Introduction to the OHDSI DataQualityDashboard

Katy Sadowski
Real World Evidence Analytics, Boehringer Ingelheim
What is data quality?

The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.

- The Book of OHDSI
What is data quality?

Inherent quality of the original source data

Quality of the OMOP CDM
• R package
• Part of the HADES suite
• 26 data quality check types --> over 4,000 individual data quality checks
• Results exported to json, csv, or database table, and can be explored in an R Shiny application
# MEDICARE CLAIMS SYNTHETIC PUBLIC USE FILES (SYNPUFS)

**DataQualityDashboard Version:** 1.0.0  
Results generated at 2021-11-25 05:41:37 in 14 hours

<table>
<thead>
<tr>
<th>STATUS</th>
<th>TABLE</th>
<th>CATEGORY</th>
<th>SUBCATEGORY</th>
<th>LEVEL</th>
<th>NOTES</th>
<th>DESCRIPTION</th>
<th>% RECORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAIL</td>
<td>MEASUREMENT</td>
<td>Completeness</td>
<td>None</td>
<td>TABLE</td>
<td>None</td>
<td>The number and percent of persons in the CDM that do not have at least one record in the MEASUREMENT table (Threshold=95%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>FAIL</td>
<td>VISIT_OCCURRENCE</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>None</td>
<td>The number and percent of records with a value of 0 in the standard concept field VISIT_CONCEPT_ID in the VISIT_OCCURRENCE table. (Threshold=0%).</td>
<td>84.86%</td>
</tr>
<tr>
<td>FAIL</td>
<td>PROCEDURE_OCCURRENCE</td>
<td>Plausibility</td>
<td>Atemporal</td>
<td>CONCEPT</td>
<td>None</td>
<td>For a CONCEPT_ID 2721063 (ANNUAL GYNECOLOGICAL EXAMINATION, NEW PATIENT), the number and percent of records associated with patients with an implausible gender (correct gender $=$ FEMALE). (Threshold=5%).</td>
<td>75.00%</td>
</tr>
</tbody>
</table>
An aggregated summary statistic that can be computed against a dataset and to which a decision threshold can be applied to determine if the statistic meets expectations.
Data quality check example

The number and percent of records with a value in the **DAYS_SUPPLY** field of the **DRUG_EXPOSURE** table less than **0**.
Data quality check example

The number and percent of records with a value in the CDM field of the CDM table less than a low value.
Data quality check example

The number and percent of records which are not mapped into a standard concept in the `CONDITION_CONCEPT_ID` field of the `CONDITION_OCCURRENCE` table.
Data quality check example

The number and percent of records which are not mapped into a standard concept in the CDM field of the CDM table.
# Data quality check types

<table>
<thead>
<tr>
<th>Check Type</th>
<th>Check Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person Completeness</td>
<td>The number and percent of persons in a database that do not have a least one record in the <strong>CDM table</strong>.</td>
</tr>
<tr>
<td>Is Required</td>
<td>The number and percent of records with a NULL value in a <strong>CDM field</strong> of a <strong>CDM table</strong> that is considered not nullable.</td>
</tr>
<tr>
<td>Is Foreign Key</td>
<td>The number and percent of records that have a value in a foreign key <strong>CDM field</strong> of a <strong>CDM table</strong> that does not exist in the <strong>foreign key table</strong>.</td>
</tr>
<tr>
<td>Is Standard Valid Concept</td>
<td>The number and percent of records that do not have a standard, valid concept in the <strong>CDM field</strong> <strong>CDM table</strong>.</td>
</tr>
<tr>
<td>Plausible Temporal After</td>
<td>The number and percent of records with a value in a <strong>CDM field</strong> of a <strong>CDM table</strong> that occurs prior to a <strong>plausible date</strong>.</td>
</tr>
<tr>
<td>Plausible Value Low</td>
<td>For a given <strong>CONCEPT_ID</strong> and <strong>UNIT_CONCEPT_ID</strong> pair, the number and percent of records with a value lower than the <strong>plausible low value</strong>.</td>
</tr>
<tr>
<td>Plausible Gender</td>
<td>For a given <strong>CONCEPT_ID</strong>, the number and percent of records associated with persons with an <strong>implausible gender</strong>.</td>
</tr>
</tbody>
</table>
An aggregated summary statistic that can be computed against a dataset and to which a decision threshold can be applied to determine if the statistic meets expectations.
Data quality check thresholds

- Fatal checks: Threshold is always 0 🕷️
- CDM convention checks: Threshold should be 0 in theory, but may need to be adjusted in practice
- Characterization checks: Threshold depends on expectations of the data source*

*and the analytic use case
Getting Started

1. Prerequisites
   - An OMOP CDM
   - An R environment set up following HADES instructions: [https://ohdsi.github.io/Hades/rSetup.html](https://ohdsi.github.io/Hades/rSetup.html)

   - By default, DQD will apply preset data quality check thresholds which may or may not be relevant for your OMOP CDM

What's New

Check-level documentation

New & improved checks

• Temporal plausibility
• Unit concept checks
• Gender plausibility

Coming soon...

• Check severity parameter
• New user interface

isRequired

Katy Sadowski

2024-02-21

Source: vignettes/checks/isRequired.Rmd

Summary

Level: Field check
Context: Validation
Category: Conformance
Subcategory: Relational
Severity: Fatal

Description

The number and percent of records with a NULL value in the cdmFieldName of the cdmTableName that is considered not nullable

Definition

This check is meant to ensure that all NOT NULL constraints specified in the CDM version are followed.

• Numerator: The number of rows with a NULL value in the column
• Denominator: The total number of rows in the table
• Calculated (CDM Constraint) × "Survived" rows in CDM table score
Thank You – 2024 Contributors

Clair Blacketer, Dima Dymshyts, Jared Houghtaling, Maxim Moinat, Nitesh Balakrishnan