OHDSI Data Quality Dashboard: Initial Design Documentation

# Statement of Work

To design and develop an open-source tool that takes as input a data object that represents the OMOP CDM specification with associated attributes that correspond to CDM conventions or expectations, that should be able to be translated into data quality tests, which can then be executed against any CDM instance to produce results in the form of a dashboard report. The underlying structure of this tool will be an R package that dynamically generates the data quality (DQ) checks of interest in SQL, based on some input that will direct how the checks are written. This package should make use of [SQL Render](https://github.com/ohdsi/sqlrender) to make the SQL executable across multiple platforms.

## Purpose

The purpose of this document is to fully specify the design of the above described Data Quality Dashboard. Upon completion it will be handed off to a development team lead by Ajit Londhe with the goal to finalize the project by the date of the 2019 OHDSI Symposium (September 15-17).

## Intended Audience

TBD

# Background

The question of data quality has been almost constant as real-world data (RWD) and real-world evidence (RWE) have become popular tools with which to influence regulatory decision making. An FDA guidance document from 2013 recommends certain topics that should be addressed by investigators such as any updates in coding practices during the course of the study and evaluation of missingness over time (1). The Sentinel Initiative[[1]](#footnote-0) used the topics outlined in the 2013 guidance document as a broad framework for what they call “Data Management and Quality Assurance” (2). As stated in the document approximately 1,200 data checks are assessed, and each is designated as a “level 1”, “level 2”, “level 3”, or “level 4”. This description of the Sentinel process is helpful though it is still high-level and does not detail the individual checks themselves.

To that end, the recently published paper by Qualls, et. al. nicely describes the 314 data quality checks employed by PEDSnet(3). These checks follow the data quality categories of conformance, plausibility, and completeness found in the Kahn, et. al. paper which is widely accepted as the framework into which data quality checks should be organized(4). Kahn and his colleagues, however, go beyond the categories and break quality checks down in to contexts, subcategories, and definitions.

It is our goal within OHDSI to apply all levels of the Kahn framework to our design of a data quality dashboard. This document will provide the description and number of checks that fall within each combination of category, subcategory, context, and definition with the full list of all checks available in the accompanying excel file.

## Previous work

There have been a number of prior and current initiatives in the OHDSI community to determine data quality. This work will build further on these initiatives:

* Achilles Heel (before: OSCAR)
* THEMIS conventions
* Dataprint for database comparison (by DeFalco)

Recent OHDSI-wide DQ initiative

# Definitions

As described in the Kahn framework, below are definitions and descriptions of the data quality terms used throughout this document.

1. Data Quality Category: A DQ concept that needs to be interpreted within a certain context (verification or validation). These categories are conformance, completeness, and plausibility.
2. Verification: Focuses on how data values match expectations with respect to metadata constraints, system assumptions and local knowledge.
3. Validation: Focuses on the alignment of data values with respect to relevant external benchmarks.
4. Conformance: A category that focuses on DQ features that describe the compliance of the representation of data against internal or external formatting, relational, or computational definitions. This contains the subcategories value conformance, relational conformance, and computational conformance
   1. Value Conformance: Seeks to determine if recorded data elements are in agreement with a prespecified, constraint-driven data architecture.
   2. Relational Conformance: Seeks to determine if the recorded data elements are in agreement with additional structural constraints imposed by the physical database structures that store data values.
   3. Computational Conformance: Focuses on the correctness of the output value of calculations against technical functional specifications.
5. Completeness: A category that focuses on features that describe the frequencies of data attributes present in a data set without reference to data values.
6. Plausibility: A category that focuses on features that describe the believability of truthfulness of data values. This contains the subcategories uniqueness plausibility, atemporal plausibility, and temporal plausibility.
   1. Uniqueness Plausibility: Seeks to determine if objects appear multiple times in settings where they should not be duplicated.
   2. Atemporal Plausibility: Seeks to determine if observed data values, distributions, or densities agree with “common” logic or a gold standard.
   3. Temporal Plausibility: Seeks to determine if time-varying variables change values as expected based on known temporal properties or across one or more external comparators or gold standards.

# Proposed Tasks

1. Run a script to get a summary statistic based on the data quality check and what it is meant to measure
2. Compare the computed summary statistic to some decision threshold
   1. This should be editable by the end user
3. Create a human readable list with the text describing the check, the summary statistic, the threshold, and whether the check passed or failed

# Data Quality Checks

Table 1 shows all checks that have been proposed to inform the DQ dashboard.

1. How should ranges be determined?
2. Are these all the checks we want for a first pass?
3. Is this how the checks should be determined?

All checks will be computed first and then a specified threshold can be applied after – this means that all checks can be run at all sites even if different sites choose different thresholds.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Category** | **Con-text** | **Subcategory** | **Definition** | **DESCRIPTION** | **Example** | **Number of Checks** |
| **Completeness** | **Validation** |  | a. The absence of data values at a single moment in time agrees with trusted reference standards or external knowledge. |  |  | **0** |
|  | b. The absence of data values measured over time agrees with trusted reference standards or external knowledge |  |  | **0** |
| **Verification** |  | a. The absence of data values at a single moment in time agrees with local or common expectations. | Should a field be evaluated for it's completeness? | How many observation periods do not have an observation\_period\_start\_date? | **185** |
|  | a. The absence of data values at a single moment in time agrees with local or common expectations. | Should a mapped field be evaluated for the percentage that was mapped to zero? | How many records in the OBSERVATION table have an OBSERVATION\_CONCEPT\_ID = 0? | **26** |
|  | b. The absence of data values measured over time agrees with local or common expectations. |  |  | **0** |
| **Conformance** | **Validation** | **Computational** | a. Computed results based on published algorithms yield values that match validation values provided by external source. |  |  | **0** |
| **Relational** | a. Data values conform to relational constraints based on external standards. | Are all required fields not null? | Are all OBSERVATION\_PERIOD\_ID values not null? | **168** |
| a. Data values conform to relational constraints based on external standards. | Are all required tables present? | Is the PERSON table present? | **2** |
| **Value** | a. Data values conform to representational constraints based on external standards. |  |  | **0** |
| **Verification** | **Computational** | a. Computed values conform to computational or programming specifications. | Are all drug eras computed at the ingredient level? |  | **2** |
| **Relational** | a. Data values conform to relational constraints. | Do foreign key fields respect the foreign key - primary key relationship? | Are all CONDITION\_CONCEPT\_IDs in the CONDITION\_OCCURRENCE table found in the CONCEPT table? | **141** |
| b. Unique (key) data values are not duplicated. | Do primary key fields repect the primary key constratins? | Are all PROCEDURE\_OCCURRENCE\_ID values unique? | **22** |
| c. Changes to the data model or data model versioning. | Are all fields present as specified in the target CDM version? | If a CDM states it is written in v6.0, are all datetime values required? | **338** |
| **Value** | a. Data values conform to internal formatting constraints. | Do the fields in a CDM follow the constraints detailed on the CDM wiki? | Are all PERSON\_ID values integers? | **338** |
| b. Data values conform to allowable values or ranges. | Do the values in <entity>\_CONCEPT\_ID fields contain concepts from the specified domain? | Do all values in the GENDER\_CONCEPT\_ID field belong to the gender domain? | **28** |
| **Plausibility** | **Validation** | **Atemporal** | a. Data values and distributions (including subgroup distributions) agree with trusted reference standards or external knowledge. | Is the prevalence of certain conditions higher than a plausible low value? | How should this be calculated? | **45** |
| a. Data values and distributions (including subgroup distributions) agree with trusted reference standards or external knowledge. | Is the prevalence of certain conditions lower than a plausible high value? | How should this be calculated? | **45** |
| b. Similar values for identical measurements are obtained from two independent databases representing the same observations with equal credibility. |  |  | **0** |
| c. Two dependent databases (e.g., database 1 abstracted from database 2) yield similar values for identical variables. |  |  | **0** |
| **Temporal** | a. Observed or derived values have similar temporal properties across one or more external comparators or gold standards. |  |  | **0** |
| b. Sequences of values that represent state transitions are similar to external comparators or gold standards. |  |  | **0** |
| c. Measures of data value density against a timeoriented denominator are expected based on external knowledge |  |  | **0** |
| **Uniqueness** | a. Data values that identify a single object in an external source are not duplicated |  |  | **0** |
| **Verification** | **Atemporal** | a. Data values and distributions agree with an internal measurement or local knowledge. | How many values are below the plausible lowest value? | How many people have a birth year prior to 1850? | **61** |
| a. Data values and distributions agree with an internal measurement or local knowledge. | How many values are above the plausible highest value? | How many people have a birth year after the current calendar year? | **54** |
| a. Data values and distributions agree with an internal measurement or local knowledge. | How many values for a particular concept are below the plausible lowest value? | How many people have a hemoglobin A1c above 15? | **8** |
| a. Data values and distributions agree with an internal measurement or local knowledge. | How many values for a particular concept are above the plausible highest value? | How many people have a hemoglobin A1c below 2? | **8** |
| b. Data values and distributions for independent measurements of the same fact are in agreement. |  |  | **0** |
| c. Logical constraints between values agree with local or common knowledge (includes “expected” missingness). | How many concepts are associated with a person of the wrong gender? | How many females in the database have a record for a prostate exam? | **288** |
| d. Values of repeated measurement of the same fact show expected variability. |  |  | **0** |
| **Temporal** | a. Observed or derived values conform to expected temporal properties. | Do fields in the same table relate to each other temporally? | How many visit\_end\_dates occur after visit\_start\_dates in the VISIT\_OCCURRENCE table? | **17** |
| b. Sequences of values that represent state transitions conform to expected properties | Do records occur within a patient's lifetime? | How many observation periods occur between a patient's date of birth and date of death? | **239** |
| c. Measures of data value density against a timeoriented denominator are expected based on internal knowledge. | Do number of records per person in a given table remain relatively constant over time? |  | **24** |
| c. Measures of data value density against a time oriented denominator are expected based on internal knowledge. |  |  | **0** |
| **Uniqueness** | a. Data values that identify a single object are not duplicated. |  |  | **0** |
| **TOTAL** | |  |  |  |  | **2039** |

Table 1: Number of types of data quality checks proposed for version 1 of the data quality dashboard

Excel file: These checks were quantified based on this file

# User Interface Design

For the first pass at a UI design to present at the 2019 OHDSI Symposium the goal will be to show a table similar to the results table that will be created by the back-end SQL. It should have at least one filter at the pass/fail level so a user can view only the checks that failed. Additionally, there should be a summary table showing the number of checks by Kahn category and the number and percent that passed. Figure 1 gives an example of how these two tables.

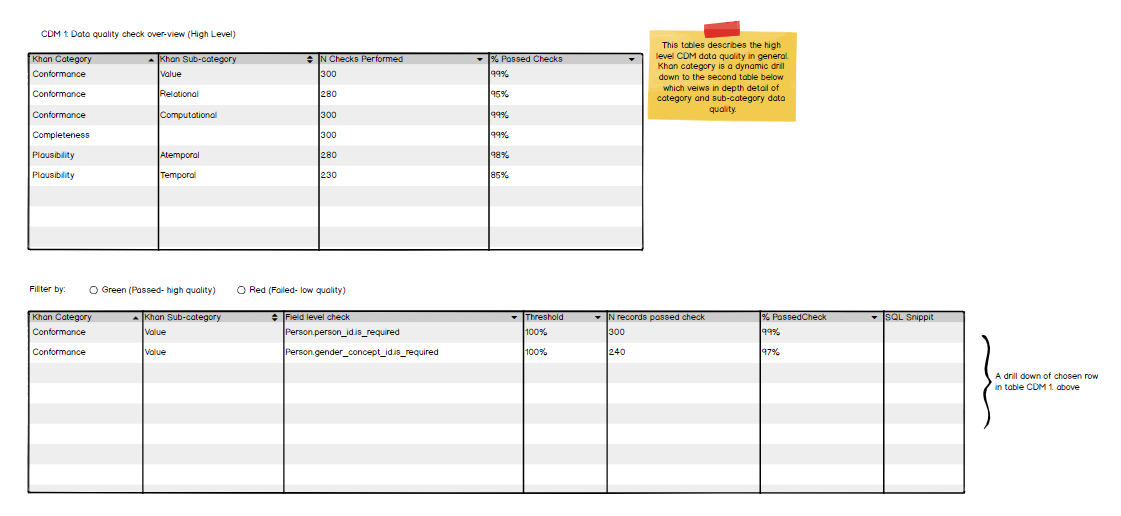


Figure 1: Example of initial UI design for the OHDSI data quality dashboard

# References

1. FDA guidance for industry: *Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data* 2013. Available from: https://[www.fda.gov/media/79922/download](http://www.fda.gov/media/79922/download).

2. Sentinel Operations Center. Sentinel Data Quality Assurance Practices 2017. Available from: https://[www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel\_DataQAPractices\_Memo.pdf](http://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_DataQAPractices_Memo.pdf).

3. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). EGEMS (Washington, DC). 2018;6(1):3-.

4. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Washington, DC). 2016;4(1):1244.

1. <https://www.sentinelinitiative.org/> [↑](#footnote-ref-0)