains of diagnoses, medications, procedures, laboratory results, clinical observations, clinical visits and clinical notes.[17] OMOP acronym stands for Observational Medical Outcomes Partnership and it was kept as the name of the model for historical reasons (previous versions of the model were developed under the OMOP project; OMOP project concluded in June of 2013).

Since October 2014, the OHDSI community has developed and maintained ACHILLES—a CDM-based data profiling tool.[18] The main function of ACHILLES is to generate high-level aggregate statistics to create a data-driven characterization of the population-level data in a database. A description and partial evaluation of an ACHILLES data quality subcomponent, called "ACHILLES Heel," is the main focus of this paper.

Data Quality Tool: ACHILLES Heel

ACHILLES Heel consists of a set of data quality rules (sometimes also referred to as "data quality checks") that generate a list of errors and warnings. Each error or warning includes an analysis ID, error message, and a count of erroneous data elements. An example output may be: "715|Distribution of days_supply by drug_concept_id; max (value=1,041 should not be > 180)," which indicates that the data contain an unusually high value of days_supply for a given drug concept. ACHILLES Heel version 1.1 (evaluated here) contains 26 rules; however, some of the rules generate multiple distinct errors or warnings.

Table 1. Data Quality Reporting Recommendations Formulated by the DQC[13] (shortened)

DATA QUALITY DOMAIN	DOMAIN DESCRIPTION	DATA QUALITY DOCUMENTATION RECOMMENDATIONS	EXAMPLE RECOMMENDATION
1. Data capture descriptions	Information on how data was observed, collected and recorded	Recommendations 1–6	#2 Data Steward: A description of the type of organization responsible for obtaining and managing the target data set (e.g., registry or state agency).
2. Data processing descriptions	Information on how data was transformed (e.g., mapping, unit conversion, derived values)	Recommendations 7–11	#8 Mappings from original values to standardized values: Documentation on how original data values were transformed to conform to the target data model format.
3. Data elements characterizations	Information on observed data features of the target data, such as data distributions and missingness	Recommendations 12–15	#13 Single element data descriptive statistics: For each variable, calculate the following descriptive statistics (count and % of missing, descriptive statistics for numerical and categorical variables, goodness-of-fit tests for anticipated distributions).
4. Analysis-specific data element characterizations	Information on data quality for a specific cohort and analysis (not on the level of the entire database)	Recommendations 16–20	#17 Data quality checks of key variables used for cohort identification: Study specific additions to recommendations #13–15.

Volume 4 (2016) Issue Number 1

ACHILLES Heel data are generated by the following two-step process.

Step 1. Precomputations

In this step, a series of Structured Query Language (SQL) queries generates aggregate interim data that are used by ACHILLES as well as by ACHILLES Heel. Each precomputed analysis has an analysis ID and a short description of the precomputed analysis, e.g., "715: Distribution of days_supply by drug_concept_id" or "506: Distribution of age at death by gender." In ACHILLES version 1.1, a total of 177 analyses are precomputed during this first step. Supplemental file S1 (available at http://dx.doi.org/10.6084/m9.figshare.1497942) provides a complete list of all ACHILLES analyses (for version 1.0). This step is computationally intensive and may take up to several hours to complete depending on the database engine used and the size of the CDM data set. The ACHILLES data model allows storing the results of this first step in a single table (ACHILLES_RESULTS) organized by up to five analysis dimensions (called "strata" within ACHILLES). Step 1 precomputed queries are not primarily driven by data quality questions but rather by the data visualization needs of the ACHILLES web application.

Step 1 precomputations (in file Achilles_v5.sql)[19] are largely guided by the CDM relational database schema and analyze most terminology-based data columns, such as condition_concept_id or place_of_service_concept_id (see a full list in the supplemental file S1 (ACHILLES version 1.0 or the commented SQL code on GitHub; GitHub is a collaboration platform for software projects).[19] The step 1 precomputations allow fast data-density visualizations and tabular views by data domain in general and by frequency of each individual event concept (such as diagnosis, procedure, medication, laboratory result, or observation; sometimes further stratified by age decile or gender). The set of precomputations may grow in future versions of ACHILLES. An overview of precomputed analyses for the current version of ACHILLES is maintained in an analysis overview CSV file.[20]

Step 2. Data Quality Rules

In the second step, the actual data quality rules are executed via SQL queries that typically utilize the data precomputed in the first stage (in file AchillesHeel_v5.sql).[19] ACHILLES Heel version 1.1 (evaluated here) does not have any overview of data quality rules, and only a review of the ACHILLES Heel SQL file (or generated output error strings) provides a comprehensive view. The list of analyses that are utilized by ACHILLES Heel version 1.1 rules is provided in the supplemental file S1 (third tab; 71 rows). The current ACHILLES Heel version 1.3 clarifies the relationship of data quality rules and analyses by introducing a rule_id and a rule overview CSV file.[21] It contains 34 additional rules added by the community. Currently, the rules are not organized in a hierarchy, but theoretically could be classified under DQC recommendation 14 (temporal data quality checks, e.g., rule 18: year of birth is in the future); DQC recommendation 15 (multiple variables cross validations, e.g., rule 2 on analysis 909: existence of drug events outside patient's observation period); and DQC recommendation 13 (single element data descriptive statistics, e.g., rule 1 on analysis 210: visit events with invalid care site ID). For rules that may generate multiple errors and warnings, it is only the combination of rule_id and analysis_id that fully communicates the data quality check being performed.

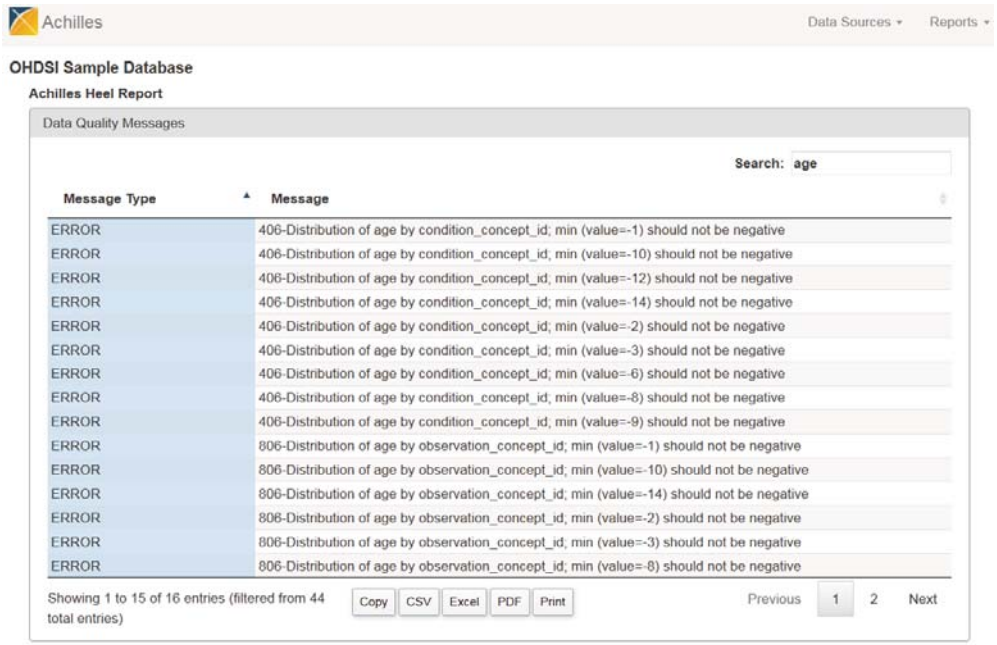
Step 2 results in a set of error and warning messages that are stored in a table "ACHILLES_HEEL_RESULTS," e.g., "ERROR: 506|Distribution of age at death by gender (count = 2); min value should not be negative." Individual errors and warnings can be subsequently displayed in the ACHILLES web

application[22] (see Figure 1). ACHILLES Heel output can also be exported for later comparison within or across sites. A live demonstration of the ACHILLES web application (including ACHILLES Heel results) is available at http://www.ohdsi.org/demos.

The three main principles guiding the construction of the Step 2 data quality queries were (1) prior experience of ACHILLES creators with data errors in CDM data sets encountered during individual data analyses; (2) identifying logical data contradictions (such as visit end date is prior to visit start date); and (3) checking for conformance to the CDM model specifications, such as use of correct target terminologies, e.g., Logical Observation Identifiers Names and Codes (LOINC) for laboratory results or Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for conditions. While the existing ACHILLES Heel set of quality rules represents a cumulative and iterative experience of multiple analysts with multiple data sets, it may not represent a comprehensive set that covers all possible aspects of data quality. Given recent increased focus on data quality in the informatics community, it relies on community contribution to the open ACHILLES Heel platform to further extend the rule set.

The current version of ACHILLES Heel is 1.3 (released June 15, 2016) . This study used previous version 1.1 released in January 2015. The tool is publicly available at GitHub as an open source tool (the current version is at https://github.com/OHDSI/Achilles with all previous versions also available within the Releases section of the GitHub repository). ACHILLES Heel is based on an earlier data quality tool—Observational Source Characteristics Analysis Report (OSCAR)—developed by the Observational Medical Outcomes Partnership (OMOP) community. In May 2015, a modified version of ACHILLES Heel (called "Heracles Heel") was developed to analyze a patient cohort instead of the whole CDM data set.

Figure 1: Screenshot Showing Viewing of ACHILLES Heel Errors and Warnings



Volume 4 (2016) Issue Number 1

Methods

ACHILLES Heel Output Comparison

We collected ACHILLES Heel comma separated value (CSV) output files from all seven sites participating in our data quality tool evaluation. As noted in Table 2, the site A data set consisted of claims data; the site B data set consisted of drug dispensing and administrative data; the sites C, F, and G data sets consisted of EHR data; and the sites D and E data sets consisted of claims plus EHR data. The ACHILLES Heel CSV output file contains no person-level information. Data from all sites were aggregated into a single list that preserved information about the site and the data set where the error was observed. We analyzed the combined list to find ACHILLES Heel's data quality assurance (QA) rules that "fired" across multiple data sets. For each QA rule, ACHILLES Heel provides the count of rows that violate that rule. A missing rule ID indicates that a given data set had no erroneous data rows for that data quality rule.

Questionnaire

In addition to ACHILLES Heel output comparison across sites, we did a structured email-based questionnaire for all participating sites to collect qualitative comments about the use and impact of ACHILLES Heel for data quality evaluation (questions 1 and 2) and general organizational data quality context (questions 3–5). The following questions were asked (with the first two questions being a general interview topic).

(1) Describe at what stage of your CDM implementation did you execute Achilles Heel analysis?
(2) Describe what was the impact of seeing the Achilles Heel results on your future ETL versions?
(3) How frequently do you refresh your CDM data and how frequently do you modify the ETL? What resources are allocated to this task? (e.g., number of man work hours per year; or percentage of employees' time is allocated to data quality).
(4) What other tools and methods does your site (or your site's specific dataset) use to assess data quality?
(5) Is there an ETL high level description document (similar to "Rabbit-In-a-Hat " documentation outputs)? Rabbit-In-a-Hat is another tool created by the OHDSI community that facilitates data mapping of local database schema to the Common Data Model schema.[23]

Data quality can be a sensitive topic to report publically or within a consortium. Kahn et al. suggested that in some cases it may lead to withdrawal from a data consortium.[13] To minimize the time required to answer the questions, we asked for very brief replies.

The main purpose of the questionnaire was to give us some limited insight (not a comprehensive view) into current data-quality related practices at a given site. A minor additional purpose was to solicit feedback that could guide future development of ACHILLES Heel.

Results

ACHILLES Heel Output Comparison

Table 2 provides an overview of health care organization sites included in our study. The types of sites included were single academic medical centers, a pharmaceutical industry research department, a clinical data research network, and a research program of a research foundation. The majority of sites provided data quality assessments for a single data set, while three sites provided data for multiple data sets. We use letters to refer to sites and numbers to refer to individual data sets, e.g., "siteD-dataset6" refers to the sixth data set at site

Table 2. Overview of Participating Sites

SITE	# OF DATA SETS	TYPE OF DATA INCLUDED
Site A	5	Claims data
Site B	1	Drug dispensing + administrative data
Site C	1	EHR data
Site D	7	Claims + EHR data
Site E	1	Claims + EHR data
Site F	1	EHR data
Site G	8	EHR data

D. Although it would be desirable to present more metadata about each data set beyond what Table 2 provides (such as the total size of the patient population, site geographic zone or continent, or other site metadata), this detailed information could lead to easy site reidentification. In consequence, this could significantly lower the number of sites participating in our evaluation, or could have other unintended consequences, such as site withdrawal from the consortium.[13] However, data set sizes (for all sites adopting the OHDSI CDM model, not just those in our study) are available on the OHDSI wiki site.[24]

Even though we preserved the site–data set affiliation, the main comparison of the ACHILLES Heel results was done at the data set level, rather than at the site level. If a site provided two sets of ACHILLES Heel results (an initial report and a revised report after ETL improvements had been made), we used the initial report. This was because we hope to identify a subset of rules that are most relevant to organizations during the initial data quality evaluations. Using the initial ACHILLES Heel report allows us to better characterize real-world data sets and to identify the most common ETL errors. It was out of the scope of our study to investigate any dynamics of the quality monitoring process and how ETL errors evolve over time.

ACHILLES Heel provides two types of data quality outputs: *errors* and *warnings*. "Errors" represent more serious data quality errors, while "warnings" point to data issues anticipated to have smaller impact. This analysis focuses only on errors and completely excludes warnings. The number of errors per data set ranged from 3 to 104,100 items. Table 3 shows the number of errors for each analyzed data set. The "ACHILLES Heel Execution Context" column indicates at which point ACHILLES Heel was executed. Although we asked each site to provide the earliest possible ACHILLES Heel results (ideally after initial ETL code was written), at many sites ACHILLES Heel was available only after the majority of their ETL coding was completed. At some sites, re-execution of ACHILLES Heel may have guided revisions of the ETL, while at other sites (indicated by the words "without Heel results") ACHILLES Heel was not re-executed at ETL development iterations.

The median number of errors was 19. ACHILLES Heel data from site A showed a much larger volume of errors compared to all remaining sites (B–G). A high proportion of site A errors (e.g., 94 percent for siteA-data set3 or 98 percent for siteA-dataset4) were caused by QA rules requiring nonnegative amounts in cost columns (copay, co-insurance, or total

Volume 4 (2016) Issue Number 1

amount paid) for drugs and procedures with further stratification by the erroneous value.

Due to multiple factors such as documentation, shifted data set priorities, research mode focused on methods research, and a 2016 upgrade to CDM version 5 with revised ETL (our study was executed in 2015 on CDM version 4, prior to this major change to version 5), we performed only a limited analysis of the large number of errors at site A. If we exclude site A's data sets 1, 3, 4, and 5 with their vastly greater number of errors (mostly due to negative copay, co-insurance, and total amount paid), the median number of errors was 17.

The merged data set of all errors from all sites contained 228,781 rows. When site A's data sets 1, 3, 4, and 5 are excluded, the merged data set has only 982 rows. Table 4 lists the most common data quality errors identified in at least 10 data sets (see second column). Supplemental file S1 (available at http://dx.doi.org/10.6084/m9.figshare.1497942) provides a complete list of all 71 errors found in at least one data set.

In Table 4, a review of the most common errors shows that many may be related to the same underlying error in birth year. It may be redundant to report the same problem multiple times; however, many data quality checks mirror the structure of the underlying analysis and this multiplicity may help identify the subdomain of the problem, e.g., all cases of implausible age-at-procedure may be clustered within a set of procedure codes (possibly pointing to the source of the error). Table 4 contains a count of data sets with error and a count of all error instances. ACHILLES Heel rules can generally be classified into two categories. In the first category are rules that operate on the data set level and generate zero or one error instance per data set (such as "Number of procedure occurrence records outside valid observation period"). In the second category are rules that operate on the data set plus an additional stratum level (e.g., condition_concept_id). For the second category of rules, the total count of instances can be more than one per data set (shown in the third column of Table 4). The supplemental file S1 provides a view of errors when ordered by the count of error instances. As noted in the Step 2 Data Quality Rules description of ACHILLES Heel data (in the 1.5.2 subsection of this paper), it would be possible to classify these errors by the DQC recommendations (or other DQC defined classifications), but the ACHILLES tool (neither version 1.1 nor the current version 1.3) is not formally attempting to do that.

Questionnaire

We received questionnaire responses from all seven organizations (100 percent response rate). Most sites provided one- or two-sentence answers for each question, with the longest response being four sentences. Staff at one site, in addition to their structured responses, provided a redacted internal report elaborating on each data error that they considered significant for follow-up.

Because we sought very brief replies, we did not analyze the responses using formal qualitative methods or qualitative analysis software. Table 5 presents the questionnaire findings classified by the type of site. The following section provides a brief summary of the survey responses.

Most sites executed ACHILLES Heel after their first ETL process. For some sites, ACHILLES Heel was created after they completed several iterations of ETL. In terms of ACHILLES Heel impact, most sites found ACHILLES Heel output very informative. Many sites turned each issue identified by ACHILLES Heel into a ticket item in their issue-tracking system for creating a modified and improved ETL code. The intent was to eliminate all or some of the ACHILLES Heel errors and warnings by revising the ETL code.

Table 3. Overview of Data Sets (Number of Heel Errors and Context Characteristics)

DATA SET	# OF ERRORS	ACHILLES HEEL EXECUTION CONTEXT	DATA SET SIZE CATEGORY (# OF PATIENTS)
siteA-data set1	104,125	after multiple ETLs without Heel results	1M+
siteA-data set2	243	after multiple ETLs without Heel results	1M+
siteA-data set3	22,289	after multiple ETLs without Heel results	1M+
siteA-data set4	58,296	after multiple ETLs without Heel results	1M+
siteA-data set5	43,089	after multiple ETLs without Heel results	1M+
siteB-data set1	39	after initial ETL	<10k
siteC-data set1	424	after multiple ETLs	1M+
siteD-data set1	19	after multiple ETLs without Heel results	1M+
siteD-data set2	13	after multiple ETLs	1M+
siteD-data set3	7	after multiple ETLs	1M+
siteD-data set4	25	after multiple ETLs	1M+
siteD-data set5	19	after multiple ETLs	10k–100k
siteD-data set6	3	after multiple ETLs	10k–100k
siteD-data set7	22	after multiple ETLs	1M+
siteE-data set1	31	after multiple ETLs	1M+
siteF-data set1	25	after multiple ETLs	1M+
siteG-data set1	17	after multiple ETLs	10k–100k
siteG-data set2	16	after multiple ETLs	10k–100k
siteG-data set3	16	after multiple ETLs	10k–100k
siteG-data set4	12	after multiple ETLs	10k–100k
siteG-data set5	14	after multiple ETLs	10k–100k
siteG-data set6	13	after multiple ETLs	10k–100k
siteG-data set7	15	after multiple ETLs	10k–100k
siteG-data set8	9	after multiple ETLs	10k–100k

Table 4. Most Common Errors Found

ERROR ID	COUNT OF DATA SETS WITH ERROR	COUNT OF ALL ERROR INSTANCES	ERROR DESCRIPTION
101*	16	n/a#	Number of persons by age, with age at first observation period; should not have age < 0
103*	15	n/a	Distribution of age at first observation period; age should not be negative
206*	13	18	Distribution of age by `visit_concept_id`; age should not be negative
406*	13	31	Distribution of age by `condition_concept_id`; min(age) should not be negative
600	13	14	Number of persons with at least one procedure occurrence, by `procedure_concept_id`; concepts in data are not in correct vocabulary (CPT4; HCPCS,ICD9P)
717	12	3173	Distribution of quantity by `drug_concept_id`; max(quantity) should not be > 600
114*	11	n/a	Number of persons with observation period before year of birth; should not be > 0
410	11	n/a	Number of condition occurrence records outside valid observation period; should not be > 0
510	11	n/a	Number of death records outside valid observation period; count should not be > 0
806*	11	25	Distribution of age by `observation_concept_id`; should not be negative
606*	10	19	Distribution of age by `procedure_concept_id`; min(age) should not be negative
610	10	n/a	Number of procedure occurrence records outside valid observation period; count should not be > 0

Notes: *Errors marked with an asterisk are all possibly related to the same underlying error in birth year. #n/a indicates that the rule operates on the whole data set and the number of instances is not applicable (the rule can generate only one instance per database, and the count of instances for that rule is always equal to the "count of data sets with error" shown in the second column).

When asked about how frequently CDM data sets are refreshed, the answers ranged from never (static CDM data; one site) to biweekly, with most sites refreshing it once a year. The amount of resources dedicated to initial data quality evaluation or ongoing data quality monitoring also varied widely. At one site where a CDM data set is tied to a health information exchange (HIE), data quality is monitored by a committee of five people that meets monthly. All sites used additional data quality tools besides ACHILLES Heel. Two sites compare the overall data volume in the source and target CDM data, and investigate significant variation in volume trends over time. Other approaches include site-specific data quality scripts written in SQL or R.

Two sites reported active use of OHDSI data mapping tool (called "WhiteRabbit" and "Rabbit-In-a-Hat") to document their ETL.[25] Other approaches include a wiki-based documentation, internal documents, and publicly available ETL documentation on the web.[26]

Discussion

ACHILLES provides a novel approach of initial data aggregation and data quality assessment. This approach allows population data-quality analyses that avoid privacy limitations imposed on patient-level data. While other data-quality evaluations scripts exist, they tend to be nonpublic and restricted to consortia members. In contrast, ACHILLES and ACHILLES Heel (including the source code) are freely available to the general public through GitHub. Another difference from prior data-quality frameworks is the provision of a basic viewer (AchillesWeb component)[22] in addition to just providing computer code that evaluates data quality. However, note that ACHILLES Heel by itself is not able to correct the underlying data quality errors. It merely identifies and quantifies potential issues, and manual revisions of the ETL process are required.

The set of data quality rules currently offered by ACHILLES Heel is likely to evolve in the future with new data rules suggested by the OHDSI community. In fact, arriving at a comprehensive and validated set of data quality rules was out of the scope of this study, and the emphasis was on demonstrating multisite data-quality evaluation and comparing outputs (given a specific set of rules). The ACHILLES architecture allows the addition of new quality checks by editing the underlying SQL file[19] without the need to change the displaying web tool or to significantly change the overall software architecture.

Evaluation of population-level data quality using a CDM approach is a relatively new area of informatics research that has been greatly facilitated by the recent broader adoption of CDMs. It was out of the scope of our study to compare ACHILLES Heel to other data quality tools, such as Mini-Sentinel or Health Care Systems Research Network, mainly because of the extensive data transformations that would have been required, and the lack of clear documentation and full public availability.

Our comparison of ACHILLES Heel output across seven sites showed that different data sets may vary widely in terms of number of errors detected (ranging from 7 errors to thousands of errors found). Classification of data quality rules into categories by more granular error type may help in dealing with this wide variation. The questionnaire revealed that, at many sites, ACHILLES Heel output can trigger targeted ETL code investigations that may translate into data quality improvements. The typical ACHILLES Heel implementation was to execute it after each data set refresh (new data) or ETL code change (new code). It also indicated that the majority of sites had some existing data-quality assessing tools that may serve as inspiration for additional future data-quality checks (contributed by sites using ACHILLES Heel) to be incorporated into the next versions of ACHILLES Heel.

Volume 4 (2016) Issue Number 1

Table 5. Data Quality Questionnaire Results

CATEGORY	QUESTION	CLAIMS DATA (SITE A)	DRUG DISPENSING + ADMINISTRATIVE DATA (SITE B)	EHR DATA (SITES C, F, G)	CLAIMS + EHR DATA (SITES D, E)
Data Quality Evaluation	Q1: Describe at what stage of your CDM implementation did you execute ACHILLES Heel analysis?	Heel was executed 1.5 years after the CDM data set was created.	Use Heel iteratively during translation and loading process	• After first iteration of the full ETL. • During ETL and the end. • Run custom DQA scripts and Heel to identify DQA issues -> communicate back to site -> sites fix ETL issues and resend data -> DQA analysis	• After each data update and each change to our CDM implementation. • Iteratively during translation and loading process.
	Q2: Describe what impact had seeing the ACHILLES Heel results on your future ETL versions?	None. ETL is static.	Able to identify serious problems with ETL and fix the issues.	• Motivated to set cut-offs for outliers leading to invalid data, and to revise our observation period logic. • Detect ETL errors, improve ETL scripts and learn pediatric-specific data issues for better understanding of data.	• Provide data quality check for each ETL version and further analysis and understanding ETL as well as data. Provide substantial feedback for future ETL versions. • Made 1,500 lines SQL codes for the ETL to fix all the bugs encountered by Heel.
General Organizational Data Quality Context	Q3: How frequently do you refresh your CDM data and how frequently do you modify the ETL? What resources are allocated to this task?	None. ETL is static.	1 time per two years 10% of an FTE	• CDM monthly • ETL after problems are detected • 15–30 days for both • Two FTEs currently <10%, but we expect to increase once our version 5 CDM stabilizes. • Monitoring committee meets monthly to discuss data quality across health information exchange (HIE).	• ETL based on feedback from Heel, changes to source data or updates to CDM model CDM on a quarterly basis • CDM: first implementation took a year, second renewal after 4 months of the first implementation, which took 2 weeks ETL: first DQM and ETL took a month (4 people), DQM took a month by a single person, second DQM & ETL took 2 weeks by a single person
	Q4: What other tools and methods are your site (or your site's specific data set) using to assess data quality?	SQL queries	Public quality measures from CMS Nursing Home Compare and with data provided by PBM at our request	• Frequently run variants of the ETL and compare the resulting ACHILLES to find the best approach. Also consult senior clinical and technical staff for validity. If it's programmer error, issue tracking software is important. • Biostatistician monitors • A suite of DQA scripts in R	• A custom system that captures benchmarks of data volumes by table within each data source. The system can compare current and prior versions to show discrepancies and variation in volume trends within each table. • Before Heel, all researchers took care of their own data quality analyses.
	Q5: Is there an ETL high-level description document?	on OHDSI website	Working on one to have public before submitting a manuscript	• WhiteRabbit and Rabbit-In-A-Hat • In the form of a wiki • Site-specific ETL documents for 8 sites and a common conventions document to populate the OMOP	• WhiteRabbit and Rabbit-In-A-Hat • OMOP ETL template and prepared own ETL SQL code

Limitations

Our evaluation revealed some limitations of ACHILLES Heel. First, the version of ACHILLES used in our evaluation did not distinguish individual data-quality checks executed on aggregated analyses. As a result of our work, we have modified the ACHILLES Heel SQL code to introduce an individual rule identifier.

Second, due to a lack of unified structure across all data quality rules for outputting how many data set rows were erroneous, we were not able to provide an average degree of given data error across data sets (e.g., how many patients on average across all data sets had data prior to birth). A revised code has been submitted to the community that proposes a pipe delimited structure that allows recording of this data in a structured fashion.

Third, all ACHILLES Heel QA rules operate on precomputed counts (ACHILES_RESULTS table), which may be a limitation for authors of new data-quality checks who want to use raw CDM data. In March 2015, the OHDSI community introduced an additional tool (called Iris) IRIS,[27] that operates directly on CDM tables and provides an alternative avenue to such authors to implement new data size and quality computations.

Fourth, our evaluation and the ACHILLES Heel tool are limited to a single data model (i.e., OHDSI CDM). Our study findings are limited to this model and may not generalize fully to other models and schemas. Researchers with data in other data models, such as Informatics for Integrating Biology and the Bedside (i2b2) relational database schema or local schemas (e.g., Stanford University[28] or Duke University[29]), must first translate the rules specified in the ACHILLES Heel main rule file (coded in SQL)[19] into their data schema or model. However, even though a translation into a different model is required, ACHILLES Heel offers an introductory set of rules (every ACHILLES Heel rule in the main rule file[19] includes comments that explain the intent of that rule) that may be used as a starting point for data quality evaluation. Moreover, the current study further highlights rules that were instrumental across several data sets that might be of higher priority to translate.

Finally, our study used a version (1.1) of ACHILLES that was current at the time of the analysis (second quarter of 2015). Since then, version 1.2 (from April 2016) introduced a rule_id for each rule and a separate column that quantifies the extent of the error (count of rows or patients with the error). Another version (1.3, from June 2016) introduced new analyses and rules, a new type of Heel output (notification), and a new table for derived analyses utilized by some of the new rules. Similarly, the version of the OMOP CDM had an impact on our study. Since our analysis, conducted exclusively on CDM version 4 data sets, many sites transitioned to CDM version 5. ACHILLES version 1.1 was the first release that introduced support for CDM version 5. New features in ACHILLES version 1.3, however, are available only to sites that migrated to the CDM version.

Generalizability and Future Work

The findings of our questionnaire offer very limited insights into how CDM data quality may be implemented nationally or internationally. We aimed at a brief pilot survey and made no attempt to link survey questions to each other. Our questionnaire is, to our knowledge, the first attempt to survey institutional resources allocated to data quality assessments (targeting CDM-shaped data maintained for research purposes instead of data in a native local data model). The implication of our work for a site that did not use any prior data-quality tools is that most likely this site will discover at least some of the common data-quality issues seen across

Volume 4 (2016) Issue Number 1

our seven sites. The availability of a tool, such as ACHILLES Heel, to evaluate data quality may lead to increased attention to data quality by sites that would otherwise implement a very limited set of data quality checks (or perform no such checks).

In terms of future work, we plan to further improve ACHILLES Heel's set of rules and the addition of the necessary precomputed analyses. We also hope—in future versions of ACHILLES Heel—to distinguish rules that check conformance to a particular CDM (from OHDSI) from more general data quality rules. Similarly, detailed rules about data plausibility—e.g., number of male patients with hysterectomies is reasonably low (in transgender patients)—require a choice of a terminology, e.g., a value set of Systematized Nomenclature of Medicine (SNOMED) or Current Procedural Terminology (CPT) codes for procedures, gender codes specific to OHDSI CDM. Future versions may flag ACHILLES Heel's data quality rules that make such assumptions to better facilitate generalizability of the rule set to other CDMs.

Note that our study focused on only 3 out of all 20 DQC recommendations (see Table 1).[13] Further studies in documenting, measuring, and evaluating data quality[30-33] are clearly needed to address the remaining recommendations. Outside the ACHILLES Heel tool, we can name several other OHDSI efforts that are relevant to additional DQC recommendations. For example, recommendation 8 ("Mapping from original values to standardized values") is in one special case addressed by the OHDSI CDM Vocabulary tables (sometimes referred to as "Athena").[34] This special case includes a data mapping scenario where source data is coded directly in one of the terminologies fully integrated within the Athena vocabulary tables—e.g., United States National Drug Codes (NDC). In this case, Athena vocabulary clearly documents mapping from NDC drug codes to RxNorm drug codes (RxNorm is one of the formal OHDSI target standards). OHDSI also provides documentation[17] that fully addresses recommendation 5 ("Description of target database model/data set structure") with their public database schema documentation.

Conclusion

ACHILLES Heel is a free, open-source data-quality tool that provides a predefined set of quality checks for data sets in CDM format. Our evaluation of ACHILLES Heel's set of QA rules identified 12 common rules that found errors across several data sets and a more complete set of 71 rules that found errors in at least one evaluated data set. Responses to our questionnaire indicate that ACHILLES Heel can provide a useful starting set of data quality rules to organizations that begin to implement formal data-quality assessments. This comparison of multiple sites triggered several new improvements to the tool.

References

1. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc*. 2015;22(1):43-50.
2. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793-795.
3. Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *EGEMS (Wash DC)*. 2014;2(1):1110.
4. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Family Practice*. 2015;16(1):1-8.
5. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011;3(79):79re71.
6. Carter J. EHR Science: EHR Data Quality (collection of relevant references; updated March 2016). 2016; http://ehrscience.com/ehr-data-quality. Accessed March 11, 2016.
7. Psaty BM, Breckenridge AM. Mini-Sentinel and regulatory science--big data rendered fit and functional. *N Engl J Med*. 2014;370(23):2165-2167.

8. Xu Y, Zhou X, Suehs BT, et al. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. *Drug safety : an international journal of medical toxicology and drug experience*. 2015;38(8):749-765.
9. Zhou X, Murugesan S, Bhullar H, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. *Drug safety : an international journal of medical toxicology and drug experience*. 2013;36(2):119-134.
10. van Vlymen J, de Lusignan S. A system of metadata to control the process of query, aggregating, cleaning and analysing large datasets of primary care data. *Inform Prim Care*. 2005;13(4):281-291.
11. de Lusignan S, Liaw ST, Krause P, et al. Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. Contribution of the IMIA Primary Health Care Informatics Working Group. *Yearb Med Inform*. 2011;6(1):112-120.
12. Ross TR, Ng D, Brown JS, et al. The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. *EGEMS (Wash DC)*. 2014;2(1):1049.
13. Kahn MG, Brown JS. Transparent Reporting of Data Quality in Distributed Data Networks. *EGEMS (Wash DC)*. 2015;3(1).
14. Health A. Data Quality Collaborative. 2015; http://repository.academyhealth.org/dqc/. Accessed May 15, 2015.
15. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in health technology and informatics*. 2015;216:574-578.
16. OHDSI. OHSDI Mission, Vision & Values. 2014; http://www.ohdsi.org/who-we-are/mission-vision-values/. Accessed Sep 10, 2015.
17. OHDSI. Common Data Model. 2015; http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:single-page. Accessed June 10, 2015.
18. OHDSI. Observational Health Data Sciences and Informatics Wiki: Achilles Documentation. 2015; http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:achilles. Accessed June 22, 2015.
19. OHDSI. Code in Structured Query Language that contains all Achilles Heel data quality rules (CDM v5 and v4 variants). 2015; https://github.com/OHDSI/Achilles/tree/master/inst/sql/sql_server. Accessed June 15, 2015.
20. OHDSI. List of Achilles precomputed analyses. 2016; https://github.com/OHDSI/Achilles/blob/master/inst/csv/analysisDetails.csv. Accessed Feb 15, 2016.
21. OHDSI. List of Achilles Heel Rules (Achilles Heel version 1.2). 2016; https://github.com/OHDSI/Achilles/blob/master/inst/csv/achilles_rule.csv. Accessed March 10, 2016.
22. Achilles Web software (interactive web site for reviewing the results of the Achilles and Achilles Heel analysis). 2014; https://github.com/OHDSI/Achillesweb. Accessed Sep 15, 2015.
23. OHDSI. Introduction to using 'White Rabbit' and 'Rabbit In A Hat' tools for facilitating of loading your local data into OHDSI Common Data Model repository. 2015.
24. OHDSI. Data network overview. 2016; http://www.ohdsi.org/web/wiki/doku.php?id=resources:data_network. Accessed Feb 16, 2016.
25. OHDSI. Observational Health Data Sciences and Informatics Wiki: Documentation. 2015; http://www.ohdsi.org/web/wiki/doku.php?id=documentation:overview. Accessed June 22, 2015.
26. OMOP. Observational Medical Outcomes Partnership: ETL Documentation. 2014; http://omop.org/CDM. Accessed June 22, 2015.
27. OHDSI. IRIS: Data size comparison tool. 2015; https://github.com/OHDSI/Iris. Accessed July 16, 2015.
28. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*. 2009;2009:391-395.
29. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. *Journal of biomedical informatics*. 2011;44(2):266-276.
30. Dentler K, Cornet R, ten Teije A, et al. Influence of data quality on computed Dutch hospital quality indicators: a case study in colorectal cancer surgery. *BMC medical informatics and decision making*. 2014;14:32.
31. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc*. 2002;9(6):600-611.
32. 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.
33. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51(8 Suppl 3):S22-29.
34. OHDSI. ATHENA standardized vocabularies. 2016; http://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/. Accessed Feb 19, 2016.