

Beyond Missingness: Systematizing Methods for Comprehensive Data Fitness Assessment in Clinical Research



OHDSI Community Call

June 30, 2026

Hanieh Razzaghi, PhD, MPH & Charles Bailey, MD, PhD

The Problem

Analysis relies on the day saying what you mean

The data say what they mean.

You can't change the data.

You can't see the data.

You can't recollect the data.

What to do?

Γνωθι δεδομενα σας

Listen to the data

Plan for the data



Study-Specific Data Quality Assessment

EDA

Curation

“Cleaning”

Network (Structural) DQA

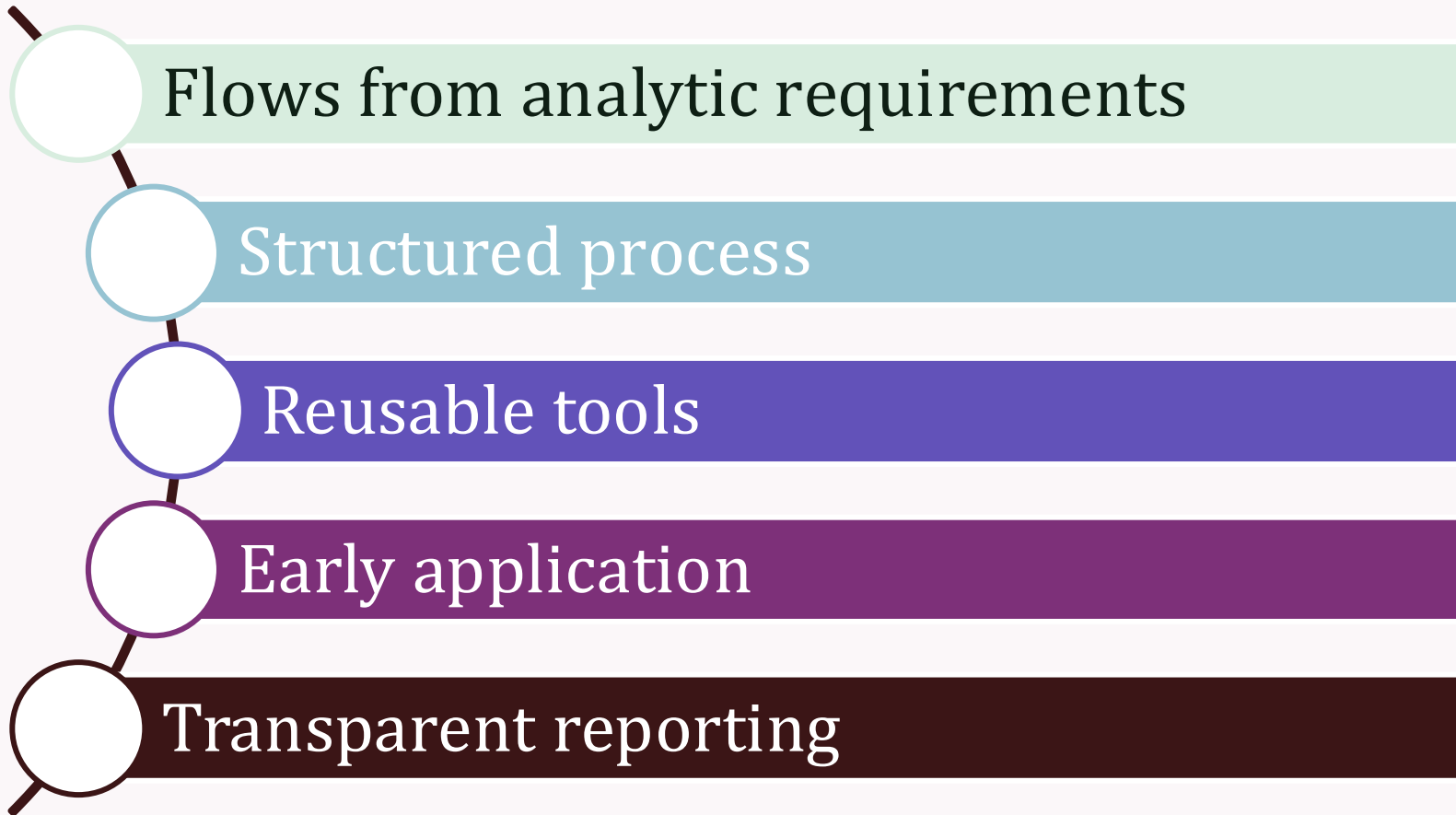
Tends to focus on structure

- Missingness
- Data model conformance
- Value sets or ranges

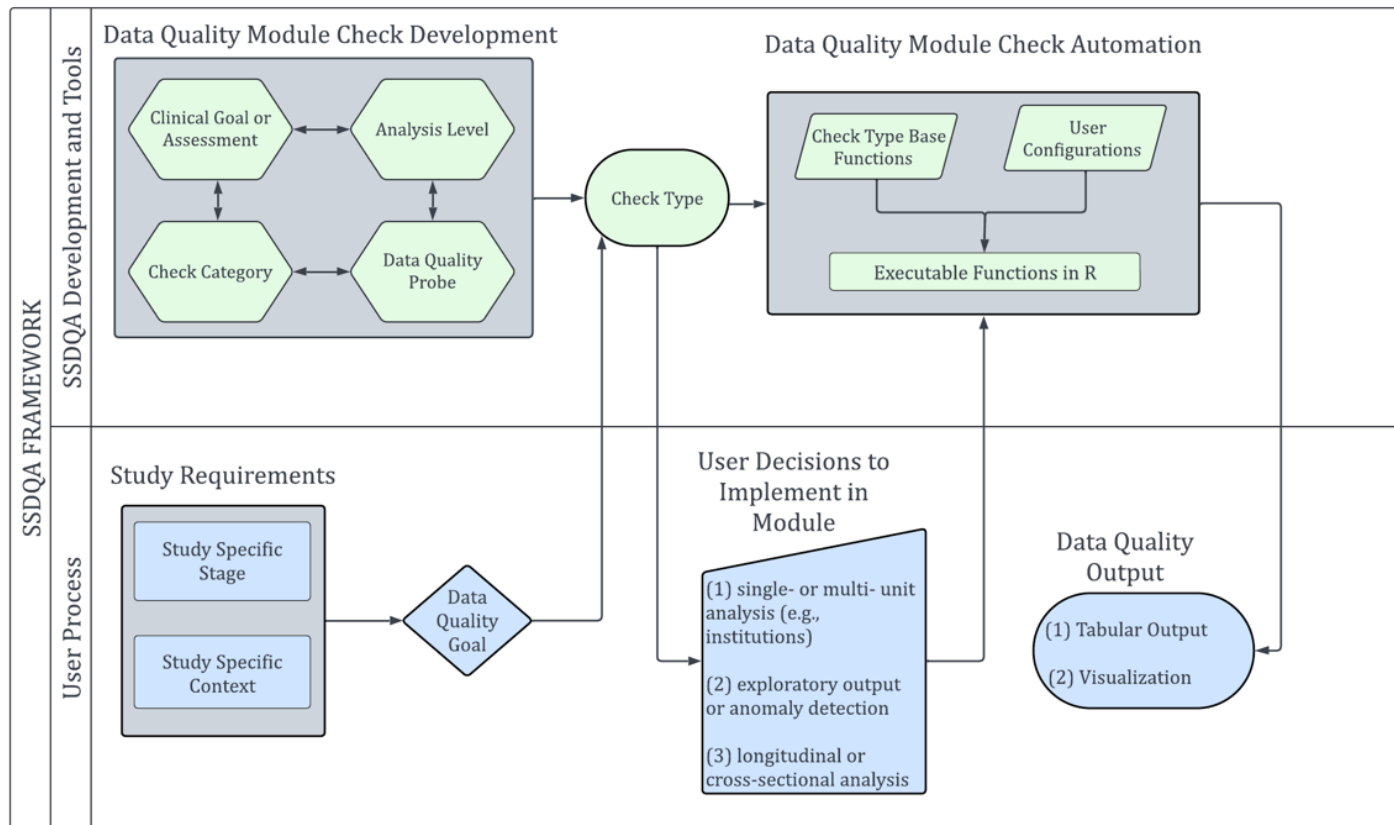
Plausibility focuses on never events

- Impossible dates
- Physiologic impossibilities

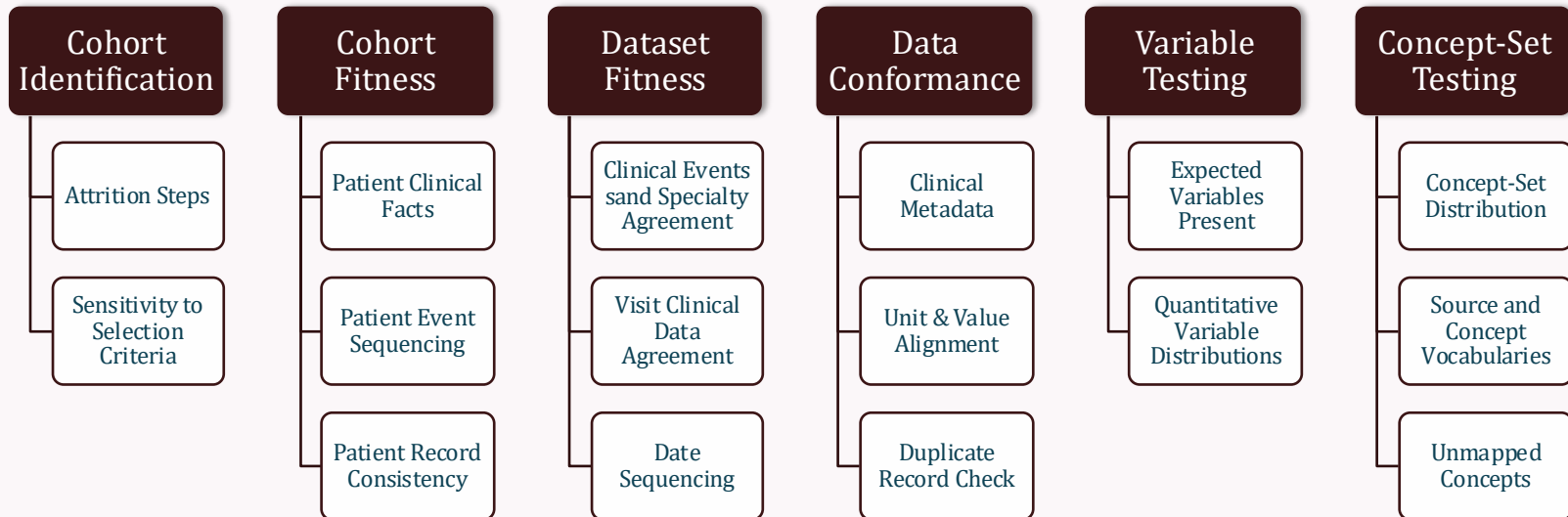
Study-Specific DQA



Conceptual Framework

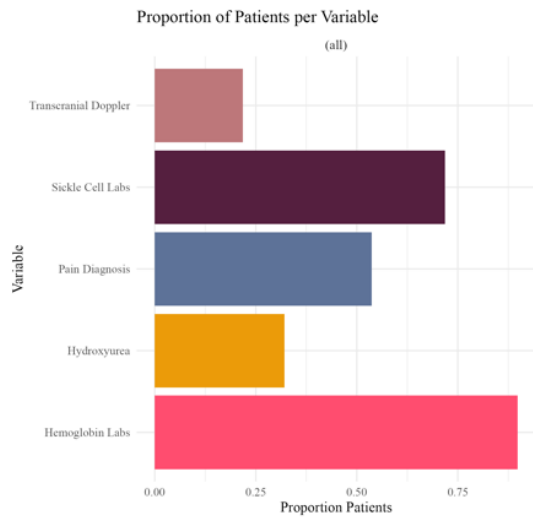


SQUBA: Study-Specific DQ Modules

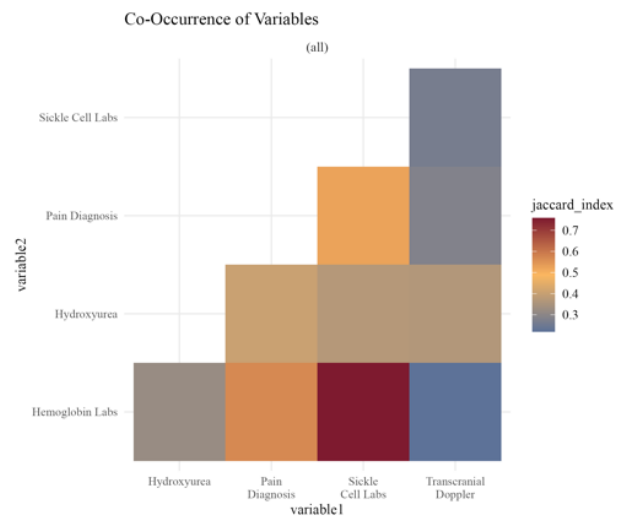


Applying SSDQA: EVP check

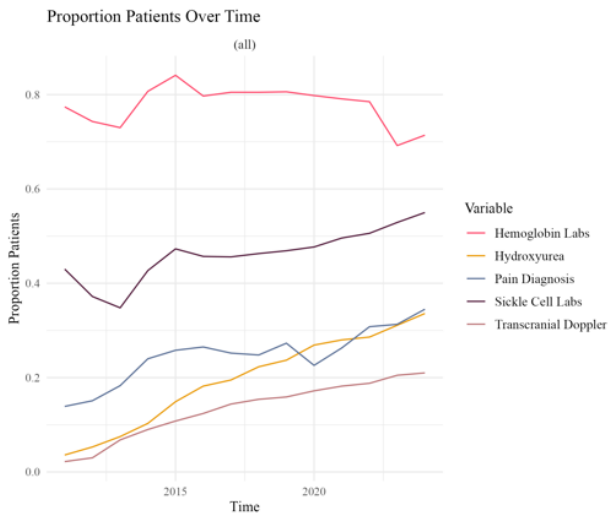
A



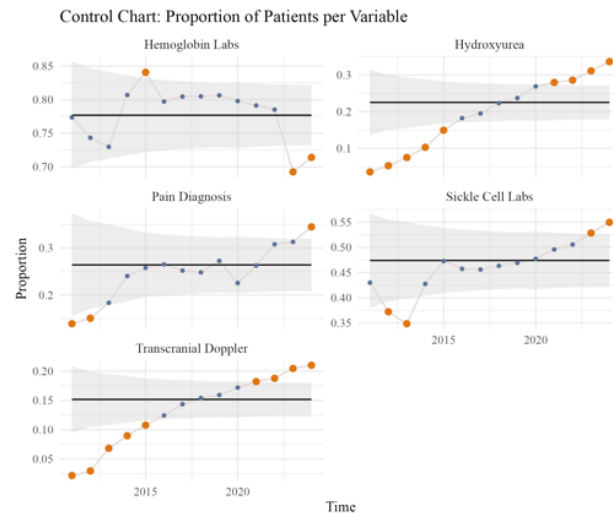
B



C

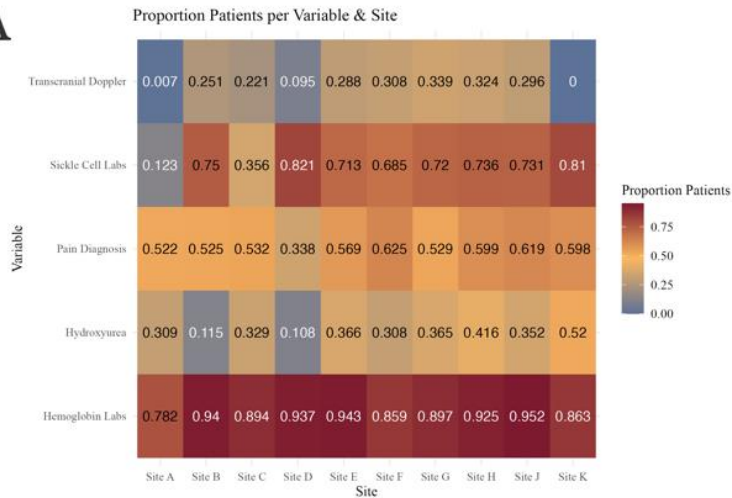


D

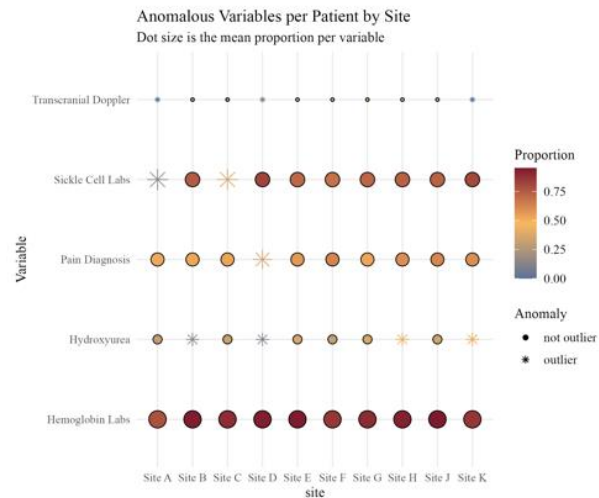


Stratifying Data

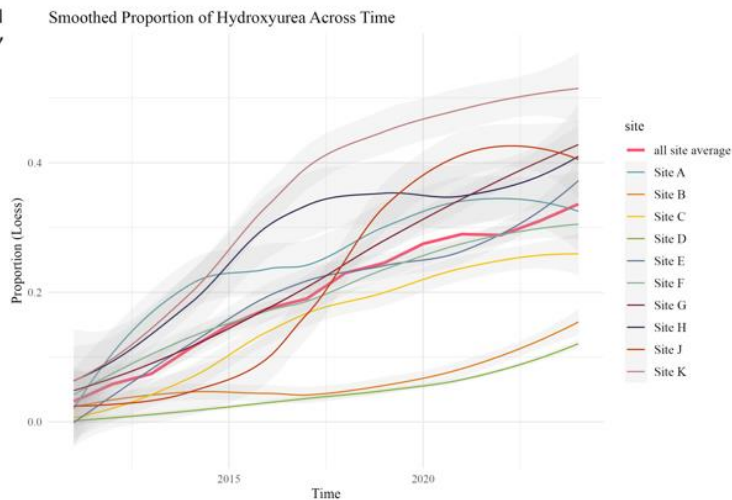
A



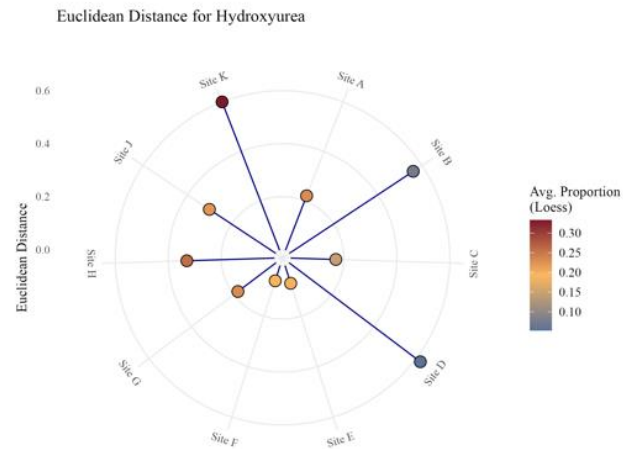
B



C



D



How does SQUBA compare to other tools?

Metric	SO	CD	DG	DOR	DOD	EC	EN	MIR	N3C	PC	SPF
Study cohorts	3	3	3	3	1	1	1	1	1	1	3
Theoretical framework	3	2	3	3	3	1	1	3	2	1	1
Check diversity	3	2	3	3	3	1	1	3	3	2	2
Check names/catalog	3	2	3	3	3	2	1	3	1	1	2
Extensibility	3	1	2	2	3	2	1	2	1	1	3
Customization	3	1	3	3	2	1	1	3	1	1	3
Data Models	3	2	0	1	2	3	2	3	3	2	1
Packaging	3	3	0	3	3	2	3	3	2	3	0
Availability	3	3	0	3	3	2	3	2	1	1	0
Dependencies	3	2	0	3	3	3	3	3	1	3	0
Coding required	3	3	0	3	3	3	3	3	3	3	0
Integration	2	3	0	2	3	3	3	3	2	3	0
Pass/Fail Checks	3	1	3	3	3	1	1	3	3	3	3
Tabular output	3	3	0	3	3	3	3	3	3	3	0
Visualizations	3	1	0	3	1	2	3	3	2	1	0
Temporal dimension	3	1	1	1	1	1	1	1	1	1	1

Labels: SQ – SQUBA; CD – OHDSI CohortDiagnostics; DG – DataGauge; DQR – dataquieR; DQD – OHDSI Data Quality Dashboard; EC – DQe-c v2; EN – ENACT Data Quality Explorer; MIR – MIRACUM data quality workbench (DQAstats/DQAgui) ; N3C- National COVID Cohort Collaborative data quality program; PC – PCORnet® Data Curation; SPF – SPFID2

Making Metadata Available



Our Network ▾ Research ▾ Youth & Families ▾ Learning Database ▾ Work with Us ▾ [Log In](#) ▾

[Collections & Domains](#) [All of PEDSpace](#) ▾ [Statistics](#)

[Home](#) ▾ [Data Quality Modules](#) ▾ [Cohort Identification](#) ▾ [Sensitivity to Selection Cr...](#) ▾ [Sensitivity to Selection Cr...](#)

[Data Quality Check](#)

Sensitivity to Selection Criteria: Multi-Site, Exploratory, Cross-Sectional Analysis



Created
[2024-12-17](#)



Click on the thumbnail above to preview images.

Files
2 Downloads ▾

Domain
[Cohort Identification](#)

Category
[Plausibility](#)

Parameters
[Multi-Site Analysis](#) [Exploratory Analysis](#) [Cross-Sectional Analysis](#)

Abstract

The Sensitivity to Selection Criteria module measures how different cohort inclusion definitions can impact the makeup of a cohort, including demographics, utilization, and outcomes. Each check will compare a user-provided base cohort definition to each provided alternate cohort definition to evaluate how the new definitions impact each measured characteristic.

Data Requirements

`base_cohort , alt_cohorts , multi_or_single_site , anomaly_or_exploratory , person_tbl , visit_tbl , provider_tbl , care_site_tbl , demographic_mappings , specialty_mappings , specialty_concepts , outcome_concepts , domain_tbl , domain_select`

Probe

[Misclassification Detection](#) [Eligibility Criteria Assessment](#) [External Benchmarking](#) [Selection Error or Bias Detection](#)

Clinical Assessment

[Targeted Patient Population](#) [Clinical Data Distributions](#)

Access Package

```
# install.packages("devtools")
devtools::install_github('ssdqa/sensitivityselectioncriteria')
```



Visualization Output

This check outputs a visualization including continuous variables (base median, first...

Thank you!

[SQUBA on PEDSpace](#)

Available in Github (SQUBA)

More descriptions with example output on PEDSnet metadata repository

ssdqa

Overview Repositories 14 Projects Packages People

SQUBA

Study-Specific Quality, Utility, and Breadth Assessment

README .md

Study-Specific Quality, Utility, and Breadth Assessment

See below for a table of the currently available modules. Each module has a distinct repository within the Github organization which can be accessed via the link in the "Module" column. The packages also pull general use functions from the [squba.gen](#) repository where needed.

Data Quality Checks

Quantitative Variable Distributions: Single Site, Anomaly Detection, Cross-Sectional Analysis

Created 2025-07-30

Data Requirements

```
cohort, qvd_value_file, omop_or_posnet, multi_or_single_site, anomaly_or_exploratory, age_groups, sd_threshold
```

Abstract

This check provides raw data and visualizations to aid a user in evaluating whether the distribution of quantitative variables aligns with clinical expectations. It can summarize the distribution of a quantitative variable (like lab result values) or patient counts (like number of patients with an outpatient visit).

How to Access This Check

- You may access the module's R package in GitHub. Or, run in R

```
install_github("ssdqa/https://github.com/ssdqa/quantvariabledistribution")
```

- Using the provided vignettes on GitHub or help in R, follow parameter input instructions for "Single Site", "Anomaly Detection", "Cross-Sectional" requirements.

Check Output

Visualization Output

This check outputs a bar graph visualizing the proportion of values that are considered outliers based on the number of standard deviations they fall from the mean. The outliers are stratified by upper and lower outlier types so the user can differentiate between outliers above and below the mean.

Files

1 Download

Tags

- Feasibility
- Bar Graph
- Event-Level Analysis
- Clinical Data Distributions
- Clinical Consistency
- Selection Error or Bias Detection