Comparing Data Quality Dashboard results from two ETL iterations: three new utilities

PRESENTER: **Anne van Winzum**

Background

The Data Quality Dashboard (DQD) has been widely used to **evaluate the quality of an OMOP CDM** data set resulting from an ETL (extract, transform, load) process[1]. In practice, during the conversion to the OMOP CDM we perform several ETL iterations. However, interpreting the differences in quality is not always straightforward.

We developed three new utilities as part of mapping of the **UK Biobank** (UKB) data under the European Health Data Evidence Network (EHDEN) COVID19 rapid data partner call[2], and in collaboration with University College London.

Methods

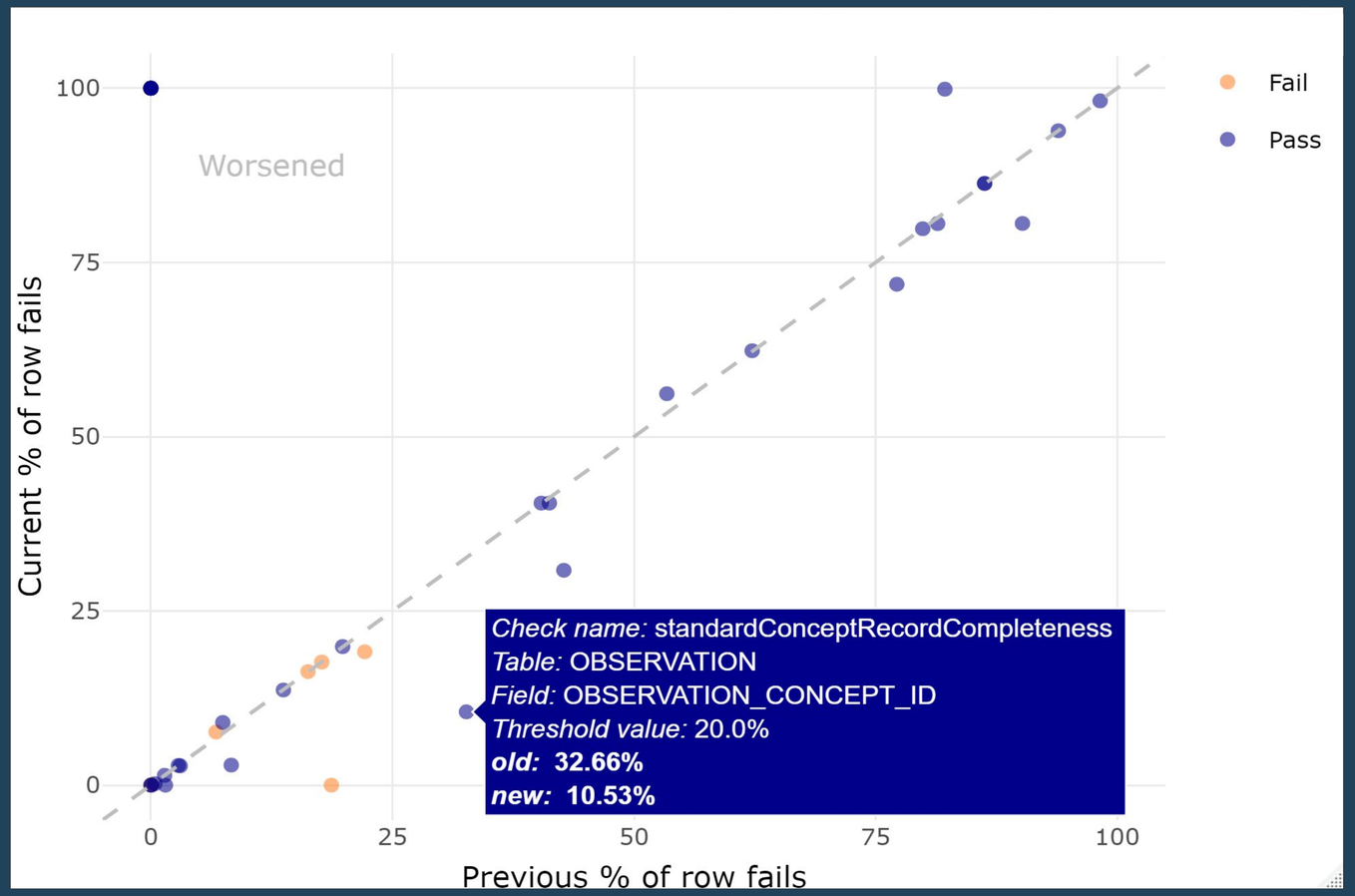
Thresholds editing

As part of the ETL iterations, we needed to change the fail-thresholds of individual checks. We created a separate table to **list the changed thresholds in a user-friendly way**. Our utility script takes this new table to produce a customized thresholds file accepted by the default DQD scripts.

Comparison of DQD results

This visualization script selects the checks for which the percentage of records that satisfy said check has changed between ETL iterations. Here, the percentage of records satisfying the checks had modestly improved (Figure 1). As an example, there is an outlier (top left corner) that prompted us to investigate and update the ETL accordingly.

Visualising and comparing DQD results is an important step to interpret the data quality and to find actionable data quality issues.



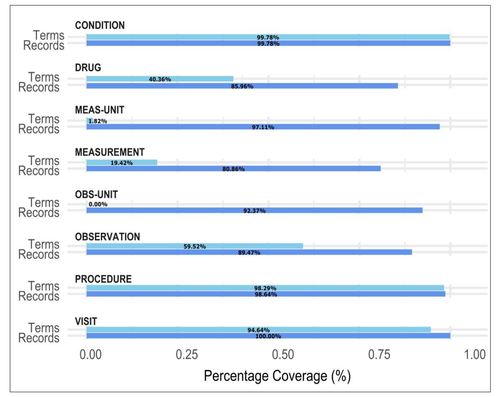
▲ *Figure 1: Each dot represents one check that has a different percentage of row fails between the iterations. This percentage of row fails is marked in the x-axis for the earliest run, and in the y-axis for the latest run.*

Scan QR to see our blogpost

On the other hand, we improved the standard record completeness in the observation table to be above 80%. Both visualizations are produced directly from DQD output.

Coverage per domain

An important part of the conversion quality is the concept mapping coverage. It is hard to get this overview from the DQD result tables alone. The new bar plot shows the **concept mapping coverage across all OMOP domains**. This ETL iteration achieved a high coverage throughout all domains and units (Figure 2) in terms of records mapped to standard concepts. The number of unique terms mapped was low for measurement and observation units (1.82% and 1.00%) and for measurement (19.42%).



▲ *Figure 2: Barplot for the mapping coverage in an ETL. In light blue: the percentage of distinct terms mapped to a standard OMOP concept; in darker blue: the percentage of records mapped to a standard OMOP concept.*

References

[1] Blacketer C, Defalco F, Ryan P, Rijnbeek P. Increasing trust in real-world evidence through evaluation of observational data quality. medRxiv 2021.

[2] EHDEN COVID19 Rapid Data Partner Call, April 2020, retrieved May 2021.https://www.ehden.eu/open-calls/04-2020-covid19-data-partner-call/

Elena G. Lara, Maxim Moinat, Anne van Winzum

The Hyve OHDSI EHDEN EUROPEAN HEALTH DATA & EVIDENCE NETWORK