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.

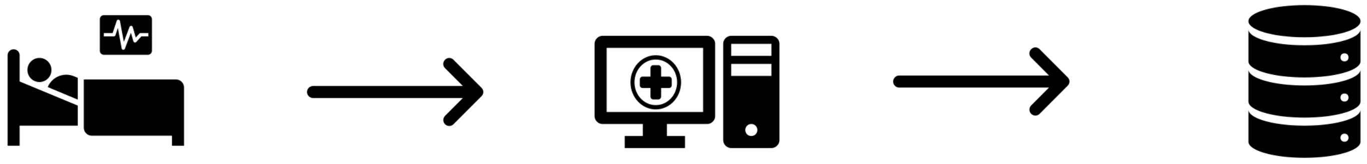
- *The Book of OHDSI*

Roebuck, Kevin. 2012. *Data Quality: High-Impact Strategies-What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*. Emereo Publishing.

What is data quality?

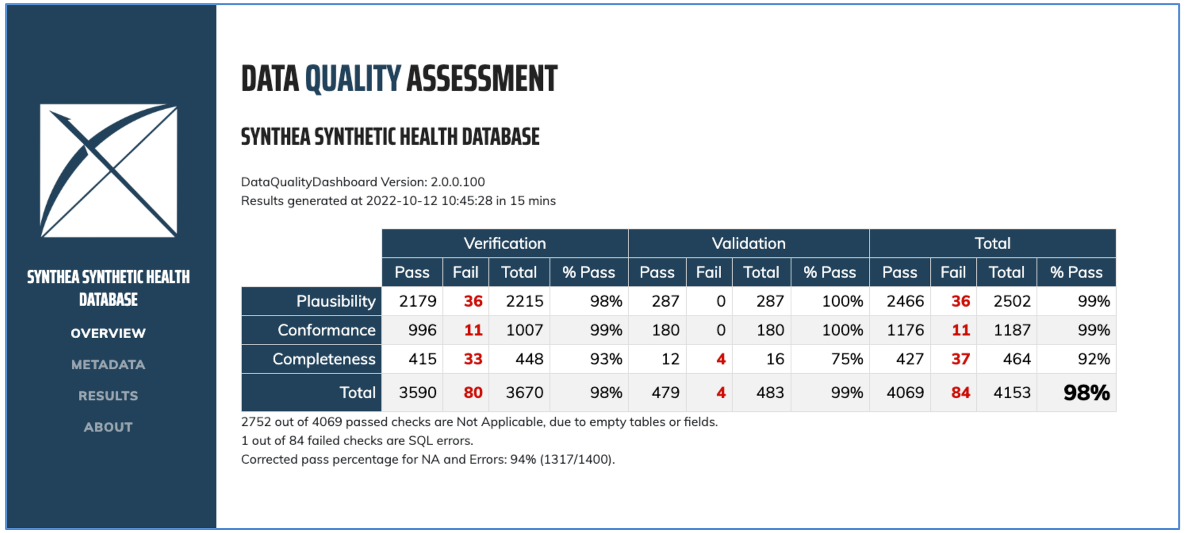
Inherent quality of the **original source data**

Quality of the **OMOP CDM**



DataQualityDashboard

- 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



DATA QUALITY ASSESSMENT

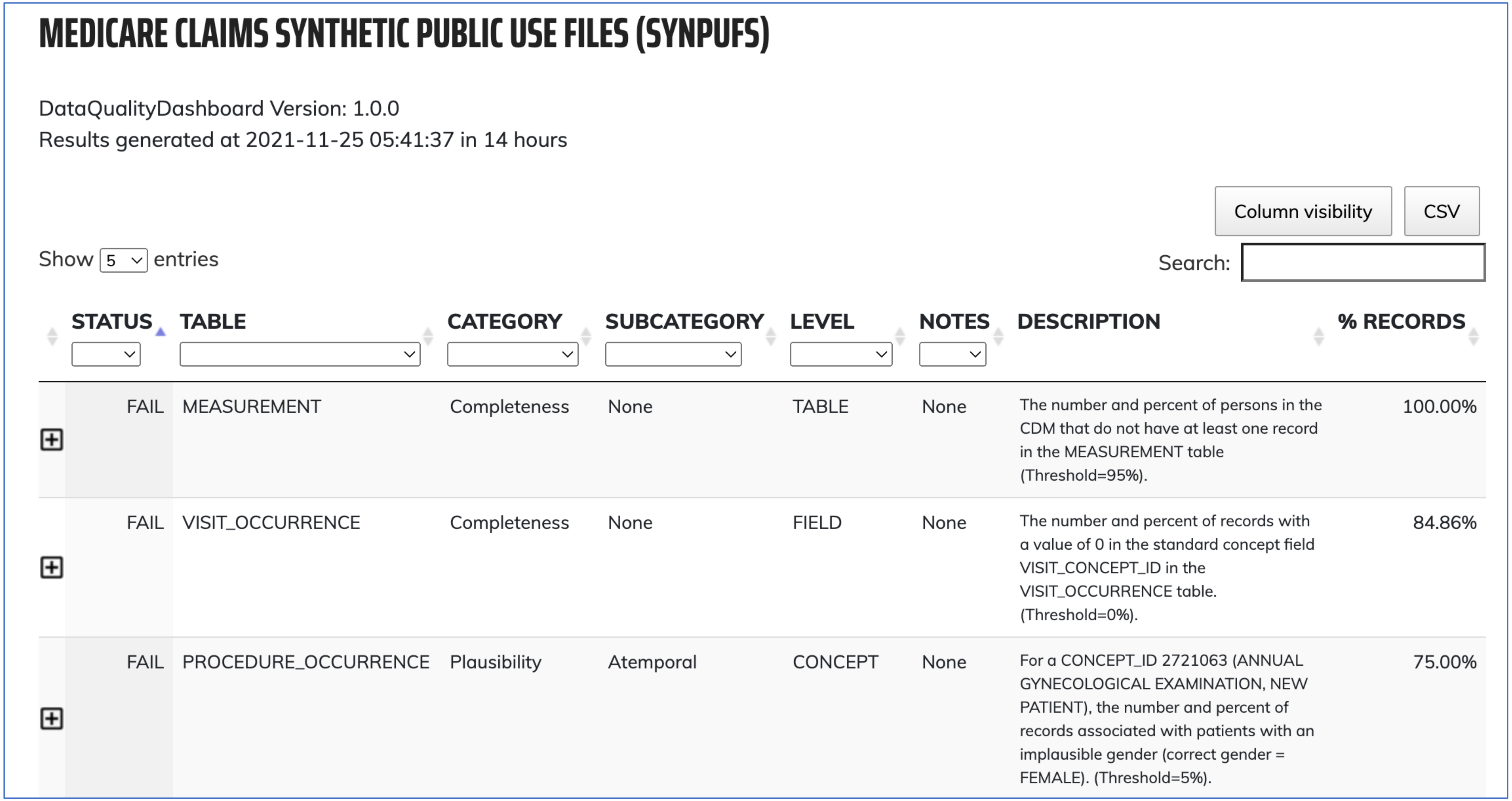
SYNTHEA SYNTHETIC HEALTH DATABASE

DataQualityDashboard Version: 2.0.0.100
Results generated at 2022-10-12 10:45:28 in 15 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2179	**36**	2215	98%	287	0	287	100%	2466	**36**	2502	99%
Conformance	996	**11**	1007	99%	180	0	180	100%	1176	**11**	1187	99%
Completeness	415	**33**	448	93%	12	**4**	16	75%	427	**37**	464	92%
Total	3590	**80**	3670	98%	479	**4**	483	99%	4069	**84**	4153	**98%**

2752 out of 4069 passed checks are Not Applicable, due to empty tables or fields.
1 out of 84 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 94% (1317/1400).

DataQualityDashboard



MEDICARE CLAIMS SYNTHETIC PUBLIC USE FILES (SYNPUFS)

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

Show 5 entries

	STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
⊞	FAIL	MEASUREMENT	Completeness	None	TABLE	None	The number and percent of persons in the CDM that do not have at least one record in the MEASUREMENT table (Threshold=95%).	100.00%
⊞	FAIL	VISIT_OCCURRENCE	Completeness	None	FIELD	None	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%).	84.86%
⊞	FAIL	PROCEDURE_OCCURRENCE	Plausibility	Atemporal	CONCEPT	None	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%).	75.00%

Data quality check definition

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.

Blacketer, Clair and Williams, Andrew. 2019. *Data Quality*. https://www.ohdsi.org/wp-content/uploads/2019/09/2-Plenary-1-OMOP-DQ-Clair-Andrew.pdf

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.

Blacketer, Clair and Williams, Andrew. 2019. *Data Quality*. https://www.ohdsi.org/wp-content/uploads/2019/09/2-Plenary-1-OMOP-DQ-Clair-Andrew.pdf

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.

Blacketer, Clair and Williams, Andrew. 2019. *Data Quality*. https://www.ohdsi.org/wp-content/uploads/2019/09/2-Plenary-1-OMOP-DQ-Clair-Andrew.pdf

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.

Blacketer, Clair and Williams, Andrew. 2019. *Data Quality*. https://www.ohdsi.org/wp-content/uploads/2019/09/2-Plenary-1-OMOP-DQ-Clair-Andrew.pdf

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.

Blacketer, Clair and Williams, Andrew. 2019. *Data Quality*. https://www.ohdsi.org/wp-content/uploads/2019/09/2-Plenary-1-OMOP-DQ-Clair-Andrew.pdf

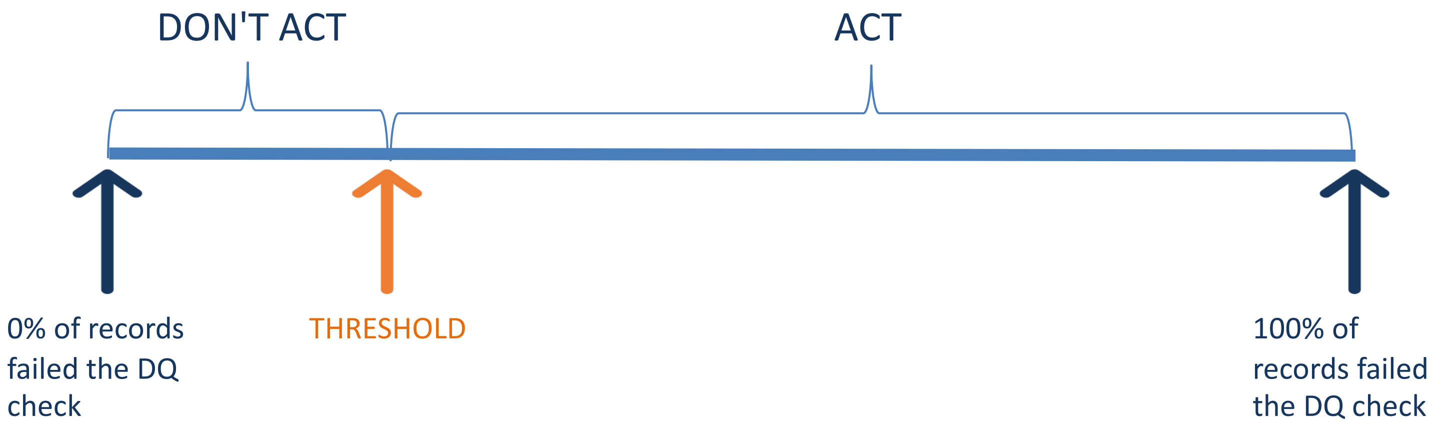
Data quality check types

Check Type	Check Description
Person Completeness	The number and percent of persons in a database that do not have a least one record in the *CDM table.*
Is Required	The number and percent of records with a NULL value in a *CDM field* of a *CDM table* that is considered not nullable.
Is Foreign Key	The number and percent of records that have a value in a foreign key *CDM field* of a *CDM table* that does not exist in the *foreign key table.*
Is Standard Valid Concept	The number and percent of records that do not have a standard, valid concept in the *CDM field* of a *CDM table.*
Plausible Temporal After	The number and percent of records with a value in a *CDM field* of a *CDM table* that occurs prior to a *plausible date.*
	…
Plausible Value Low	For a given *CONCEPT_ID* and *UNIT_CONCEPT_ID* pair, the number and percent of records with a value lower than the *plausible low value.*
Plausible Gender	For a given *CONCEPT_ID*, the number and percent of records associated with persons with an *implausible gender.*

Blacketer, Clair and Williams, Andrew. 2019. *Data Quality*. https://www.ohdsi.org/wp-content/uploads/2019/09/2-Plenary-1-OMOP-DQ-Clair-Andrew.pdf

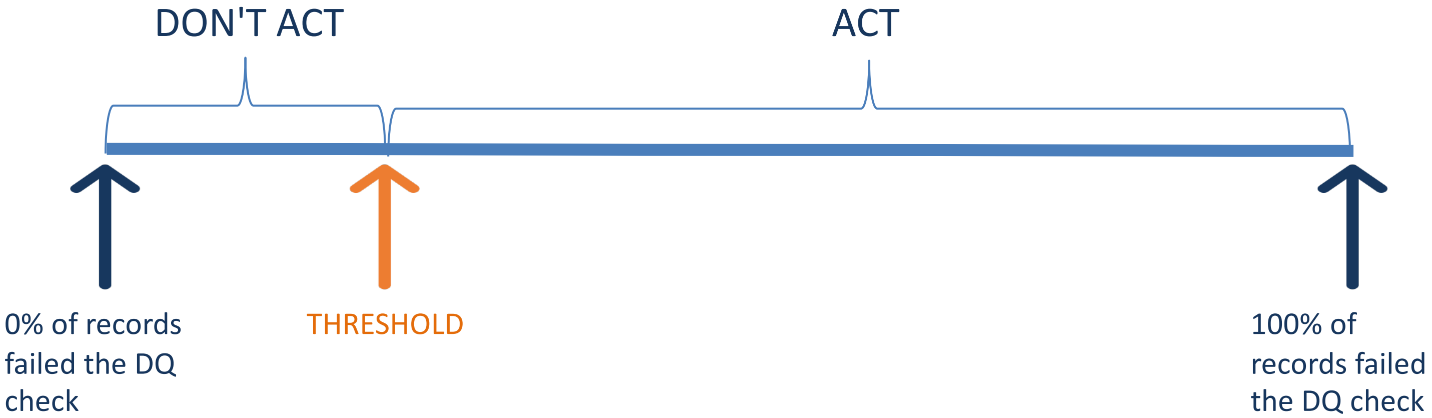
Data quality check thresholds

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**.


DON'T ACT
ACT
0% of records failed the DQ check
THRESHOLD
100% of records failed the DQ check

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*



DON'T ACT

ACT

0% of records failed the DQ check

THRESHOLD

100% of records failed the DQ check

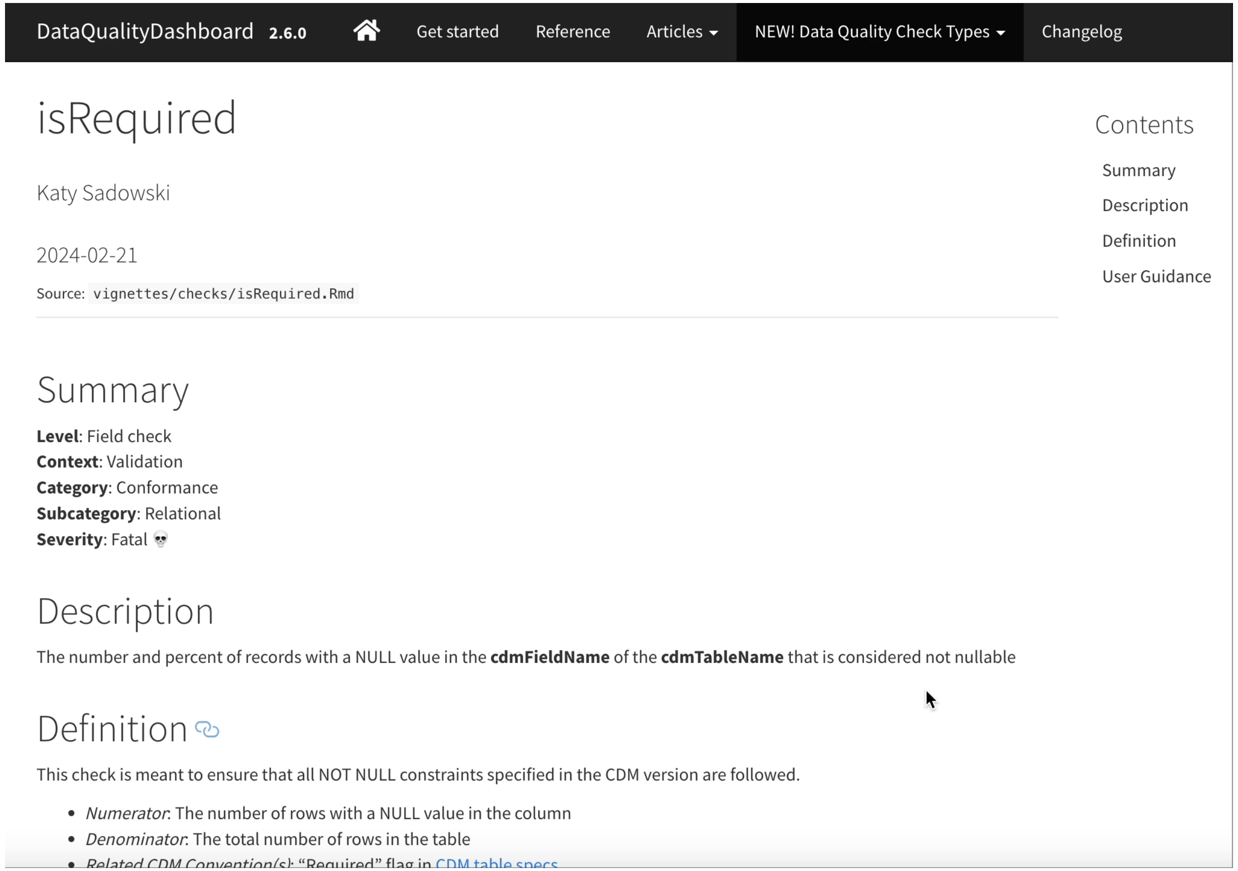
*and the analytic use case

Getting Started

1. Prerequisites
   - An OMOP CDM
   - An R environment set up following HADES instructions: https://ohdsi.github.io/Hades/rSetup.html
2. Install and execute DataQualityDashboard following DQD instructions: https://ohdsi.github.io/DataQualityDashboard/articles/DataQualityDashboard.html
   - By default, DQD will apply preset data quality check thresholds which may or may not be relevant for your OMOP CDM
3. Configure data quality check thresholds: https://ohdsi.github.io/DataQualityDashboard/articles/Thresholds.html

What's New

Check-level documentation



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
- *Related CDM Convention(s)*: “Required” flag in CDM table specs

New & improved checks

- Temporal plausibility
- Unit concept checks
- Gender plausibility

Coming soon...

- Check severity parameter
- New user interface 🤩

Thank You – 2024 Contributors

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