DataQuality: DataQualityDashboard: Thresholds Network Study

Authors: Vojtech Huser

**Date:** 10 Nov 2017

**Acknowledgement:** The analysis is based in part on work from the Observational Health Sciences and Informatics collaborative. OHDSI (<http://ohdsi.org>) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

The authors declare no conflict of interest.

Contents

[Amendments and Updates 2](#_Toc30750785)

[Background 2](#_Toc30750786)

[Objective 2](#_Toc30750787)

[Data sources 2](#_Toc30750788)

[D1 - HIV Cohort 3](#_Toc30750789)

[D2 3](#_Toc30750790)

[D3 - Stanford University 3](#_Toc30750791)

[Data collection 3](#_Toc30750792)

[Population 3](#_Toc30750793)

[Methods 3](#_Toc30750794)

[Table and figure shells 3](#_Toc30750795)

[References 12](#_Toc30750796)

# Amendments and Updates

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author(s) | Comments |
| 2 | 23 Jan 2020 | Vojtech Huser | Addition of more percentiles (no dependency on Achilles) |
| 1 | 10 Nov 2019 | Vojtech Huser | First version, based on Achilles |

# Background

There is lack of knowledge bases facilitating data quality assessment (DQA) of laboratory data. Previous related work look at units[1]. In order to facilitate machine learning and AI methods over observational databases, ensuring high quality input data is crucial.

# Objective

We aim to study how to optimize data quality assessment. One area of focus of laboratory data. Numerical values as results of lab tests may contain extreme values that are implausible. We plan to design a method to arrive at rule based assessment of data quality for laboratory data. (primary analysis).

# Data sources

The analyses will be performed across a network of observational healthcare databases. All databases have been transformed into the OMOP Common Data Model. The complete specification for OMOP Common Data Model is available at: <http://omop.org/cdm>.

The following datasets will be included in this analysis:

* HIV cohort dataset (D1)
* Dataset 2 (J&J) (D2)
* Stanford U (D3

## D1 - HIV Cohort

HIV Cohort dataset is data representing cohort of patients sourced from Great Plains Collaborative.

## D2 - Dataset 2

## D3 - Stanford University

# Data collection

Data collection will be performed through a shared analysis program combining R and SQL. This program will be distributed to the data partners, executed locally against the data in OMOP CDM format, and results will be returned to the central coordinating site (NIH).

# Population

Included in the analysis will be all persons in a given dataset with at least one laboratory resuls with units and non-null numerical value.

# Methods

We first consider lab test-unit pairs and characterize numerical values for such pair. For pairs where multiple datasets provide data, we arrive at thresholds values suitable for data quality assessment (DQA).

.

# Table and figure shells

In this section the tables and figures that will be generated are described, and examples *showing fake/random data* are shown.

Table 1: Thresholds by lab test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Min Extreme | Min Plausible | Max Plausible | Max Extreme |
| Lab test 1-unit A |  |  |  |  |
| Lab test 1- unit B |  |  |  |  |
| Lab test 2 – unit A |  |  |  |  |
| … |  |  |  |  |

# References

1 Huser V. Real Word Database for Validation of Units for Clinical Laboratory Tests. In: *Proc AMIA Symp*. 2018. https://github.com/vojtechhuser/ThemisConcepts (accessed 10 Dec 2019).