Name:	Vojtech Huser
Affiliation:	National Institutes of Health
Email:	vojtech.huser@nih.gov
Presentation type (s):	Poster

Facilitating analysis of measurements data though stricter model conventions: Exploring units variability across sites

Vojtech Huser, MD, PhD[1]

[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract

Strict data model specifications lead to data that are easier to analyze. Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) currently does not mandate specific units for laboratory measurements. We explored variability of measurement units by laboratory test in several OMOP datasets and assessed feasibility of producing stricter conventions for units in measurement data.

Introduction

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has emerged as one of the leading models for capturing healthcare data. Recent evolution of the OMOP CDM focused on better modelling of Electronic Health Record, in addition to claims data. The measurement table within the OMOP CDM allows capture of laboratory results structured by measurement concept id and contains model components (table columns) for result as coded value or numerical value (with units captured as structured concept). Our work focuses on possibly improving how laboratory measurement data are standardized for analysis. OMOP CDM specifications consist of syntax standardization (table format specification) and conventions that further restrict valid OMOP dataset.[1] OHDSI Achilles Heel tool tests for conformance to a small subset of those specified conventions but further extensions of the Achilles Heel rule knowledge base are needed to cover all existing conventions.[2] Some research networks, such as PEDSnet (a Clinical Data Research Network funded by Patient-Centered Outcomes Research Institute), chose to implement an even stricter and longer set of conventions.[3] Tight model specification in other areas of CDM, such as the requirement of only standard concepts in data (depending on the data domain) allowed analyses that can be executed without site-specific analytical code customization. This is because the data standardization is performed by sites during the Extract-Transform-Load (ETL) process. This greater upfront effort during ELT facilitates less complex statistical code later during data analysis.

We considered a scenario where an analysis needs to process a given laboratory results (e.g., LDL cholesterol or weight) and either units are standardized to a single unit per measurement (e.g., kg for weight or mg/dL for LDL cholesterol) or can appear in multiple units (e.g., pounds or mmol/L). We explored site variability in measurement units and ways to advance the measurement data CDM data restricting conventions or Achilles Heel rules that would facilitate greater data standardization across sites.

Methods

Using selected Observational Health Data Sciences and Informatics (OHDSI) data partners from our ongoing Data Quality study[4], we created an additional limited data extraction (using Achilles pre-computations; no patient-level data were needed) that extracts units for a subset of highly prevalent measurements. The extraction code and sample

extracted data for a dataset are available at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/DataQuality/extras/units. Each participating site can customize two extraction thresholds to decrease the number of items reported (threshold 1: minimum number of tests instances for the test to be included in the extract; threshold 2: minimum ratio for a unit (within a test) for the unit to be included in the extract). Rare units and rare tests were thus excluded to minimize the extract size and possibly increase participation of more sites.

Preliminary Results and Discussion

We extracted data for three CDM datasets representing different organizational settings. Table 1 demonstrates few examples of measurements with two or more units (with ratio >0.2). To facilitate conversion, we tested two UCUM-based web services providers if automated conversion for all measurements would be feasible. We also performed validation of all UCUM units (CONCEPT_CODE) within the OMOP Vocabulary using one validator and found 120 units (out of 973 defined units) that are not valid.

Comparison of units by site shows that using empirical data, it would be possible to generate a list of permissible units by measurement_concept_id to reduce data variation or to support analysts in conducting the conversion at analysis time. To facilitate unit standardization, we have extended Achilles Heel rule set with a new rule that generates a warning if data for a measurement test indicate presence of more than 5 distinct units (included in Achilles version 1.4.0). We have initiated forum discussion to standardize several common measurements (inspired by data quality checks of the Sentinel network).[5] Besides units, currently, for number of measurements, there are two valid standard concepts for several vital signs (LOINC code and SNOMEDCT code). Our study is limited by only including a limited set of OMOP datasets and by working with only a subset of units.

Table 1: Example of units found for a subset of measurements (units with ratio <0.2 are not listed)

Measurement Concept ID	Measurement Name	Ratio	Unit Concept ID	Unit Name
3000034	Microalbumin urine	0.44	8859	microgram per milliliter
3000034	Microalbumin urine	0.28	8840	milligram per deciliter
3000819	Albumin/Creatinine [Mass Ratio] in 24 hour Urine	0.73	9072	microgram per milligram of creatinine
3000819	Albumin/Creatinine [Mass Ratio] in 24 hour Urine	0.25	9017	milligram per gram of creatinine
3007220	Creatine kinase [Enzymatic activity/volume] in Serum or Plasma	0.43	8645	unit per liter
3007220	Creatine kinase [Enzymatic activity/volume] in Serum or Plasma	0.24	8510	unit
44816571	Ethanol [Presence] in Saliva (oral fluid) by Confirmatory method	0.71	8840	milligram per deciliter
44816571	Ethanol [Presence] in Saliva (oral fluid) by Confirmatory method	0.29	8713	gram per deciliter

Acknowledgement: We would like to thank collaborating analysts at selected Data Quality study sites (Juan Banda, Seng Chan You [final analysis is pending for the site])

References

1. OHDSI. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM): Measurement table. https://github.com/OHDSI/CommonDataModel/wiki/MEASUREMENT#conventions. Accessed Aug 25, 2017.
2. OHDSI. Ahilles Heel Rule Knowledge Base Overview. https://github.com/OHDSI/Achilles/blob/master/inst/csv/achilles_rule.csv.
3. PEDSnet. Data Transformation conventions (for OMOP CDM). 2017; https://drive.google.com/drive/folders/0By1tgphRY1wpN1ExVkx6ZmlxdFE?usp=sharing.
4. Huser V. Data Quality Study Protocol. 2016; http://www.ohdsi.org/web/wiki/doku.php?id=research:dqstudy. Accessed Aug 2, 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.