Enabling Transparency Across a Global Data Network

Clair Blacketer[1,9], Mui VanZandt[2], Jose Posada[3], Nigam Shah[3], Seojeong Shin[4], Gyeol Song[5,6], Yunpeng Li[7], Mornin Feng[8], Marcel De Wilde[9], Peter Rijnbeek[9]

[1] ***Janssen Research & Development, Raritan, NJ, [2]IQVIA, USA, [3]Stanford School of Medicine, Stanford, CA, [4]Ajou University Graduate School of Medicine, Suwon, Republic of Korea, [5]EvidNet, Inc., Seongnam-si, Gyeonggi-do, Korea, [6]Department of Clinical Research Design & Evaluation, SAIHST, Sungkyunkwan University, Seoul, Korea, [7]SmindU, China, [8]Saw Swee Hock School of Public Health, Singapore, [9]Erasmus Medical Center, Rotterdam, NL***

Introduction

Large data networks now allow for observational research to be conducted on a global scale(1). To support this research, it is of the utmost importance that the contents of those data are assessed and communicated in an efficient way. The Data Quality Dashboard (DQD)[1], developed by the OHDSI Community, provides a comprehensive view of a database which can be used to determine its fitness-for-use in relation to clinical questions of interest.

Data Quality Dashboard

The Data Quality Dashboard was developed to leverage the structure of the OMOP Common Data Model(2) to enable systematic data quality assessment and to communicate the results of the assessment. It functions by using an underlying control file that applies 20 parameterized data quality checks across the data model, writing and resolving up to a total of 3,351 individual checks. These checks are organized into categories and contexts based on the Kahn data quality framework(3) and compared against thresholds set a priori to determine whether they pass or fail.

Methods

The DQD was run on a total of 39 databases representing data from 3 different regions (US, Asia-Pacific, EU) across 8 sites, all members of the OHDSI community. The results were initially compared based on the total percentage of checks that passed. 5 of the 3,351 checks were then chosen to represent the Kahn categories of plausibility, conformance, and completeness as a proof-of-concept for how the DQD results could inform large-scale network research. The checks chosen were:

1. The percent of records in the DRUG_EXPOSURE table with a NULL value in the DRUG_EXPOSURE_END_DATE field. [Conformance]
2. For the condition “Inflammatory disorder of male genital organ”, the percent of records associated with female patients. [Plausibility]
3. The percent of persons in the database that do not have at least one record in the MEASUREMENT table. [Completeness]
4. The percent of persons with a non-standard gender concept [Conformance]
5. For measurement records of cholesterol in serum with a unit of milligram per deciliter, the percent with a value greater than 266. [Plausibility]

Results

Across the 39 databases on average 97% of data quality checks passed. Only three of the 39 databases had any records with a NULL in the DRUG_EXPOSURE_END_DATE field of the DRUG_EXPOSURE table (Check 1). As seen in Figure 1 the distribution of the percent of “Inflammatory disorder of male

[1] https://github.com/OHDSI/DataQualityDashboard

genital organ" condition records associated with female patients (Check 2) varied widely by region, with the highest in US at 14.2% and the lowest in Asia-Pacific 0.33%. Check 3, the percent of persons that do not have at least one record in the MEASUREMENT table differed by database, with the lowest at 4.4% and the highest at 82.8%.

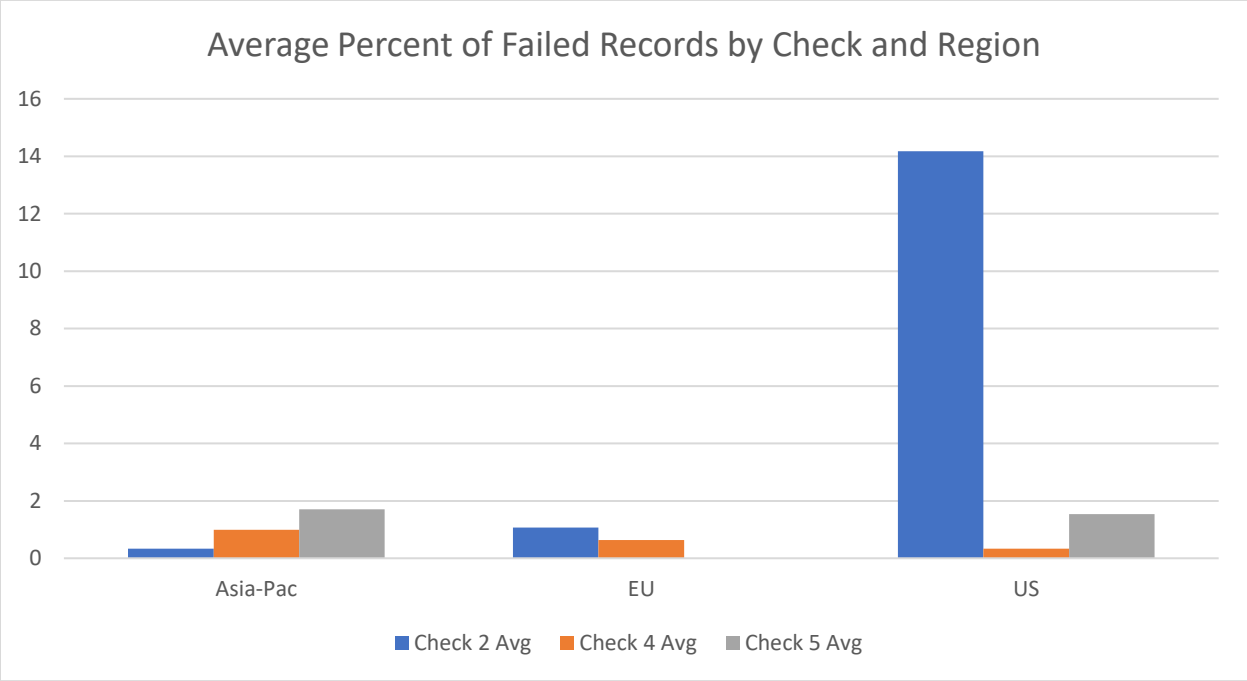

Figure 1. Average percent of failed records by data quality check and region

Both Check 4 and Check 5 were similar across the regions ranging from 0.33% to 0.99% and from 0% to 1.53%, respectively.

Discussion

The DQD gives an unprecedented look into a database, especially when considering a network study. The over 3,000 data quality checks can help understand if a database is fit to answer certain clinical questions. For example, if a study wants to look at the length of drug exposures, the DQD exposed that at least three databases of the 39 sampled have some records missing the DRUG_EXPOSURE_END_DATE, which will inhibit the calculation of exposure time. Some checks vary by region, as seen with Check 2 and some are database dependent, as seen with Check 3. Overall, the Data Quality Dashboard provides a comprehensive, transparent view of a database that informs network research and enables more efficient and informed collaboration.

References

1. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(27):7329-36.
2. OMOP Common Data Model: OHDSI; [02/01/2020]. Available from: www.github.com/OHDSI/CommonDataModel.
3. 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.